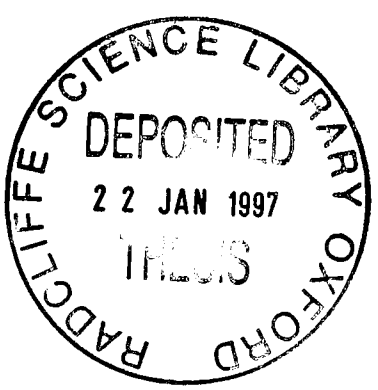
# The Sequence Characterisation of α-Globin Haplotypes in Human Populations

Jemma Townsend Norwich

Wolfson College
University of Oxford

Thesis submitted for the degree of Doctor of Philosophy
Hilary Term 1996



# The Sequence Characterisation of α-Globin Haplotypes in Human Populations

**Jemma Townsend Norwich**
Wolfson College, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy, Hilary Term 1996

## Abstract

Genetic variation at the human α-globin gene complex has only been characterised to date as haplotypes - a linked set of 7 RFLPs and 2 VNTRs - and not at the sequence level. However, an understanding of the extent of sequence variability at this locus, and its relation to underlying mutational processes, is of interest to studies of both molecular and population genetics.

This variation has been investigated by the development of a PCR haplotyping and direct sequencing protocol. Sequencing of 6 RFLP-containing regions of the haplotype in different human populations demonstrates that all (-) RFLPs have the same sequence, regardless of geographical origin. Additional sequence variation is found surrounding the RFLP sites, indicating that the overall level of variation at this locus is comparable to that at other loci, although a higher mutation rate is shown in an inter-species comparison. The additional polymorphic sites are concentrated in only three regions; a survey comprising a total ~280kb indicates that this sequence polymorphism is generated and influenced by localised gene conversion, which was not evident at the RFLP level. PCR analysis of the 2 VNTRs in human populations demonstrates considerable size variation in human populations, not detected by previous haplotyping protocols. There is also extensive internal sequence variation in the VNTRs (in a total survey of ~50 kb), apparently accumulating in a polar fashion and governed primarily by gene conversion, as reported at other VNTR loci. In addition, flanking sequence variation is identified which undergoes conversion, although not in concert with changes in the nearby repeat structure.

Consideration of the association of variation, at both the unique and repeated sequence regions, with RFLP α-haplotypes demonstrates that haplotypes differ in age, and allows relationships between them to be determined. The resulting haplotype network has similarities to that determined from β-globin gene variation; the implications for recent human evolution are discussed.

## Acknowledgements

I am grateful to the following people for invaluable contributions:

- John Clegg, my supervisor, for continual support and enthusiasm for this project
- Rosalind Harding, Tony Boyce and Jeremy Martinson for discussions about data
- Doug Higgs, June Roberts-Thomson, Adrian Hill, Laurent Excoffier and Trefor Jenkins for the DNA samples used in this study
- Doug Higgs and Jonathan Flint for information about the α-globin gene cluster sequence and global haplotype frequencies
- Jacques Rochette, Malia Fullerton, Janette Lamb, Jeremy Martinson, and David Rees for technical advice
- Jeremy Martinson, David Rees and Rosalind Harding for proof-reading some drafts of this thesis
- Tony Brett for assistance with computer and logistical problems
- Malcolm Shepherd and Peggy Willoughby for general lab support.

# Table of Contents

**The Sequence Characterisation of α-Globin Gene Haplotypes in Human Populations**


Chapter 1: The Study of Human Genetic Variation 2

The Types and Study of Genetic Variation 3
An Historical Perspective 3
Types of Human Genetic Polymorphism 4
Mechanisms Responsible for Observed Patterns of Genetic Variation 6
The Contribution and Testing of Neutral Theory 6
Linkage, Selection and the Consequences for Neutral Variation 8
Recombination 10
The Synthesis of Molecular and Population Genetics: The Globin Genes 19
Organisation of the Human Haemoglobin Genes 20
Mutational Mechanisms Affecting Globin Gene Variation 26
Globin Genes and Human Population Affinities 29
The Questions to Be Addressed 33

Chapter 2: The α-Globin PCR Haplotype and The Study of African Diversity 35

Haplotyping by PCR 35
Development of the PCR-Haplotype 35
The Problem of GC-rich Templates 36
Modification of Protocol for Efficient Amplification 37
Determination of α-Haplotypes by PCR 41
A Study of African Diversity: α-Haplotypes in Kenya 43
α-Haplotype Diversity of Luo-Speakers from Kenya 43
α-Haplotypes and African Prehistory 50
African Diversity and the Mechanisms Producing Diversity in Populations 52

Chapter 3: Sequence Variation at the α-Globin Gene Cluster 58

The Method of Direct Sequencing the α-Globin Complex 58
Design of Sequence Primers 59
Methods to Produce Single-Stranded Template 59
The Achievement of High-Quality Sequencing 60
Sequence Variation of the Unique Sequence Regions 64
Identification of Sequence Polymorphism 64
Sequence Polymorphism and Divergence at the α-Gene Cluster 67
Evolution of *Alu* Repeats in the α-Globin Gene Cluster 72
The α-gene Z Homology Boxes 77

Chapter 4: Variation at the Hypervariable Regions of the α-Haplotype 84

Characterisation of Variation at the IZHVR Locus 86
Size Variation 86
Sequence Variation at the IZHVR 92
Characterisation of Variation at the IVS1.HVR Locus 96
Size Variation at the IVS1.HVR in the ψζ1 and ζ2/ζ1 Genes 96
Sequence Variation within the IVS1.HVR 98
Mutational Mechanisms at α-Globin HVRs 102
Neutrality 102
Mutational Constraint and Polarity 105
HVRs and α-Haplotypes 107

Chapter 5: α-Globin Haplotype Relationships 110

Sequence Data from Non-Repetitive Regions and Haplotype Context 110
HVR Variation and Haplotypes 113
Size Variation at the IZHVR 113
Size and Sequence Variation at the ζ Genes 117
Determination of an α-Haplotype Network of Relationships 122
Principles of the Approach 122
The Relationships of α-Haplotypes 124
The Origins of Full Haplotypes 130

Chapter 6: α-Globin Gene Sequence Variation and Human Evolution 136

Human Evolution and Genetics 137
Hypotheses of Human Evolution and an African Origin for Modern Humans 137
α-Globin Molecular Divergence Times 139
Coalesence Times and Human Evolution 143
α-Haplotype-Related Sequence Variation and Recent Population Expansions 145
Future Studies 147

Appendix 1: Standard Materials and Methods 152
Appendix 2: Kenyan α-Haplotypes and α-Genotypes 164
Appendix 3: Sequencing Primers 166
Appendix 4: α-Globin Gene Cluster Sequences 168
Appendix 5: Non-α-Haplotyping PCRs 190

References 193

## Abbreviations

| | |
|---|---|
| bp - | base pair |
| cM - | centi-Morgan |
| CT - | coalescence time |
| DNA - | deoxyribonucleic acid |
| dsbs - | double-stranded breaks |
| DMSO - | dimethysulphoxide |
| HLA - | human leukocyte antigen |
| HVR - | hypervariable region |
| IVS1.HVR - | HVR in intron 1 of ψζ1/ζ1 genes |
| IVSII.HVR - | HVR in intron 2 of ψζ1/ζ1 genes |
| IZHVR - | interzeta hypervariable region |
| IAM - | infinite alleles model |
| kb - | kilobase ($10^3$ bp) |
| kyr - | thousand years |
| Mb - | megabase ($10^6$ bp) |
| mRNA - | messenger RNA |
| mtDNA - | mitochondrial DNA |
| myr - | million years |
| mya - | million years ago |
| $N_e$ - | effective population size |
| PCR - | polymerase chain reaction |
| PZ/Z - | pseudozeta(ψζ)/zeta (ζ) genes |
| RNA - | ribonucleic acid |
| RFLP - | restriction fragment length polymorphism |
| ss - | single-stranded |
| USCE - | unequal sister-chromatid exchange |
| U - | enzyme units |
| VNTR - | variable number tandem repeat |

# The Sequence Characterisation of α-Globin Haplotypes in Human Populations

# Chapter 1

## The Study of Human Genetic Variation

Genetic disease and congenital malformations occur in approximately 2-5% of live births, and genetic factors are likely have an important role to play in individual susceptibility to many common diseases [Weatherall, 1991]. A human genome map will therefore be of undoubted importance for the study of genetic disease by enabling one to associate genes with phenotype and to localise genes for disorders about which little is known. Ultimately, an entire human genome sequence will be determined, but given the financial burden and potential ethical implications of this goal, it is appropriate to question its usefulness. A single sequence overlooks the fact of individual variability; just how much of a simplification one sequence will therefore be, will depend on exactly how much variation is present between individuals, which is unclear at present.

This particular question has historically concerned population and evolutionary geneticists, since genetic variation is what the evolutionary process operates on. Inherited diseases represent only the "tip-of-the-iceberg" of human genetic variation; it is clear that much "normal" variation, with no apparent phenotypic effect also occurs. The nature of this variation, how much variation is present in human populations and why it is maintained are long-standing questions of population genetics, and not simply for the concern of purists: understanding variation is at the very heart of understanding of what constitutes "normal" or "abnormal". Further, little is known about why genetic diseases are present in human populations at all, considering that selection might have been expected to eliminate them. In order to address this issue, the forces of population genetics theory (selection, drift, migration and mutation) must be invoked [Flint *et al.*, 1993].

## The Types and Study of Genetic Variation

### An Historical Perspective

Following the transformation of Darwin's theory of evolution into "neo-Darwinism", by the incorporation of genetics, great efforts were made to provide an empirical basis to supplement what was a mathematical theory of population genetics. These attempts were difficult, however, because investigators' lifetimes are too short to observe population genetic processes. This all changed in the mid-60s with the advent of molecular techniques [Nei, 1987].

Initially, amino acid sequences were obtained, and this led to the extremely influential observation that the number of amino acid substitutions between species increased with time since divergence in an approximately linear manner. This "molecular clock" meant that evolutionary times of divergence between species could be estimated. However, the method is time-consuming and expensive, so other methods were developed. The first method is based on the relationship between the extent of immunological reaction (often against serum albumin) and number of amino-acid substitutions and was used to clarify the phylogenetic relationships of humans and apes, indicating a close relationship between chimpanzees, gorillas and humans [Sarich and Wilson, 1967]. Further, the divergence between humans and its closest relatives was estimated to be ~ 5 myr ago, somewhat different from the prevailing view of ~ 30 myr, from fossil data. DNA hybridisation, another alternative technique, estimates the underlying sequence differences between two species from the thermostability of their hybrid DNA. Applied to hominoids, the close relationship between chimpanzees and humans was again shown, and the split estimated to have occurred ~ 7 myr ago [Sibley and Ahlquist, 1984].

The development of a simpler protein electrophoresis method [Hubby and Lewontin, 1966; Lewontin and Hubby, 1966; Lewontin, 1991] moved the emphasis of study away from species relationships to genetic polymorphism within species. This method only detects a fraction of amino-acid changes in proteins (those causing a new change in protein charge, which is estimated to be ~ 25% of amino-acid changes), but still demonstrated extensive protein variation in a large number of organisms, and has enabled human population variation to be characterised [Nei and Roychoudhury, 1982]. More recently, the technical breakthroughs of gene cloning, restriction enzyme methods and rapid DNA sequencing has not only revolutionised molecular biology and uncovered many hitherto unexpected aspects of genic structure (such as exons/introns and pseudogenes), but has also provided molecular and population genetics with a wealth of data with which to address fundamental questions. For practical reasons, many studies of nuclear sequence variation have been done with *Drosophila* species [Aquadro, 1992; Brookfield and Sharp, 1994; Moriyama and Powell, 1996], since the problems of diploidy[1] can be overcome by the use of inbred strains and chromosomal extraction. However, the development of increasingly more sophisticated PCR technologies means that, increasingly, similar surveys of human variation are now becoming possible (see [Fullerton *et al.*, 1994]).

**Types of Human Genetic Polymorphism**

It is only a small minority of DNA polymorphisms that are responsible for the amino-acid differences detected at the protein level. Studies have shown that DNA variability is also due to other nucleotide substitutions (not causing amino-acid changes), deletions, insertions and inversions. Variation in eukaryotic genomes also

[1]The allelic sequences of multiply-polymorphic diploid individuals are difficult to resolve.

occurs at a chromosomal level, with spontaneous meiotic errors resulting in alteration of the number of chromosomes, or in chromosomal translocations. Genetic mutations may result from spontaneous errors during replication or repair, environmental agents (such as ionising radiation) or sequence rearrangements facilitated by recombination. In coding DNA, nucleotide substitutions are the major type of genetic mutation (insertion/deletion mutations are likely to result in frameshift mutations) and, because of the degeneracy of the genetic code, may result in the same amino-acid (synonymous substitutions) or in a different amino-acid (replacement substitution).

In non-coding DNA, short deletion and insertion polymorphisms (one to a few bp) are quite common (see Efstratiadis *et al.*, [1980]), representing errors in DNA replication, although larger length polymorphisms may result from unequal crossing over (recombination-see later in this chapter) and DNA transposition at repetitive sequences. Repetitive sequences are estimated to constitute ~ 50 % of the human genome [Strachan, 1992] and may be interspersed or tandemly repetitive. There are two interspersed repeat families in the human genome: SINEs (short interspersed repeats) and LINEs (long interspersed repeats) (reviewed in Weiner *et al.*, [1986]). The *Alu* SINEs (most have an *Alu I* enzyme recognition sequence) are the most abundant repeat sequence family in primate genomes, being present at copy numbers of greater than half a million per haploid genome [Strachan, 1992]. They are thought to have mobilised through primate genomes by retroposition; that is, generation of an RNA polymerase III transcript, reverse transcription and integration at staggered breaks in the genome. *Alu* repeats in the human genome are divided into subfamilies of distinct genetic ages (for example, Britten *et al.*, [1988]; Jurka and Smith, [1988]). While most *Alu* repeats were inserted into the genome millions of years ago, some are young and are not yet fixed in human populations, thus representing a novel form of polymorphism (see Batzer *et al.*, [1994]). An

answer to the question of what mutational changes occur in the human genome, although much work has concentrated on coding sequences and thus point mutations, must encompass not only substitutions, but also other mechanisms such as occur at repetitive sequences.

## Mechanisms Responsible for Observed Patterns of Genetic Variation

### The Contribution and Testing of Neutral Theory

The discovery of extensive protein variation led to controversy over the mechanisms of maintenance of genetic variation in natural populations. Prior to electrophoretic studies, a relatively small number of polymorphic characteristics (such as blood groups) were known and polymorphism was believed to be maintained by selection (see Lewontin, [1991] for review). Extensive polymorphism and the finding that the rate of sequence change during evolution varies considerably with the DNA region examined, so that the more functionally important the region is (for example, histone genes) the lower the rate, both required explanation and led to the formulation of the neutral theory [Kimura, 1983]. This postulates that genetic drift (change in allele frequencies due to the chance sampling of alleles from generation to generation) is the dominant force driving molecular evolution: most mutations are deleterious and therefore rapidly eliminated from the population by selection. The remaining variation, or polymorphism, is neutral (with no phenotypic effect) and transient, and the amount in a population at equilibrium is determined by population size and mutation rate: in a small population, drift is more influential than in a large population. The role of selection is thus to eliminate less fit genotypes and save beneficial ones, and although it may speed up gene substitution in a population, molecular evolution occurs primarily by mutation pressure [Nei, 1987].

Accumulating sequence data is allowing progress to be made in assessing the relative contributions of mutation, selection and drift to variation within populations, but the neutral theory remains controversial for a number of reasons. From a theoretical point of view, many problems are due to the fact that molecular evolutionary studies rely on comparative analyses and therefore are concerned with patterns rather than process [Gillespie, 1987; Kreitman, 1987]. Further, although many patterns are expected from neutral theory, such as the relative abundance of synonymous substitutions compared to replacement substitutions in regions of high functional constraint, the same patterns can also be accommodated by a selectionist argument (see Gillespie, [1987]). However, selection models have not been extensively developed, mainly because of the comparative complexity of the mathematics; in their absence and with recent modifications of the neutral theory (particularly the dependency of mutational neutrality on population sizes, see Kreitman, [1987] for discussion), all manner of results can be compatible with neutrality.

This is not to detract from the inestimable value of the neutral theory in producing an elegant null hypothesis against which genetic data can be evaluated. Also, the recent development of neutral tests has ensured that more suitable use of DNA sequence data can now be made: neutral theory predicts that heterozygosity[2] and the rate of evolutionary change are both proportional to the neutral mutation rate, and therefore should be correlated [Kreitman, 1987]. Tests of this prediction, in *Drosophila* species, have indicated that some loci do behave neutrally while others do not [Brookfield and Sharp, 1994; Hudson *et al.*, 1987; McDonald and Kreitman, 1991] [Moriyama and Powell, 1996]. When neutral expectations are not met, molecular mechanisms, including recombination, and demographic factors (such as population size variation) are of importance, in addition to selection [Moriyama and Powell,

[2]Heterozygosity is the proportion of heterozygotes for a given locus in a population.

1996]; in most instances, it is likely that a complex interplay of molecular and population genetic forces is responsible for the patterns of variation, making the task of separating such influences into their relative importance near impossible.

**Linkage, Selection and the Consequences For Neutral Variation**

When genes are on different chromosomes, segregation into germ cells occurs independently; when close together on the same chromosome, they are linked and inherited as a unit. The degree of linkage will depend on how far apart the loci are and thus the likelihood of being disrupted by the process of recombination at meiosis. The principle of distance between genes being a function of the probability that linked genes will be separated by recombination, has been known for more than 50 years from the time of T. H. Morgan's work on *Drosophila*. With the advent of the appropriate DNA recombinant technology, the same principles could be applied to humans and have allowed detailed genetic maps to be produced, with most markers identified so far having a spacing of a few cM (in humans, 1 cM[3] is roughly equivalent to $10^6$ bp) [Dib *et al.*, 1996; Weissenbach *et al.*, 1992].

Linkage disequilibrium has been identified genome-wide in humans [Peterson *et al.*, 1995] and, at the sequence level, can extend for many hundreds of kb (for examples: Charmley *et al.*, [1990]; Harvey *et al.*, [1995]; Thomas *et al.*, [1995]). The first detailed study of RFLPs, in the human β-globin gene complex, indicated that one RFLP was present every 500 bp [Jeffreys, 1979] and therefore such large stretches of linkage disequilibrium usually include more than one RFLP or polymorphic marker; these linked markers constitute highly informative haplotypes.

[3]One centiMorgan (cM) or map unit is a distance measurement that corresponds to ~1% frequency of recombination between two loci.

Selection for even the occasionally advantageous mutation can influence neutral polymorphism, as a result of linkage. If balancing selection[4] is operating on a mutation, polymorphism at linked neutral sites, in a population at equilibrium, will be maintained; alternatively, if directional selection is occurring at a selected site, linked polymorphism will be lower, following what is known as a "selective sweep". Linked polymorphisms are said to "hitch-hike" [Maynard-Smith and Haigh, 1974] and the size of the hitch-hiking effect, either reducing or maintaining polymorphism, will largely depend on the recombination rate [Kaplan *et al.*, 1989]. Historically, in the absence of molecular data, there has been debate about the dominant role of selection as either a purifier or maintainer of polymorphism [Lewontin, 1991]; clearly, this depends on whether selection tends to be balancing or directional.

However, the results of observed linkage on genetic variation, also depend on the underlying mutational mechanisms of recombination, and to demographic factors, such as population size (see Jorde, [1995] for review). To take specific examples, sickle cell anaemia and thalassaemia represent some of the commonest single gene disorders; they are so common in many populations because of carrier protection against *Plasmodium falciparum* malaria [Allison, 1964; Flint *et al.*, 1986; Siniscalco *et al.*, 1966]. However, the linkage between the various underlying mutations and other neutral sites (haplotypes) shows geographical differentiation and thus the influence of drift. Thalassaemias are a group of hereditary anaemias in which various defects in haemoglobin synthesis are present as a result of defective globin genes. Mutations causing thalassaemia in the two human globin gene clusters, $\alpha$- and $\beta$-, are linked to particular haplotypes, but the specific linkage depends on which population is studied.

[4]The frequency of the advantageous heterozygote increases until the incidence is balanced by loss of homozygotes from the population.

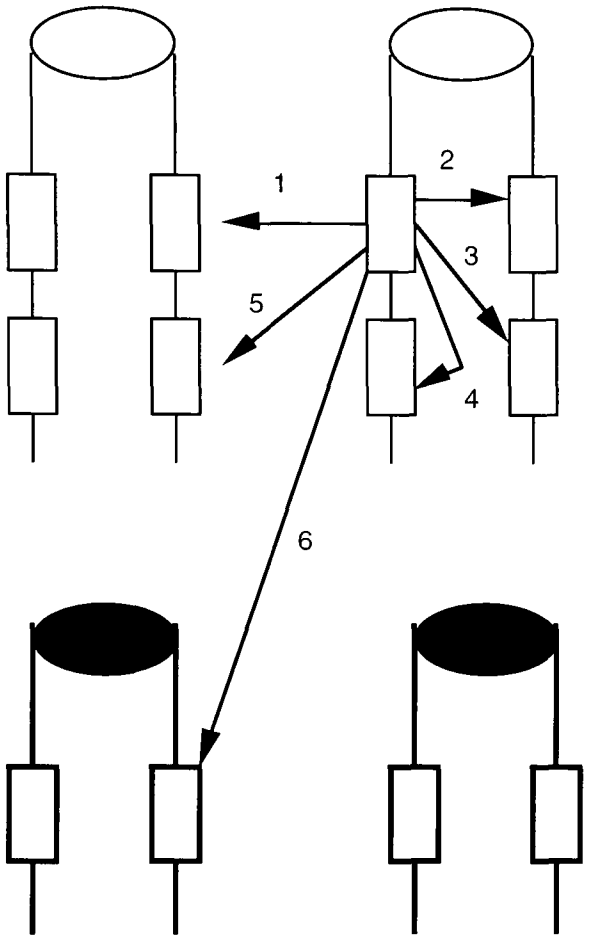
The linkage relationship between RFLPs and human mutations is therefore complex, as has also been observed in studies of *Drosophila* [Schaeffer and Miller, 1993]. Due to the high frequency of thalassaemia, there has been considerable incentive to attempt disease prevention by the implementation of prenatal diagnosis. Such diagnosis can be performed directly by restriction enzyme digestion of foetal DNA for sickle cell and some β-thalassaemia mutations [Weatherall, 1991], or by RFLP linkage [Kan and Dozy, 1978]. However, because of geographical differences in linkage, family studies are required to detect the chromosome carrying the thalassaemia mutation. This highlights the importance of population context, something which the development of neutral theory naturally emphasises. The actual observed patterns of variation depend also, however, on recombination. Although effective absence of recombination and consequent linkage affects levels of variation, the action of recombination is important by destroying linkage and therefore the effects of a mutation either drifting towards, or being selected towards, fixation.

## Recombination

### Models of Recombination

Recombination itself is not well-understood, despite its utility in the construction of genetic maps and its apparent consequences for genetic variation. Recombination produces new associations of genetic information in both mitotic and meiotic cells, although only in meiotic recombination does the process have evolutionary consequences. There are a number of types of recombination depending on the interacting homologous loci and their relative locations in the genome (Figure 1.1). These interactions can be reciprocal (crossover) or non-reciprocal (gene conversion).

Figure 1.1 Types of Recombination




If meiotic recombination occurs after DNA synthesis and involves only genes represented in one copy per haploid genome, recombination can occur either:
1) between genes located on non-sister chromatids of homologous chromosomes (classical recombination), 2) between genes located on sister chromatids (sister strand recombination). If the gene is duplicated on one pair of homologues, recombinational interactions could include: 3) unequal sister-strand, 4) intrachromatid, 5) quasiclassical. If genes are repeated on non-homologous chromosomes: 6) heterochromosomal recombination. 3-6 can be termed "ectopic" recombination; 2-4 constitute intrachromosomal recombination. After Petes *et al.*, [1991].

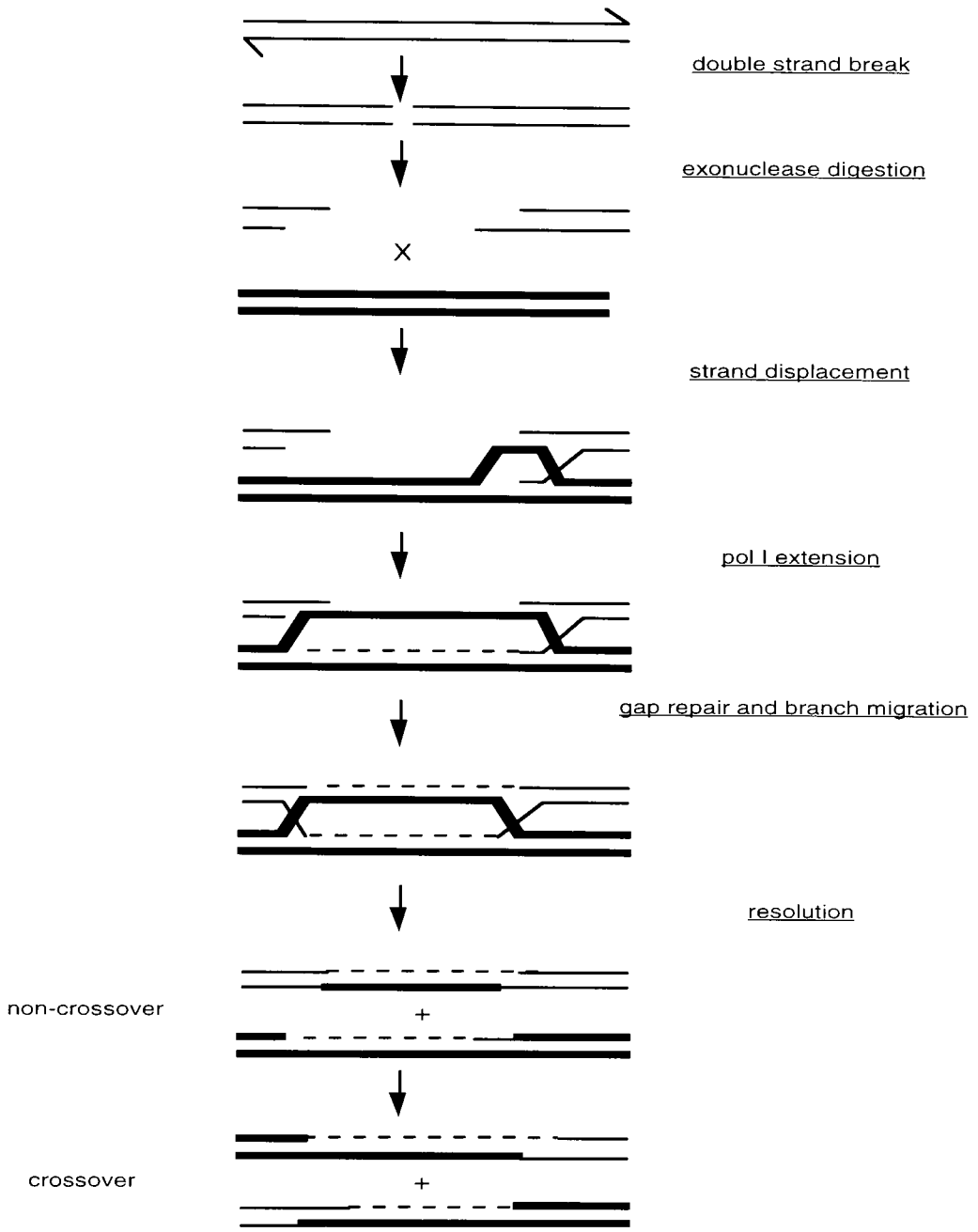
Eukaryotic recombination has been studied most intensively in yeast since, unlike other eukaryotes, all the haploid meiotic products are retained in an ascal sack and can be directly analysed [Petes *et al.*, 1991]. However, the development of sperm typing has had a dramatic effect on mammalian recombination studies [Chakravarti, 1994; Hubert *et al.*, 1994]; prior to this, recombinant mutations had to be identified in family studies (provided mutations occur frequently enough to be detectable), or linkage of genetic markers was studied to determine where and how often recombination occurs. Sperm typing has the advantage of allowing an unlimited number of haploid meiotic products to be collected and analysed and, although the techniques involved are technically demanding, it is now possible for fundamental contributions to be made towards the understanding of recombination from humans and other mammals. The problem remains, however, that sperm are the product of a number of mitotic as well as meiotic divisions, so that observed recombination is not unambiguously meiotic in origin. In addition, double crossovers and gene conversion are difficult to distinguish without access to the products of a *single* meiotic division; fortunately, the existence of positive interference does suggest that double crossovers are less likely than gene conversion [Kwiatkowski *et al.*, 1993], validating its inference from observations of apparent genetic transfer in mammalian systems.

Conversion and crossover are associated events in yeast and other organisms [Borts and Haber, 1987; Petes *et al.*, 1991], with the degree of association depending on the locus studied. Most models of recombination explicitly interpret this as signifying a mechanistic connection, with conversion portrayed as an intermediate in the production of crossovers. Meiotic (but not mitotic) conversion is usually polar with frequencies higher at one end of a gene than at the other [Malone *et al.*, 1994; Nicholas *et al.*, 1989; Petes *et al.*, 1991] suggesting initiation of conversion at a particular site, with frequencies changing in a distance-dependent manner away

from the site. The distance over which conversion is observed to occur, the conversion tract, varies in length according to locus and organism: an average of ~1.5 kb has been observed in yeast [Borts and Haber, 1989], 50 bp-1 kb in *Drosophila* [Curtis and Bender, 1991; Hilliker *et al.*, 1994] and ~1-300 bp in human HLA genes [Zangenberg *et al.*, 1995] and γ-globin genes [Powers and Smithies, 1986].

Models differ in the mode of initiation and the precise nature of the conversion process, but otherwise are fairly similar and consecutive models generally retain many features of earlier models [Stahl, 1994]; indeed, until recombination was studied at the molecular level in yeast [Nicolas *et al.*, 1989], it was difficult to determine which model was more appropriate. The initiating event in recombination has been envisaged as a single-strand nick [Meselson and Radding, 1975], although the evidence now weighs in favour of the double-strand break repair model [Szostak *et al.*, 1983] (Figure 1.2). In the latter, following a double-strand break (dsb) on one chromosome, formation of points of exchange, or Holliday junctions, branch migration and heteroduplex repair, result in either gene conversion or crossover (with or without associated conversion). As originally conceived, the model suggested that most gene conversion would occur by repair of sequence differences between the duplexes contained within the double-strand gap. However, the finding that the 3' ends (Figure 1.2) can be very extensive [Sun *et al.*, 1991] indicated that heteroduplex formation may also be extensive, and mismatch repair therefore of considerable importance in gene conversion, as emphasised in earlier models [Meselson and Radding, 1975]. Other evidence comes from yeast mutants, with defective repair mechanisms, which show a decreased number of conversion events with a concomitant increase in post-meiotic segregation products (where the two strands of one DNA duplex have different sequences; reviewed in Petes *et al.*, [1991]) and an absence of a conversion gradient [Alani *et al.*, 1994]. Further, increased heterology of the participating duplexes has been shown both to increase the

Figure 1.2: The Double Strand Break Model for General Recombination




Enzyme activity creates a gap from a dsb, allowing one of the 3' ends to displace the strand on the other duplex. DNA polymerase action extends the 3' end of the invading strand and increases the size of displacement loop. The other single-stranded end pairs with the displaced loop and further DNA synthesis occurs. Ligation of the free ends results in two Holliday junctions which may migrate to form heteroduplex DNA. Resolution, by cutting junctions, may occur in a number of possible places; if isomerization (twisting) occurs, which changes the strands that are crossing, a crossover type event rather than simple gene conversion will result.

likelihood of meiotic conversion events and aberrant recombination versus crossover events [Borts and Haber, 1987] (although this result has not been obtained at all yeast loci [Malone *et al.*, 1994; Schultes and Szostak, 1990]) and to reduce mitotic crossover rates [Datta *et al.*, 1996]; this evidence also indicates an important role for mismatch repair.

There is considerable evidence that eukaryotic recombination operates primarily according to the dsb-repair model. Initiation of conversion in yeast has been shown to be associated with the formation of dsbs [Liu *et al.*, 1995; Nicholas *et al.*, 1989; Sun *et al.*, 1989]; also, double Holliday junctions have been observed following single strand exchange [Schwacha and Kleckner, 1995], conversion tracts form on both sides of the initiating site [Schultes and Szostak, 1990], and the duplex which suffered dsbs was found to be the recipient (not donor) of the information transferred during conversion [Nicholas *et al.*, 1989], all of which is consistent with the dsb, but not single-strand nick model. Dsbs also correlate with the observed cytogenetic distributions of chiasmata (the physical points of crossover), they form just prior to the formation of recombination products, and in yeast mutants for the protein RAD50, known to be required for yeast recombination, dsbs do not form [Petes *et al.*, 1991].

## The Non-Random Nature of Recombination

### *Physical Mapping*

It has become apparent from the comparisons of genetic and physical maps that recombination does not occur randomly along chromosomes. Mammalian chromosomes are longitudinally heterogeneous in structure and function, which is the basis of specific banding patterns produced by a variety of chromosomal

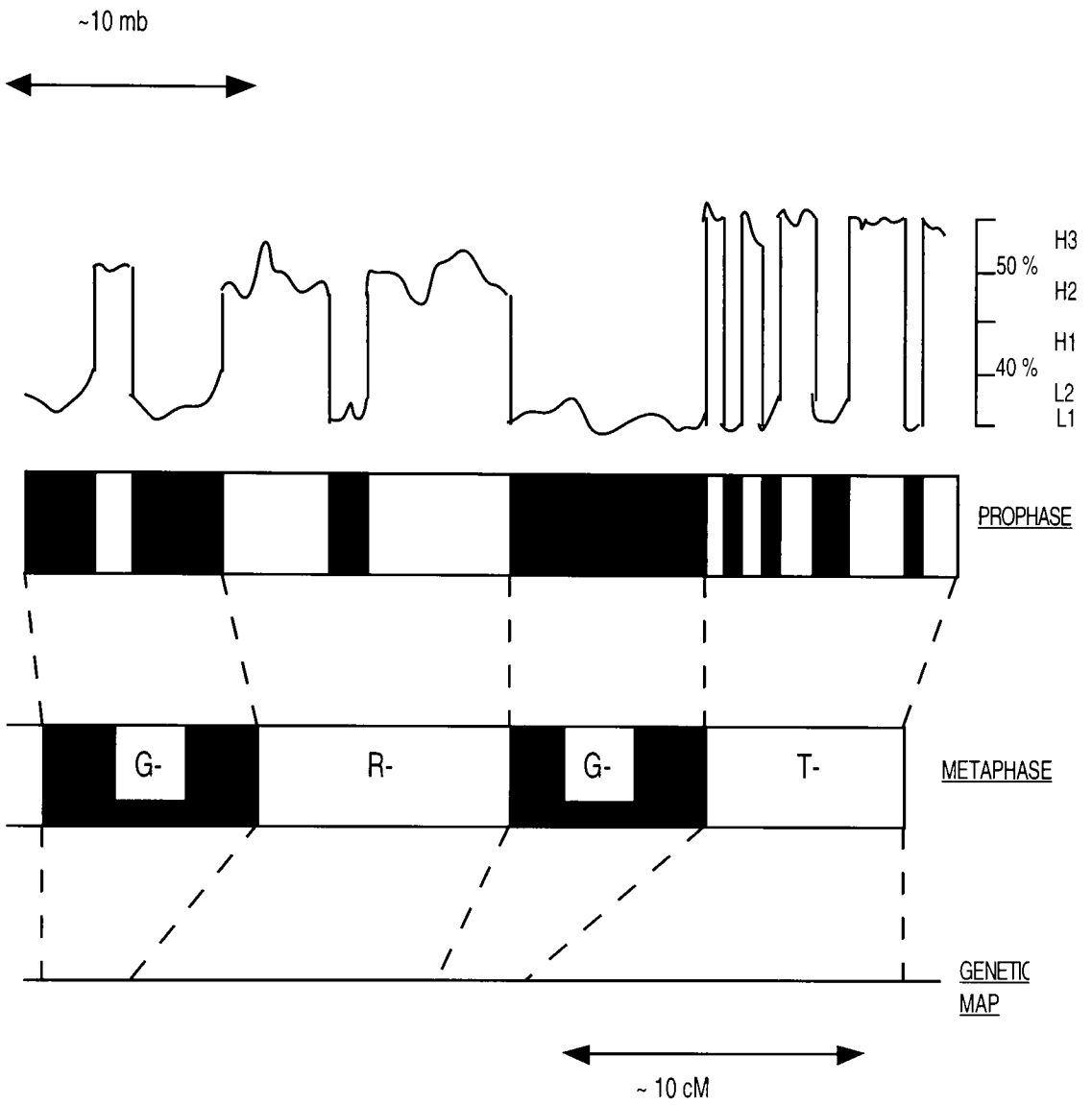
staining techniques. "G" (Giemsa) and "R" (Reverse) bands have complementary patterns at metaphase (Figure 1.3) and represent two different functional compartments of human euchromatin. They are characterised by differences in a number of interdependent properties, including gene density, replication timing, base composition, repeat sequence density, chromatin packaging, likelihood of cancer breakpoints and radiation-induced aberrations, and number of chiasmata (for reviews, see Craig and Bickmore, [1993]; Holmquist, [1992]).

R bands contain the most genes (including almost all house-keeping genes), are GC-rich, have the greatest number of chiasma and a high concentration of *Alu* repeats. Subdivision of R bands into four "flavours" is possible on the basis of the specific GC- and *Alu*-richness [Holmquist, 1992]; the bands with the highest GC and *Alu* repeat content are preferentially located at the telomeres (T-bands) [Saccone *et al.*, 1992] and, although these bands represent only about 15% of all bands, contain more than 60% of mapped genes [Holmquist, 1992]. At prophase, when band resolution is higher than at metaphase, R bands can be seen to consist of a number of fused sub-bands of varying characteristics, including some G-type sub-bands (Figure 1.3). These sub-bands are considered to be equivalent to the long stretches of DNA of homogeneous base composition known as isochores [Bernardi, 1989](legend to Figure 1.3).

*Recombination Hotspots*

There is an underlying genomic rate of recombination which is characteristic for any one organism; for example, ~1 cM/Mb in humans, as discussed earlier, and 0.39 cM/kb in yeast [Petes *et al.*, 1991]. However, rates are generally higher near the ends of chromosomes (close to the telomeres in T-bands) and lower near the centromere, and comparatively high rates are also observed at the level of individual loci, such as

Figure 1.3: Base-Composition along the genome and its relation to cytogenetic and genetic maps




This schematic model is based on that of Holmquist, [1992]. Dark and light boxes represent G (Giemsa +) bands and R and T (Reverse or G-, Telomeric) bands, respectively, at metaphase; these bands have characteristic structural and functional properties. At a higher level of resolution at prophase, bands consist of a number of subbands; the properties of the metaphase band represents an average of the fused subbands. T bands consist of the largest number of subbands. The prophase bands are considered to represent isochores, large regions of sequence, homogeneous in base composition. There are five isochore families (H3, H2, H1, L2, L1) of varying % GC; most genes are found in the H3 isochore in telomeric regions.

those in the human β-globin gene complex, insulin and Duchenne muscular dystrophy genes [Chakravarti *et al.*, 1984; Chakravarti *et al.*, 1986; Grimm *et al.*, 1989]. Molecular analysis of such recombination "hotspots" indicates that these are regions of preferential formation of dsbs and hence initiation of recombination [Nicholas *et al.*, 1989; Sun *et al.*, 1989; Wu and Lichten, 1994]; regions that conform to the overall genome recombination rate also show dsbs, but at a lower rate [Lichten and Goldman, 1995; Wu and Lichten, 1994]. This is further evidence for the dsb-repair model of recombination discussed earlier.

Many of the initiating sites of hotspots occur within the promoter regions of eukaryotic genes [Nicholas *et al.*, 1989; Petes *et al.*, 1991; Wu and Lichten, 1994] which has led to the suggestion that meiotic recombination is related to transcription, as seems to be the case for mitotic recombination where transcription relates to increased formation of dsbs and recombination [Baker and Read, 1995; Thomas and Rothstein, 1989]. Since formation of dsbs is concentrated within restricted regions such as the promoter, but is not apparently sequence-specific [Liu *et al.*, 1995; Massy *et al.*, 1995], the association between recombination and transcription is likely related to chromatin structure [Wu and Lichten, 1994]. Transcriptionally active DNA is nucleosome-free and in an open conformation which may make the DNA more accessible not only to transcriptional protein machinery, but also that of recombination. Open chromatin is also more sensitive to DNase I and nucleases and therefore possibly to the enzymes which form dsbs.

The importance of chromatin context for recombination is demonstrated by studies of the M26 hotspot, located in the *ade6* gene of *Schizosaccharomyces pombe*, created by a simple G to T mutation within a seven bp sequence. If the gene is moved to a different region of the genome, it no longer behaves as a hotspot [Ponticelli and Smith, 1992]. It is therefore possible that *cis*-acting sequences, at the promoter, alter

chromatin structure and are responsible for this hotspot's activity [Zahn-Zabal *et al.*, 1995]. In eutherian mammalian systems, recombination hotspots are often sex-specific, with telomeric and other hotspots showing a higher rate in males than females. Interestingly, the positions of loci with sex differences in recombination correlate with the regions of the genome which show mono-parental methylation and expression (imprinting) [Paldi *et al.*, 1995]. Since methylation affects chromatin structure, this seems to confirm a chromatin effect on local and globalised rates of recombination.

*A Role for Gene Conversion?*

Correct chromosomal segregation at meiosis requires the presence of at least one chiasma per chromosome, which behaves somewhat akin to mitotic centromeres. Prior to this, however, homologous chromosomes must recognise each other, and pair up on the spindle. Until recently, the central dogma of meiosis could be summarised as follows: synapsis to recombination to segregation. It is now clear that pairing and synapsis are mechanistically and temporally distinct processes [Hawley and Arbel, 1993]. Recombination is envisaged as playing a role in the homology searches that must occur before correct chromosomal pairing is possible. However, the recombination intermediates must resolve as gene conversion rather than as crossover, so that inadvertent translocations do not occur if mispairing occurs [Carpenter, 1987]. Small gene conversion events have been suggested to represent the "footprints" of aborted recombination during homology searching [Carpenter, 1987; Powers and Smithies, 1986]. Once correct pairing takes place (perhaps initiated at preferential places in the genome, which are recognisable as recombination hotspots), the synaptonemal complex can allow synapsis to be completed and, at this stage, some recombination intermediates may mature as crossovers and ensure disjunction.

Recombination at VNTRs and Other Repeats Sequences

Tandemly repetitive loci, first reported in 1980 [Wyman and White, 1980], are numerous throughout the genome and exhibit high levels of polymorphism caused by variation in the number of tandem repeats in an array. The repeated sequence in different arrays may range in size from one to several hundred base pairs, and division is made between arrays composed of repeats of different size: 1-5 bp repeats are found in microsatellites, ~10-100 bp minisatellites [Jeffreys *et al.*, 1985a], VNTRs [Nakamura *et al.*, 1987] or HVRs, and 100 bp upwards constitutes satellite DNA. To some extent, this classification is arbitrary, especially given evidence that some microsatellites and VNTRs are related to each other [Armour *et al.*, 1995] (and also possibly between microsatellites and *Alu* repeats, [Arcot *et al.*, 1995]).

VNTRs tend to be restricted to subtelomeric, GC-rich regions that are characterised by high rates of recombination, and are often found close to other types of repeat [Armour *et al.*, 1989]. Satellite DNA is also confined, but to centromeric regions, whereas microsatellites tend to be more widely distributed throughout the genome, with a possible tendency to mirror the distribution of minisatellites [Weissenbach *et al.*, 1992]. The localisation of VNTR arrays does not limit their applications, however; gene mapping can utilise VNTRs [Nakamura *et al.*, 1987] because many genes are also located to subtelomeric regions, as indicated earlier. The high polymorphism associated with VNTRs means that they are also widely used in paternity testing and forensic science [Jeffreys *et al.*, 1985b] and in population genetics [Deka *et al.*, 1991; Flint *et al.*, 1989; Martinson *et al.*, 1993].

Some GC-rich VNTR repeat sequences show a similarity both to each other, indicating a consensus or core sequence, and to the Chi ($\chi$) sequence of

bacteriophage and *E.coli* [Jeffreys *et al.*, 1985a]. The χ sequence is recognised by the RecBCD protein which is required for recombination in *E.coli.*, and the similarity between the core sequence and χ, coupled with the high levels of VNTR polymorphism, led to the suggestion that the rate of production of new alleles must be high and, since the presumed mechanism for allelic length changes was unequal crossover, that minisatellites are hotspots for meiotic recombination [Jeffreys *et al.*, 1985a]. Many of the mammalian hotspots for recombination have been reported to be associated with repeated sequences, including VNTRs and microsatellites, for example [Chakravarti *et al.*, 1984; Chakravarti *et al.*, 1986; Steinmetz *et al.*, 1986; Wahls *et al.*, 1990b], and the core sequence does stimulate homologous recombination in human cells [Wahls *et al.*, 1990a]. Male and female germ lines have similar rates of evolution at some VNTR loci and higher rates of length mutation are observed in germline than in somatic tissues [Jeffreys *et al.*, 1988b], also suggesting a role for meiotic recombination. However, not all VNTRs are meiotic recombination hotspots, and some minisatellites do not possess homology to the core sequence and are AT-rich, such as that 3' to the apolipoprotein B gene [Knott *et al.*, 1986], yet are still highly polymorphic. It is also unclear whether VNTRs *cause* high rates of recombination or "just happen" to be present in regions with high recombination [Cullen *et al.*, 1995; Dover, 1989; Jarman and Wells, 1989; McGinnis and Spielman, 1994].

It has become apparent that when new VNTR array sizes are produced, sister-chromatid, rather than interchromosomal exchange, is responsible since flanking markers are very rarely exchanged [Higgs *et al.*, 1986; Martinson *et al.*, 1994; McGinnis and Spielman, 1994; Renges *et al.*, 1992; Wolff *et al.*, 1988; Wolff *et al.*, 1989]. Since mutant alleles are generally similar in size compared to their progenitors [Jeffreys *et al.*, 1988a], replication slippage may also be important for repeated sequence evolution. Slippage is the mispairing of two DNA strands of a single DNA

duplex [Levinson and Gutman, 1987], and is the dominant force in microsatellite evolution [Valdes *et al.*, 1992]. Recent studies of internal mapping and sequencing of VNTRs, however, indicate that size and sequence changes are often complex, frequently involving gene conversion [Armour *et al.*, 1993; Jeffreys *et al.*, 1994]; exchange resolved as conversion could be interallelic as well as intrachromosomal, since crossover of flanking markers would not always be observed.

It has also recently become clear that microsatellite array evolution is more complex than previously assumed under simple models of replication slippage. The discovery of the involvement of unstable trinucleotide arrays undergoing large-scale tandem repetition and "dynamic" mutation [Richards and Sutherland, 1994] in a number of genetic diseases, including fragile-X syndrome, suggests that other loci may be responsible for the mutability of repeat arrays. In the genome-wide instability of microsatellites seen in cancer cells, there is an inability to repair errors made during replication [Strand *et al.*, 1993], highlighting a role for complex protein (mismatch repair) systems modifying microsatellite mutation [Datta *et al.*, 1996; Kunkel, 1993]. At present, the best conclusion is that it is likely that a number of mutational processes operate on tandem repeat arrays [Dover, 1993], not all of which are fully understood.

## The Synthesis of Molecular and Population Genetics: The Globin Genes

A molecular and population genetic perspective is central to explaining fully why the genome is variable, and therefore giving a context for understanding disease variation. It is significant that the preceding discussion has referred to the various disorders of haemoglobin; since molecular biology has been applied to human disease, the structure and function of haemoglobin and its variants have been studied most intensely, and in population contexts. Thus, not only is much known

about haemoglobin gene function, but also analysis at the DNA level has enabled progress to be made towards understanding the roles of population genetic forces. Further, the detailed studies of haemoglobin variation have led to contributions towards an understanding of the evolutionary history of the human species through studies of population affinities (for reviews, Flint *et al.*, [1993]; Hill and Wainscoat, [1986]).

**Organisation of the Human Haemoglobin Genes**

Haemoglobin is found in a wide variety of organisms, from plants to vertebrates, and is believed to have been ancestrally encoded by one gene, which probably consisted of three exons, separated by two introns (or possibly four exons and three introns, see Stoltzfus and Doolittle, [1993]). The positions of the introns relative to the coding sequence is conserved in globin genes of all species. Vertebrate haemoglobin consists of a heterotetramer formed by two pairs of polypeptide chains encoded by the α- and β-like globin genes. The genes represent an ancient duplication event of the ancestral globin gene, early in vertebrate history (~ 500 Myr ago [Efstratiadis *et al.*, 1980]); another such duplication event produced the myoglobin gene [Weller *et al.*, 1984]. In amphibians, the α- and β-genes are present on the same chromosome, but in birds and mammals, they are present on different chromosomes.

The human α-globin gene complex is located in the sub-telomeric region of chromosome 16 (16p13.3; [Buckle *et al.*, 1988]) and is present within a long (>200 kb) GC-rich isochore [Ikemura and Aota, 1988], with the properties of T bands; that is, high levels of recombination, *Alu* repeats and early replication [Craig and Bickmore, 1993; Higgs *et al.*, 1989]. Conversely, the human β-globin gene complex is located interstitially on chromosome 11, within an AT-rich region (G band) with very

different properties to the α-globin complex [Ikemura and Aota, 1988; Vyas *et al.*, 1992]. One consequence of the different genomic environments is that the two clusters are subject to different regulatory mechanisms. Co-ordinate regulation of the α- and β-globin gene clusters must occur for a balanced production of α and β chains to be achieved, but there are distinct mechanisms by which the two clusters achieve tissue specificity and control of transcription. Both α- and β globin gene clusters require upstream regulatory elements, but the β-globin element has to alter the chromatin structure to allow transcription to take place, whereas the α-globin element does not, since the cluster lies within constitutively open chromatin [Craddock *et al.*, 1995; Dillon and Grosveld, 1993; Sharpe *et al.*, 1992; Vyas *et al.*, 1992].

Differences in regulation are not the only consequence of the two clusters being present in different isochores; the many other properties of different chromosomal bands are likely to have considerable influence on the levels of variation that exist within the two clusters. Comparison of the intergenic regions of the two clusters in mammalian species has already suggested that the α-globin gene complex may be mutating faster, and it has been postulated that this is a consequence of the different chromatin contexts (and replication timing) of the clusters [Hardison and Miller, 1993].

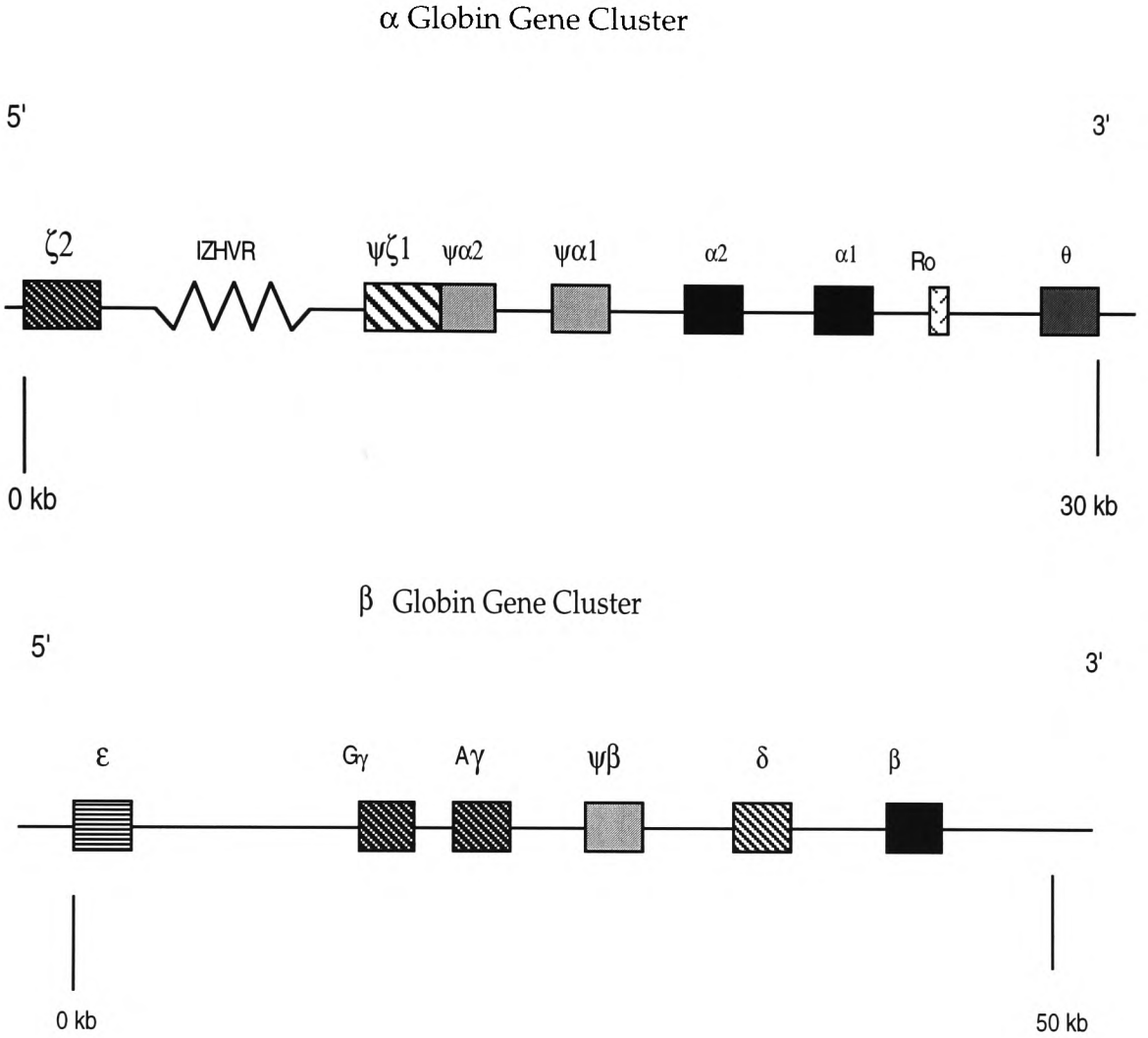
All species that use haemoglobin to carry oxygen have different forms of the molecule at various developmental stages according to the different requirements during development. This is due to further gene duplication events of both α- and β-globin genes to form gene clusters, with some of the duplicated genes undergoing mutation to acquire modified functions. The number of, and spacing between, these duplicated genes in each cluster varies between different species, and some genes have become inactivated to form pseudogenes in some species, although they may

remain functional in others. Gene families of this sort are not uncommon, and a large percentage of actively expressed human genes are members of families showing a high degree of sequence similarity, but which vary in function [Strachan, 1992]. Such gene families have led to the widely accepted scenario that following duplication, gene copies are freed from functional constraint and evolve neutrally until a new function happens to be acquired [Ohno, 1970]. However, it is perhaps more likely that new functions evolve under selection, and that once neutrally evolving pseudogenes arise by mutation, they remain inactive [Hughes, 1994].

In the α-globin gene cluster, duplication formed the ζ genes (~400 Myr ago; [Proudfoot *et al.*, 1982]), and in the β–globin gene complex, more recent duplications have formed the γ (from β) and ε (from γ) genes [Efstratiadis *et al.*, 1980](Figure 1.4). These duplicated genes are involved in haemoglobin (Hb) switching during development, from embryonic Hb (ζ2γ2 - Hb Portland; ζ2ε2 - Hb Gower 1; α2ε2 - Hb Gower 2) → foetal Hb (α2γ2 - Hb F) → adult Hb (α2β2 - HbA, major; α2δ2 - HbA2, minor). The genes are ordered in the α- and β-gene clusters 5' to 3' in the order of transcription during development (Figure 1.4) [Efstratiadis *et al.*, 1980; Lauer *et al.*, 1980].

The proto-mammalian α-globin gene cluster is believed to have been 5' ζ-5' α-3' α-θ 3' [Higgs *et al.*, 1989]. Sequence comparisons of primates indicate that the human α1 gene is derived from the 3' α gene and that the ancestral 5' α gene has become the human pseudogene, ψα1. Thus, a further duplication event of the α1 gene occurred to produce the human α2 gene, and the two genes retain an almost identical sequence (reviewed in Higgs *et al.*, [1989]). The anciently duplicated ζ genes are ~60% homologous in sequence to the α-genes, although they possess considerably longer introns due to the presence of GC-rich repeated sequences [Proudfoot *et al.*, 1982]. The upstream ζ2 gene is functional [Pressley *et al.*, 1980] but the ψζ1 gene

Figure 1.4: Organisation of the Human Globin Gene Clusters




Top: The α globin gene cluster, located on chromosome 16, consists of functional embryonic ζ2 and duplicated adult α-genes. ψζ1 (and the converted alternative, ζ1, not shown), ψα2, ψα1 and Ro are pseudogenes. The θ gene is of unknown function (see text). Bottom: The β-globin gene cluster, located on chromosome 11, consists of foetal ε and embryonic γ genes and adult major, β, and minor, δ, genes. The ψβ gene is non-functional.

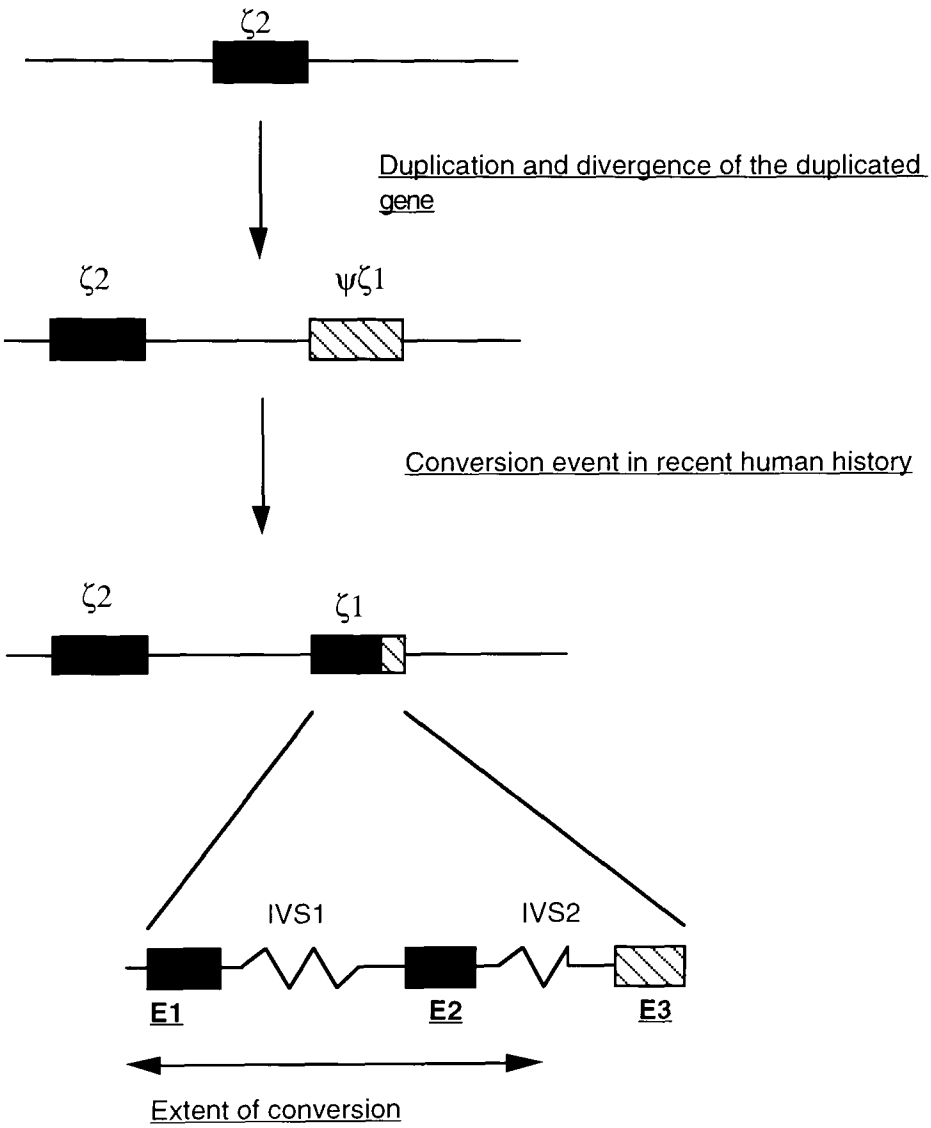
cannot produce a functional polypeptide due to a stop mutation in its coding sequence [Proudfoot *et al.*, 1982].

The ψζ1 gene is not, however, fixed in human populations, with most populations being polymorphic for a ζ1 gene, which is produced by a partial gene conversion of ψζ1 by ζ2 (Figure 1.5) [Hill *et al.*, 1985]. Although this conversion removes the inactivating mutation of ψζ1, the ζ1 gene is not transcriptionally active; this is presumably because of the absence of necessary upstream regulatory sequences (although see Proudfoot *et al.*, [1984]) or because the conversion does not include some unknown regulatory sequence in the second intron or exon three [Hill *et al.*, 1985]. In addition to ψζ1/ζ1, there are three other pseudogenes in the α-globin gene complex, two of which are related to the α-genes: the ψα1 gene described above [Proudfoot *et al.*, 1982] and the more highly divergent ψα2 gene which is positioned ~65 bp from the polyadenylation site of ψζ1 [Hardison *et al.*, 1986] (Figure 1.4). The third pseudogene, Ro, is present at the 3' end of the cluster and is related in sequence to a small cytoplasmic RNA [Jurka *et al.*, 1988]. Close to Ro is an α-like gene, θ (Figure 1.4), which has a functional sequence in primates, is transcribed and apparently maintained by selection [Hsu *et al.*, 1988; Leung *et al.*, 1987; Marks *et al.*, 1986; Shaw *et al.*, 1987] but, due to mutation, has a predicted non-globin gene product [Clegg, 1987a]. Naturally occurring deletions of θ have no detectable phenotype [Fischel-Ghodsian *et al.*, 1987], and therefore, the gene's function remains unclear.

Globin Haplotypes

DNA analysis of the α- and β-globin gene clusters has revealed substantial variation in individuals and populations [Higgs *et al.*, 1989]. This variation primarily consists of restriction enzyme polymorphisms (RFLPs) [Kan and Dozy, 1978], but length

Figure 1.5: Gene Conversion of ψζ1 Gene to Produce the ζ1 Gene




The functional α-like ζ2 gene was duplicated during mammalian evolution to produce a ψζ1 gene, which has remained largely similar in sequence to ζ2, although it has acquired a number of point mutational differences (including one which inactivates the gene) and alterations in size of repeated sequences contained within the introns. A gene conversion event partially alters the ψζ1 gene to a ζ2 sequence, forming the ζ1 gene. The conversion extends from 5' to the gene through to the repeat sequence in the second intron [Hill *et al.*, 1985].

polymorphisms and variation in gene copy number and type are also present [Higgs *et al.*, 1989]. This extensive polymorphism creates highly informative loci for gene mapping, and studies of molecular mechanisms and population genetics. Since these various polymorphisms are located close to one another in the globin gene complexes, they are in linkage disequilibrium and effectively behave as single allelic haplotypes.

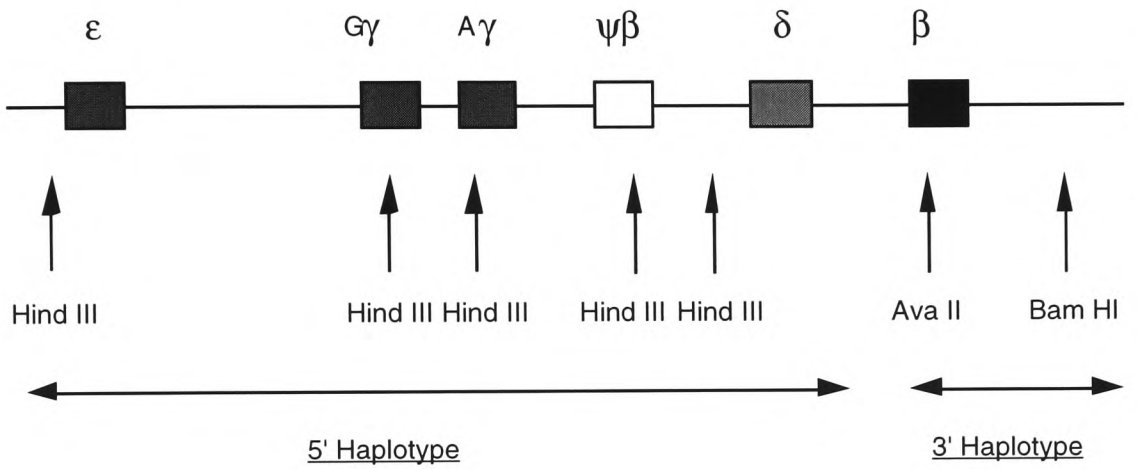
The major incentive for the development of these multilocus haplotypes has been to aid in the analysis of deleterious mutants, particularly in the β-globin gene cluster [Antonarakis *et al.*, 1985], since β-thalassaemia is more severe than most α-thalassaemias. The β-haplotype is commonly defined using seven RFLPs; recombination is not uniform across the cluster, however, with a region between the δ and β genes exhibiting unusually high levels of recombination [Chakravarti *et al.*, 1984; Treco *et al.*, 1985]. The consequence of this is that there are two groups of linked β-globin RFLPs positioned either side of the hotspot, known as the 5' and 3' haplotypes (Figure 1.6).

A systematic search for polymorphisms carried out to develop a multilocus haplotype in the α-globin gene complex revealed extensive variation, with 14 of 203 restriction enzyme sites (1132 bp) studied being polymorphic [Higgs *et al.*, 1986]. Only common polymorphisms (with a population frequency greater than 0.05) were used to construct the α-haplotype, but a number of other less frequent, population-specific polymorphisms were identified (Figure 1.7). Of a total of nine common polymorphisms, eight are dimorphic and one is trimorphic; seven of the dimorphic sites are simple RFLPs, with one (PZ/Z) caused by the gene conversion event of the ψζ1 gene discussed earlier (Figure 1.5). This latter dimorphism is related to a change in size of the converted gene because of differences in the lengths of the repeat sequences contained within the introns of the ζ genes. The trimorphic

polymorphism (IZHVR; Figure 1.7) is also caused by differences in length of a repeated sequence, and is classified into small, medium and large alleles. The α-globin cluster VNTR loci, by being part of the α-haplotype, clearly do not behave as recombination hotspots [Higgs *et al.*, 1986].

---

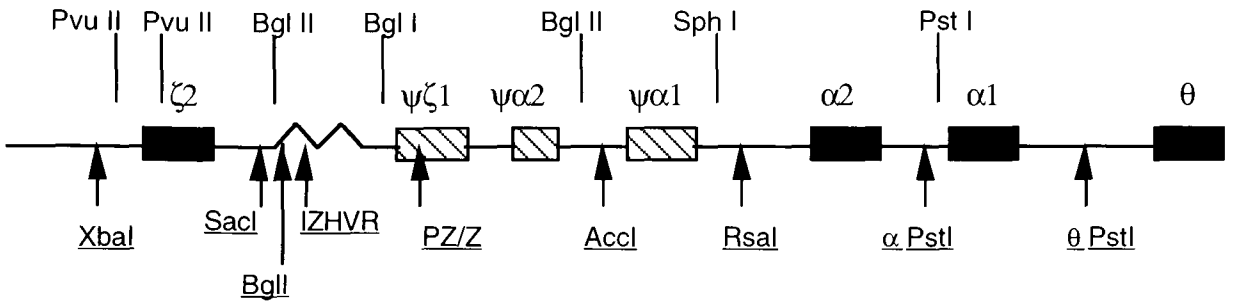
Figure 1.6: β-Globin Haplotypes




---

The inclusion of length variations in the α-haplotype increases the heterozygosity and informativeness relative to the β-haplotype, but also increases its complexity [Higgs *et al.*, 1986]. Unlike the β-globin gene complex, recombination is apparently uniform across the α-complex; of a possible 768 ($2^8$x3) haplotypes, only ~ 60 have been found so far in world populations (A.J.Boyce, personal communication), indicating linkage disequilibrium between all haplotyping sites. α-haplotypes are divided, relatively arbitrarily, into groups (I-VIII) according to the pattern of sites at the 3' end of the gene cluster: PZ/Z, *Acc I*, *Rsa I*, α *Pst I* and θ *Pst I* [Higgs *et al.*, 1986](Figure 1.7). Subtypes are named according to the pattern of the 5' sites (*Xba I*, *Sac I*, *Bgl I*, IZHVR; Figure 1.7), and are denoted by a letter following the haplotype group; for example, the Ia is a subtype of the group I haplotypes (Figure 1.7).

Figure 1.7: α-Globin Haplotypes




| | Polymorphic Site | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Haplotype | Xba I | Sac I | Bgl I | IZHVR | PZ/Z | Acc I | Rsa I | α Pst I | θ Pst I |
| Ia | + | + | - | M | PZ | + | + | - | - |
| IIa | - | + | - | L | PZ | + | - | - | - |
| IIIa | - | - | + | M | Z | - | - | - | - |
| IVa | + | - | - | S | PZ | + | - | + | + |
| Vc | + | - | - | S | PZ | + | - | - | + |
| VIa | - | + | - | L | PZ | - | - | - | - |
| VIIa | - | + | - | M | Z | + | - | - | - |
| VIIIa | + | - | - | S | PZ | + | - | + | - |

The α-globin gene cluster contains nine commonly polymorphic sites; the IZHVR is a tri-allelic length polymorphism, being classified into size classes small (S), medium (M) and large (L). The remaining sites are dimorphic, with all, apart from the PZ/Z site which is caused by length variations of repeat sequences (see Figure 1.6), being RFLPs [Higgs *et al.*, 1986]. Above the cluster are shown RFLPs which have been identified as being rarely polymorphic, e.g. the *Pst I* site 5' to α1 is found only in African populations and only at low frequency [Higgs *et al.*, 1986]. Haplotypes form eight main groups on the basis of 3' sites and are named according to the pattern of 5' sites. The most common haplotypes from each group are shown here [Higgs *et al.*, 1986; Martinson *et al.*, 1995; O'Shaughnessy *et al.*, 1990; Ramsay and Jenkins, 1988].

α-haplotypes contain RFLPs which reflect the process of point mutation, but, as discussed previously, population genetic forces (drift and selection) also operate on mutational variation, and recombination can shuffle point mutations into different patterns. The observed patterns of variation will therefore depend on a complex interplay of all these forces; this interplay is illustrated by another well-characterised locus, the major histocompatibility locus (HLA in humans) which demonstrates extremely high allelic diversity. This diversity could be due either to diversifying selection (e.g., as protection against parasites) or to an unusually high rate of point mutation or recombination. It has been estimated that the rate of recombination is very high relative to point mutation [Steinmetz *et al.*, 1986; Zangenberg *et al.*, 1995], with recombination reassorting a number of internally linked regions, but it is clear that both selection and recombination operate together at this locus [Erlich and Gyllensten, 1991; Martin *et al.*, 1995; Ohta, 1995].

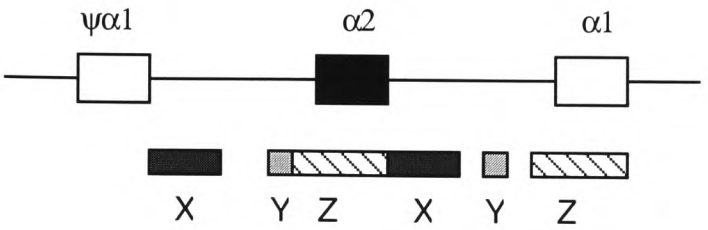
As with β-globin haplotypes, most human populations have only a few common α-haplotypes, some of which are common in many populations, and a small number of rarer haplotypes [Higgs *et al.*, 1986]. In the original survey of haplotype variation in nine populations, five high frequency haplotypes were described [Higgs *et al.*, 1986]. These α-haplotypes are not clearly related to one another by single point mutations, although minor (rarer) haplotypes may have arisen by recombination between the more common haplotypes [Higgs *et al.*, 1986]. Therefore, although there is no hotspot for recombination in the α-cluster, recombination clearly does influence variation. In addition, the regions of the haplotype consisting of length variation are also likely to be subject to recombination rather than simple point mutation.

Both the functionally specialised gene pairs, ζ and α, show concerted evolution; that is, the tandem genes within a species are more homogeneous than they are between species. Human allelic α2 and α1 genes have maintained almost perfect homology over a ~1.5 kb region, including the 5' flanking region of human α-genes. Likewise, despite the non-functionality of the ψζ1 gene, it has only six base pair changes compared to ζ2 [Proudfoot *et al.*, 1982]. The patterns of sequence correction differ between species, however: homology is maintained only in the regions required for gene expression in the goat duplicated α-genes [Schon *et al.*, 1982], whereas in the chicken, the two α-genes, $\alpha^D$ and $\alpha^A$, are divergent [Dodgson *et al.*, 1981]. When concerted evolution is observed, it is due to recombinogenic activity between the duplicated genes, which facilitates the fixation of variation and changes gene frequencies in a non-Mendelian fashion [Dover, 1982; Dover, 1986].

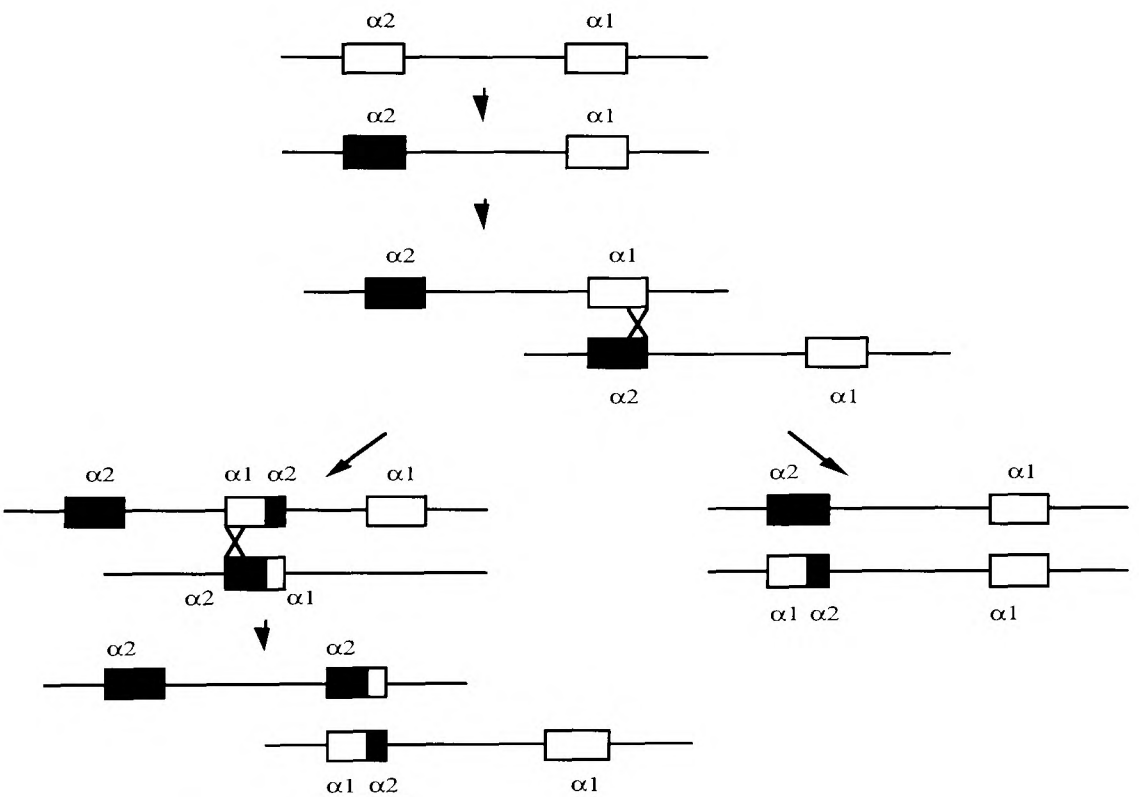
Following duplication, genes will be highly homologous and therefore likely to undergo both crossover and conversion events. Evidence for crossover comes from the α-thalassaemias: whereas most β-thalassaemia is caused by point mutation, α-thalassaemia is mostly due to deletions [Higgs *et al.*, 1989]. $\alpha^+$ thalassaemia, in which the production of α-globin is reduced, is usually due to the loss of one α-gene (to form the -α chromosome) caused by a number of different crossover events, and facilitated by sequence homology. There are three homologous regions, separated by small regions of non-homology, surrounding and containing the α-genes: the X, Y and Z boxes (Figure 1.8) [Higgs *et al.*, 1984; Lauer *et al.*, 1980], which can misalign between α2 and α1. Subsequent crossover will result in -α and ααα chromosomes (Figure 1.8); the -α deletion can occur by misalignment of the Z sequences, which are 3.7 kb apart, to form the $-\alpha^{3.7}$ deletion; subdivision of this deletion (into types I, II and III) is made on the basis of where in the Z box the crossover event takes place

Figure 1.8: The Concerted Evolution of the Human α-Globin Genes

A) α-Gene Homology Boxes




B) Recombination between α-genes facilitated by Homology




A) The α-genes are contained within regions of homology known as the X, Y and Z boxes. B) Duplication of the ancestral α-gene produced two genes which undergo sequence correction. Misalignment allows exchange between the genes: as gene conversion (right) in which gene number is unaltered, or as reciprocal exchange (left) which alters gene number. The first round of the latter exchange produces an α-gene deletion chromosome and a triplicated α-gene chromosome. These represent intermediates for a second crossover, which restores gene number to two. Both conversion and crossover will result in concerted evolution, if the products become fixed in the population.

[Higgs *et al.*, 1984]. Misalignment of the X boxes and crossover is also observed in human populations, to result in the $-\alpha^{4.2}$ deletion [Higgs *et al.*, 1989]. From global surveys of α-globin gene variants, it has been observed that the frequency of the -α chromosome is elevated by selection in malarious regions [Flint *et al.*, 1986]; the frequency of the ααα chromosome is also quite high in some populations suggesting that it also may be selected.

The -α and ααα chromosomes may act as intermediates for further crossover events, as evidenced from chromosomes with multiple copies of α-genes, with apparently normal phenotypes, in human populations [Higgs *et al.*, 1989]. It has been suggested that crossover between -α and ααα chromosomes, resulting in a chromosome with two genes (which would be mosaics of α1 and α2) could be a mechanism of maintaining sequence homology (Figure 1.8) [Higgs *et al.*, 1984]. Similar intermediates (-ζ and ζζζ) are seen at the ζ-gene pair suggesting that the same process, known as crossover fixation, could also be responsible for the homology seen between the two human ζ genes. Given these crossover events, it is likely that gene conversion also occurs, as suggested in the primate haptoglobin gene cluster [Erickson and Maeda, 1994], in other gene families [Baltimore, 1981] and, from comparative sequencing studies, between the γ genes [Powers *et al.*, 1984; Powers and Smithies, 1986; Slightom *et al.*, 1980] and ζ genes [Hill *et al.*, 1985](Figure 1.5). The likely role for conversion in sequence homogenisation illustrates that the process is not just responsible for sequence diversification, as emphasised earlier at the HLA loci [Erlich and Gyllensten, 1991].

The α-genes show preferential accumulation of sequence differences in the second intron and 3' flanking region (that is, in the Z box sequence; Figure 1.8); it has therefore been suggested that the presence of a seven base pair insertion/deletion in the second intron of α2 and α1 inhibits branch migration and thus gene conversion

beyond this point [Hess *et al.*, 1984]. Although this explanation would account for the observed polarity, the fact that crossovers do occur beyond the second intron [Higgs *et al.*, 1984] questions its validity. The actual explanation for the sequence polarity observed in the homology blocks (including the similar polarity in sequence divergence observed seen across the X boxes [Hess *et al.*, 1984]) may be related to the presence of *Alu* repeats at the boundaries of the blocks with non-homologous sequence [Hess *et al.*, 1983; Sawada and Schmid, 1986], and which may disrupt conversion (see Murti *et al.*, [1992]). In fact, although most recombination is likely to occur between the homologous α-genes, deletions are known to occur which remove both α-genes (causing $\alpha^o$-thalassaemia), and which appear often to be the result of non-homologous recombination between *Alu* sequences [Nicholls *et al.*, 1987]. Many of the end-points of these deletions occur between ζ2 and IZHVR, and between the α1 gene and a repetitive sequence, the 3' HVR [Nicholls *et al.*, 1987], suggesting that regions of the cluster, other than the α-genes, are also prone to recombination events.

**Globin Genes and Human Population Affinities**

Information on global distributions of globin RFLP haplotype variation for the purposes of disease studies, also allows questions of human evolution to be addressed. For anthropologists and population geneticists, an accomplishment equivalent to an entire human genome sequence would be a definitive human population tree[5], including branching order, timing and geography. Some progress towards this aim has been made. An early study of protein data suggested a primary division between African and all other populations [Nei and Roychoudhury, 1982], and this was confirmed with the population phylogeny constructed using 5' β-globin

[5] Or even better, a gene genealogy or representation of the relationships between alleles of a gene, since we would not have to assume that our definition of a "population" is meaningful (that is, a biologically defined, interbreeding related group) , as required in a population phylogeny.

haplotypes [Wainscoat *et al.*, 1986] - the first study using a nuclear DNA locus. Nuclear loci are more appropriate for human evolution studies since, unlike the recently favoured mtDNA locus, they represent male and female genetic contributions.

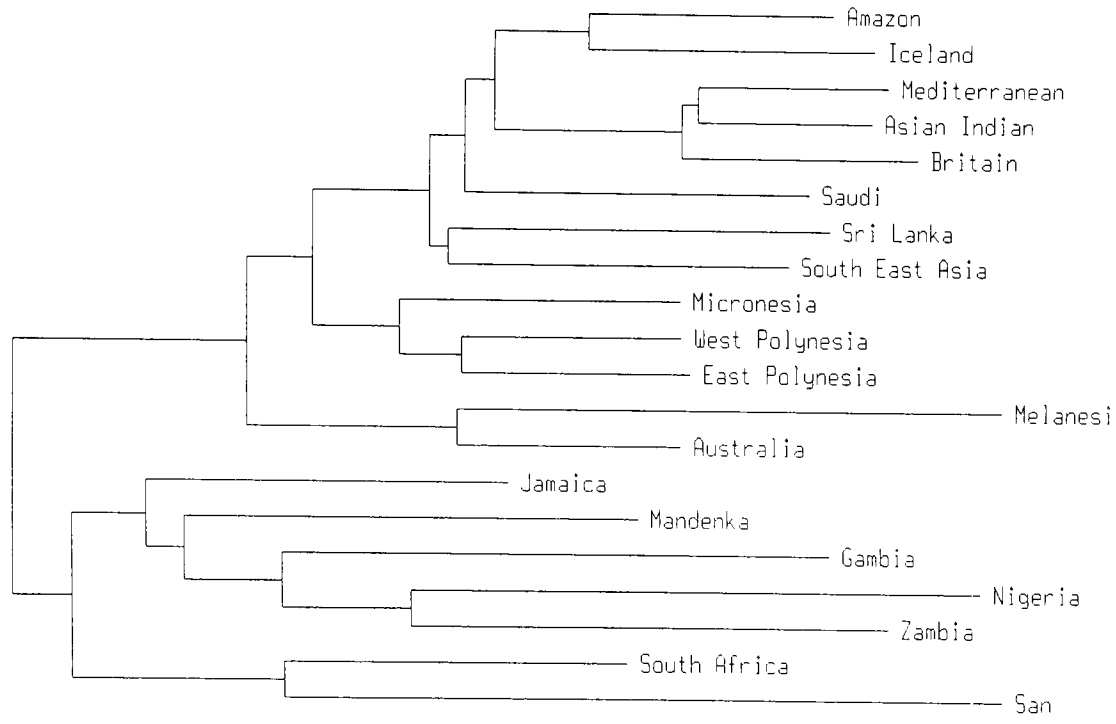
African populations possess predominately one β-haplotype, which is not found elsewhere [Wainscoat *et al.*, 1986]. This uniqueness, coupled with greater diversity in sub-Saharan African populations is consistent with the suggestion, from fossil data, that humans originated in Africa and migrated to the rest of the world from there [Stringer and Andrews, 1988]. Further studies at the DNA level, using a variety of loci, are confirming these conclusions, and detailing some of the specific migrations that have taken place during human history; for examples, β-globin haplotypes and sequences [Fullerton, 1995; Wainscoat *et al.*, 1986], microsatellites [Goldstein *et al.*, 1995], an *Alu* insertion polymorphism [Batzer *et al.*, 1994], numerous nuclear loci [Bowcock *et al.*, 1991; Cavalli-Sforza *et al.*, 1988; Deka *et al.*, 1995; Excoffier *et al.*, 1987; Martinson *et al.*, 1995; Nei and Roychoudhury, 1993; Olerup *et al.*, 1988; Ramsay and Jenkins, 1988; Wainscoat *et al.*, 1986], Y chromosome [Jobling and Tyler-Smith, 1995; Spurdle *et al.*, 1994] and mitochondrial DNA [Cann *et al.*, 1987; Chen *et al.*, 1995; Horai *et al.*, 1995] . From a population genetics perspective, if selective neutrality of these systems, regularity of genetic change and consistency of population size are assumed, this greater diversity can be considered to represent the greater age of African populations.

One of the best examples of applying genetics to questions of human migration and history comes from the Pacific. Since there have been clear waves of expansion indicated by a wealth of archaeological and linguistic data, the Pacific represents a good testing ground for the techniques required for the determination of a human evolutionary tree. When the correlation between the frequency of α-thalassaemia

and malarial incidence was noted [Flint *et al.*, 1986], it was clear that in Melanesia all α-thalassaemia deletions were of one type: $-\alpha^{3.7III}$, a deletion not found in other populations. The deletion is selected up to high frequency on a haplotype which is characteristic of the region, the IIIa, indicating that the deletion has arisen locally. The consequence of this is that it allows studies of the movements of peoples across the Pacific. For example, archaeological and linguistic data indicate that migrants from Southeast Asia crossed the Pacific into Polynesia ~1000-2000 BC [Bellwood, 1989], but there has been debate over whether they made contact with Melanesians on the way. Polynesia has no malaria but populations possess the Melanesian $-\alpha^{3.7III}$ deletion, which is evidence of interbreeding between Melanesians and the proto-Polynesians. Further, Fiji, at the boundary of Polynesia and Melanesia, and possibly representing a region from which Polynesians colonised the rest of the eastern Pacific, has a mixture of α-haplotypes from Southeast Asia and Melanesia [Trent *et al.*, 1988]. Other globin gene variants indicate that the Polynesians probably also had contact with Micronesians as they migrated eastwards [O'Shaughnessy *et al.*, 1990].

Thus, α-haplotypes and gene copy variants have proven useful for investigating the recent history of the Pacific. Recent studies are indicating that more ancient migration events and affinities can be addressed using α-haplotypes [Martinson *et al.*, 1995; Ramsay and Jenkins, 1988; Roberts-Thomson *et al.*, in press; Zago *et al.*, 1995]. A larger data set of α-haplotypes in world populations is being compiled in this laboratory, so that ultimately a comprehensive population phylogeny will be possible at this locus. Initial results indicate that the α-haplotype data can produce trees which suggest plausible relationships (Figure 1.9) - in comparison with those trees drawn from variation at other loci, and from what is known from studies of fossils, linguistics and archaeology.

Figure 1.9: Population Tree Based on α-Haplotypes




Assuming that haplotypes with the same RFLP cutting pattern are equivalent in all populations, this tree can be drawn (A J Boyce, personal communication) using the neighbor-joining method of Saitou and Nei, [1987].

However, there are some potential problems with the use of haplotype data in this way. For example, the ancestral β-haplotype is found only in Africa, and presumably has been lost in other world populations; non-Africans possess haplotypes which are either rare in Africa or appear to be recombinants between other haplotypes [Wainscoat *et al.*, 1986]. If Africa is to be postulated as the source population for all other human variation, it is expected that those haplotypes outside of Africa should be mutationally related to African haplotypes, perhaps by recombination [Giles and Ambrose, 1986]. The only possible way to determine whether this is the case (and thus to speculate about human evolution) is to have sequence information related to haplotypes. The same requirement applies to the phylogenies to be made with α-haplotypes; such trees are built on the assumption that a particular haplotype is the "same" in all populations in which it is found and therefore is an accurate predictor of relationships.

It is not clear whether this assumption is correct; (-) alleles at RFLP sites may not necessarily possess the same underlying sequence, since a (+) site could be lost by any number of mutations, including insertion/deletions and point mutation at any of the sites in the enzyme recognition sequence. The repetitive sequence α-haplotyping sites (IZHVR and PZ/Z) also may not be equivalent between populations, especially given that such repeated sequences are known to show extensive size variation, as indicated earlier. The only way to determine whether the assumptions made in forming population phylogenies are correct, is to have sequence information sufficient to determine whether shared haplotypes in populations indicates affinity or convergence. Without such a certainty of relationship, the confidence in the phylogeny produced from these highly informative RFLP haplotypes is reduced and if unexpected relationships are suggested between populations (as does occur, A. J. Boyce, personal communication; Figure 1.9 suggests an unexpectedly close relationship between Icelandics and

Amazonian Indians), it is impossible to determine whether this is due to violation of the original assumptions or due to the way in which the phylogeny has been constructed.

## The Questions to Be Addressed

Depending on one's perspective, the resolution of population phylogeny represents the conclusion of studies of human genetic variation geared towards the discovery of the underlying causes of genetic disease. There can be no doubt that both medical and population genetics have much to gain from each other, and the interests involved in the study of the amount, consequences and maintenance of genetic variation are diverse. The aims of this study encompass many of the varied questions that originate from the different fields of biological investigation that have been outlined so far. As a starting point, however, the question is quite specific: can we assume that α-haplotypes are equivalent between different populations for the purposes of phylogenetic study?

In order to answer this question a detailed understanding is required of a number of fundamental issues concerned with the nature of, and forces operating on, variation at the sequence level in the α-globin gene complex. Therefore, this study aims to discover the quantitative and qualitative aspects of the sequence variation, its production and how it relates to α-haplotypes; specifically, the questions here are: are (-) RFLP alleles always the same at the sequence level? Are RFLP haplotypes representative of the underlying sequence variation at this locus? How do subtypes arise and how are the major haplotypes related? An understanding of the molecular mechanisms responsible for producing variation is perhaps one of the most important aspects of this study since it will ultimately allow a more comprehensive and informative genetic tree to be produced.

The fact that sequence variation is so well-characterised at the β-globin locus (for a recent study see [Fullerton *et al.*, 1994]), means that extremely informative comparisons are possible with the α-globin cluster. Since the globin complexes are situated in such different compartments of the human genome, an understanding of variation at the DNA sequence level can be related to broader characteristics such as replication timing, sequence composition, and recombination rates. The fact that the α-globin gene complex does not contain a hotspot, as the β–globin gene cluster does, will also help to gain an understanding of the consequences of linkage disequilibrium, and how recombination operates within "cold", small (<100 kb) regions. Comparison between the globin gene clusters is thus potentially very informative in providing answers to questions of how different regions of the genome vary in terms of polymorphism, and given the different properties of their genomic environments, how these differences may be related to mechanism.

# Chapter 2

## The α-Globin PCR Haplotype and a Study Of African Diversity

### Haplotyping by PCR

α-globin haplotyping to date has been performed by restriction enzyme digestion, Southern blotting and hybridisation with a number of α-globin cluster specific probes [Higgs *et al.*, 1986](Appendix 1). This method, although reliable, requires a large amount of DNA: in the region of 60 μg or more. This represents the method's main limitation, and is not a trivial one given that the population samples which would be of greatest interest are often available in only small quantities. In many cases, the DNA is also of inferior quality and this can make it unsuitable for Southern blotting. In order to extend the survey of world diversity at this locus it could be argued that it is crucial to be able to reduce the amount of starting DNA. The obvious way to do this is to produce an alternative methodology based upon PCR.

PCR haplotyping also has a number of other advantages: the method is faster, non-radioactive, less labour intensive, produces less ambiguous results and, as discussed in following chapters, there is an opportunity to provide a finer resolution of haplotypes: both at the hypervariable loci contained within the haplotype and at the sequence level. The usefulness and applicability of the α-globin haplotype is thus increased considerably.

### Development of the PCR-Haplotype

The polymerase chain reaction (PCR) [Saiki *et al.*, 1985] as an *in vitro* method of synthesising a particular segment of nucleic acid, has had a dramatic impact on

molecular biology. Enhancements of the method, which began most notably with the use of thermostable DNA polymerases [Saiki *et al.*, 1988], have fostered the development of a number of diverse PCR applications [Arnheim and Erlich, 1992; Erlich, 1989; Innis *et al.*, 1990].

The PCR consists of two oligonucleotide primers that hybridise to opposite strands and flank the region of interest in the target DNA. A number of cycles involving heat denaturation of the template, annealing of the primer and extension of the annealed primers by a DNA polymerase result in the accumulation of the specific fragment of interest. The accumulation is exponential (until the later stages of the reaction when a "plateau" stage is reached) since the products from one extension can serve as template during the next cycle.

There is a "standard" protocol for PCR on which most applications are based [Innis and Gelfand, 1990; Saiki, 1989; Saiki, 1990]. However, although the principle of PCR is straightforward, the reaction is relatively complicated and incompletely understood, and no single set of conditions works in every application. Modifications of the "standard' protocol are usually necessary and this optimisation process is of particular importance when the application is to be of a repetitive and analytical nature. The precise alterations that have to be made are influenced by a number of factors, including the primer sequences used and sequence characteristics of the region to be amplified. To some extent, the optimisation process is empirical and this is particularly true when dealing with a non-standard application, such as the amplification of the GC-rich α-globin cluster.

**The Problem of GC-Rich Templates**

Incomplete denaturation of template is thought to be a major component in many PCR failures [Innis *et al.*, 1990], and particularly for GC-rich amplifications. The

melting temperature (Tm) of DNA is defined as the temperature at which half of the DNA is denatured or single-stranded, and this temperature is related to the sequence composition of the DNA. In the DNA molecule, guanine (G) binds to cytosine (C) by three hydrogen bonds, whereas adenine (A) binds to thymine (T) by only two hydrogen bonds; consequently breaking GC bonds requires higher temperatures. GC bonds stabilise secondary structure within the template and this can inhibit access to the template by the primers and extension of the polymerase along the template. GC-rich amplification protocols, including the protocol described in this study, are all primarily concerned with overcoming these particular problems.

**Modifications of Protocol for Efficient Amplification**

The aim was to develop a highly reproducible method that uses little DNA. Some of the modifications made and described here were essential not only for specific amplification, but also for optimising the yield, which was important for further applications of the protocol (particularly direct sequencing, Chapter 3).

DMSO Buffer

Commercially available, standard amplification buffers (e.g. as supplied by Cetus), did not result in a visible product with any primer set used in this study. Use of a buffer containing the denaturant dimethylsulfoxide (DMSO) proved necessary, as shown in amplification systems designed to type α deletion chromosomes [Bowden *et al.*, 1992; Bowden *et al.*, 1992; Bowden *et al.*, 1992; Baysal and Huisman, 1994]. DMSO has been found to aid amplifications of other GC-rich DNA [Bookstein *et al.*, 1990], extension through regions of secondary structure [Shen and Hohn, 1992] and is used in long PCR [Cheng *et al.*, 1994]. It is believed to alter the degree of product strand separation at a particular denaturation temperature [Innis *et al.*, 1990;

Varadaraj and Skinner, 1994]. DMSO is toxic, so glycerol has been suggested as a non-toxic alternative to DMSO [Lu and Negre, 1993]. However, in experiments in this study, replacing DMSO with glycerol did not result in visible products. The DMSO buffer used is shown in Table 2.1 and is based upon that used by Dutton *et al.*, [1993], Dutton *et al.*, [1993], Jeffreys *et al.*, [1988b], Baysal and Huisman, [1994], Bowden *et al.*, [1992] and Martinson, [1991].

---

Table 2.1: α-Globin PCR Haplotype Buffer Composition

| | Final Concentration |
|---|---|
| Tris-Cl(pH 8.8)* | 66.7 mM |
| $(NH_4)_2SO_4$ | 16.7 mM |
| $MgCl_2$ | 1.0-3.0 mM |
| $Na_2EDTA$(pH 8.0)* | 66.7 μM |
| Gelatin | 0.01% |
| DMSO | 10.0% |
| β-mercaptoethanol | 140 mM |

* pH at 20°C. Final concentration refers to a 100μl PCR volume.

---

Vent Polymerase as an alternative to Taq Polymerase

One problem of using DMSO is that it is inhibitory to Taq polymerase, the most commonly used thermostable polymerase (isolated from *Thermus aquaticus*). 10% DMSO reduces Taq polymerase activity by 50% at 70°C [Gelfand, 1989]. An inhibitory effect on Taq was observed in this study. Therefore, amplifications were routinely performed with Vent (New England Biolabs), which is a highly thermophilic polymerase isolated from *Thermococcus litoralis*, and which amplified more efficiently than Taq in the DMSO buffer. Since Vent polymerase is considerably more thermostable than Taq (with a half-life of 6.7 hours at 95°C versus 40 mins for Taq), higher temperatures (95-98°C) and longer denaturation steps (> 1 min for 30-40 cycles) could be used in the PCR without dramatically

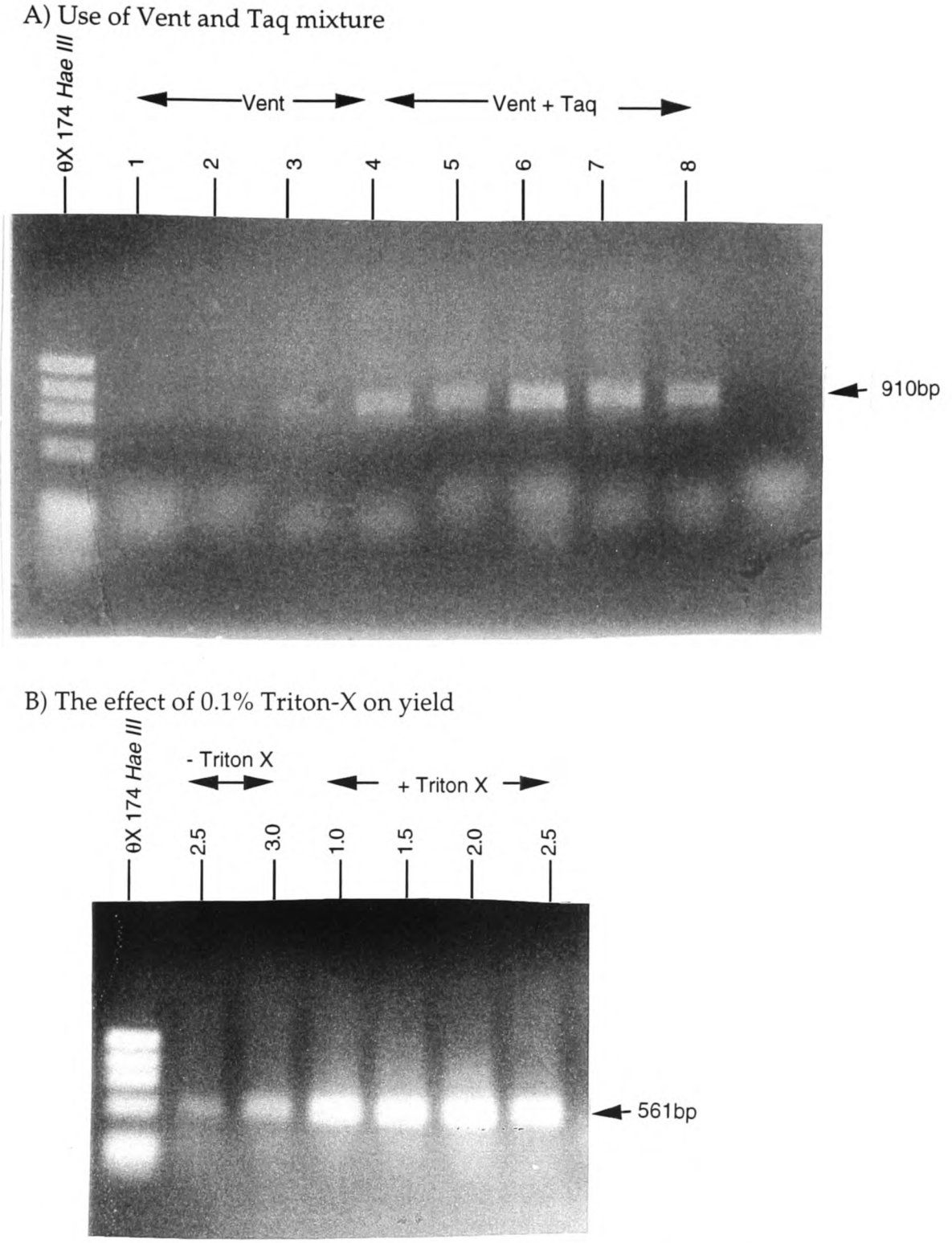
affecting enzyme activity. This allowed more efficient denaturation of the GC-rich template, which was facilitated further by the use of thin-walled reaction tubes.

Vent, unlike Taq, possesses a 3'-5' exonuclease or "proof-reading" activity which results in a higher fidelity (a property that has been exploited in the design of long PCR [Cheng *et al.*, 1994]). However, this activity necessitated the careful optimisation of reactions to ensure that 3' degradation of primers or product did not occur significantly during the PCRs. For example, the time of extension at 72°C was limited to as close to the manufacturer's recommended 1 kb $min^{-1}$ as possible. Exonuclease activity is enhanced by the presence of excess dNTPs, so the amount of dNTPs (Pharmacia) was used only within a range of 200-400 mM (each dNTP). It should be noted that Vent is relatively sensitive to changes in reaction conditions and since the buffer used contained reagents which may deteriorate (e.g. the reducing agent, β-mercaptoethanol) it was important maintain fresh stocks of buffer and component reagents. A modified form of Vent, which has the 3'-5' exonuclease activity removed, called Vent (exo-) is reported to be less sensitive to changes in reaction conditions (New England Biolabs), and therefore is a possible alternative. It was also found that if amplifications with Vent were sub-optimal, a mixture of Taq and Vent polymerases in the reaction mixture sometimes worked more efficiently (Figure 2.1a).

Primer sequences

Target regions of ~1 kb were chosen around each of the haplotyping sites (seven RFLP, two HVRs), with the RFLP sites in a position within the chosen PCR product such that, upon digestion, fragments would be easily visualised upon agarose gels. Suitable primers (20-30 bp) were designed using available sequence ([Proudfoot *et al.*, 1982; Goodbourn *et al.*, 1983]; D R Higgs, personal communication; J Flint,

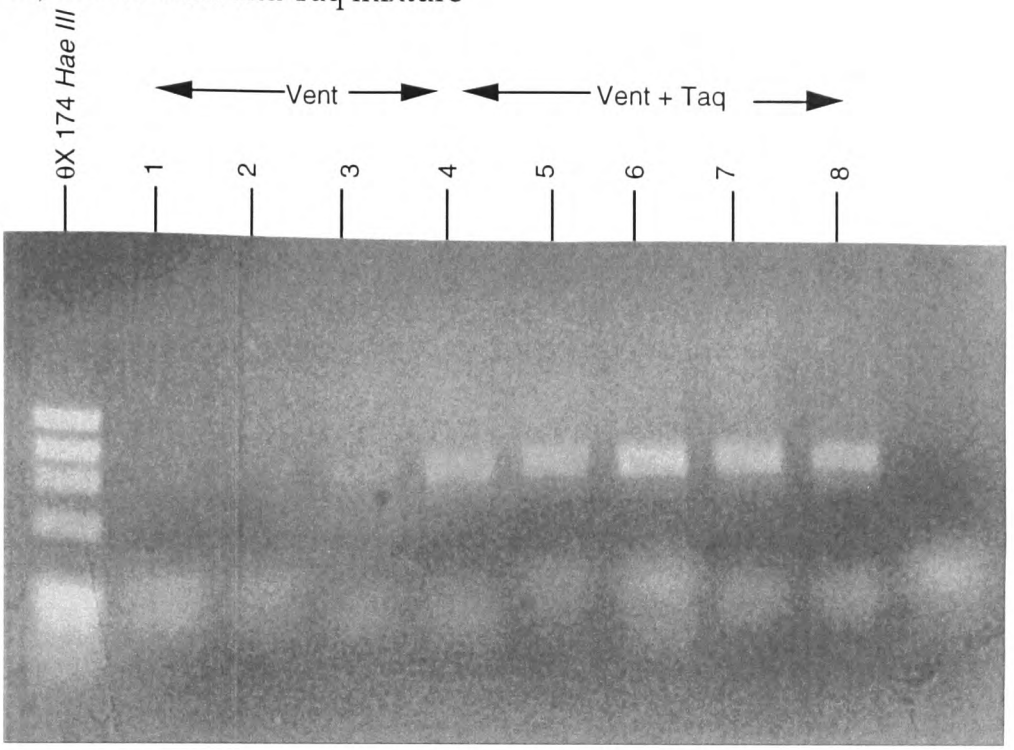
Figure 2.1: Strategies in the Amplification of the α-Globin Cluster



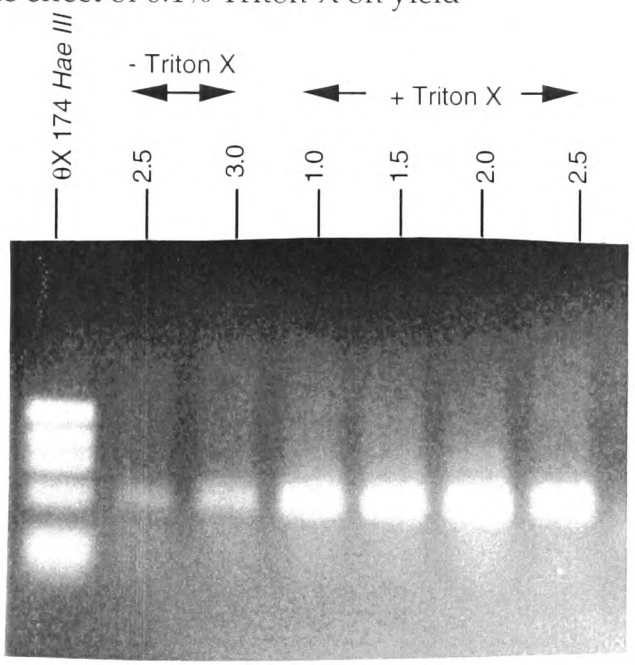

A) Sub-optimal Xba PCR (Table 2.2): reactions 1-4 are duplicates using only Vent polymerase; reliability is improved using Vent/Taq mixture (0.5 U of each) in duplicate reactions 5-8. B) Sac PCR yield increased with Triton-X, and over a wider range of buffer [$Mg^{2+}$] (concentrations shown in mM). Both gels are 1.5% agarose.

Figure 2.1: Strategies in the Amplification of the α-Globin Cluster

A) Use of Vent and Taq mixture




B) The effect of 0.1% Triton-X on yield




A) Sub-optimal Xba PCR (Table 2.2): reactions 1-4 are duplicates using only Vent polymerase; reliability is improved using Vent/Taq mixture (0.5 U of each) in duplicate reactions 5-8. B) Sac PCR yield increased with Triton-X, and over a wider range of buffer [$Mg^{2+}$] (concentrations shown in mM). Both gels are 1.5% agarose.

personal communication) and with the aid of the computer program, PRIMER (GCG), which allowed accurate primer Tm (which were designed to be 55-70°C) and levels of complementarity (both self-, and between 3' ends of the primer pair) to be calculated. The identity of products was confirmed by amplifying and digesting, and directly sequencing, samples of known RFLP status (from previous studies using Southern blotting).

Alterations to improve yield of amplifications

Amplifications worked with only 20-30 pmoles of each primer, albeit rather erratically, but 50 pmoles of each primer per 100 µl reaction were generally used because of enhanced yields. With each set of primers designed, the $Mg^{2+}$ concentration in the DMSO buffer, which affects specificity and yield, and details of the cycling conditions, were optimised. Initially, the cycling conditions that were attempted were excessively long in order to overcome the inefficiencies of the PCR; for example, long denaturation times (much longer than 1 minute per kb) were used. However, given the optimisations listed above, efficient amplification was possible using shorter times for denaturation, annealing and extension steps, although two initial cycles of longer denaturation, annealing (at a temperature 1°C lower than that used for the rest of the reaction) and extension often increased the end yield. Yield was also enhanced by the use of Triton-X (Figure 2.1b). Triton-X and other non-ionic detergents have been reported to aid PCR efficiency in a number of studies [Dutton *et al.*, 1993; Innis *et al.*, 1990; Varadaraj and Skinner, 1994], possibly by regulating the Tm values of primers or template by reducing secondary structure [Varadaraj and Skinner, 1994].

Specific problems of amplifying minisatellite loci

VNTR repeats tend to form extensive secondary structure, such as hairpin structures, so that they present particular problems for the PCR. The α-haplotype includes two minisatellite loci (the IZHVR and PZ/Z); the IZHVR possesses large repeat arrays, with the largest alleles up to ~2 kb, which is an additional challenge to PCR, as noted for other large HVRs [Jeffreys *et al.*, 1988]. An initial primer pair positioned close to the HVR amplified different sized alleles with different efficiency depending on the conditions used. Trials compared the usual DMSO buffer (Table 2.1) with a buffer containing no DMSO but large amounts of BSA [Jeffreys *et al.*, 1988b; Martinson, 1991]. Amplification (although only of medium sized alleles) was most efficient in the DMSO buffer and not the BSA buffer, especially when the gene 32 protein (gp32) of phage T4 was used; gp32 is a thermostable single stranded DNA binding protein which can inhibit secondary structure formation and aid amplifications of longer PCR products [Schwarz *et al.*, 1990]. Primers positioned further away from the HVR array, allowed amplification of all S, M, L alleles (refer to Chapter 1). It was observed that larger alleles amplified slightly less efficiently than smaller alleles, with the effect most evident in samples heterozygous for two differently sized alleles [Arnheim and Erlich, 1992; Jeffreys *et al.*, 1988b], but the effect was not problematic for the purposes of haplotyping.

**Determination of α-Globin Haplotypes by PCR**

The primers and optimal cycling conditions used for α-haplotyping by PCR are summarised in Table 2.2. All reactions utilise the buffer as described above and in Table 2.1. Each amplification is named after the haplotyping site which it contains. To type the nine sites in the α-haplotype, eight amplifications are necessary (two of the nine haplotyping sites, IZHVR and *Bgl I*, are contained within the same PCR

Table 2.2: PCR Haplotyping Protocol

| Site | Primer Sequence (5'-3') | Position in cluster | Stage | Denaturation | Annealing (Minutes) | Extension | Volume | Amt DNA | Buffer [Mg2+] | dNTPs | Additional components (final concentration) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Xba I | XF2 : CTCAACCAGATGCCAGCCAAACAGAC<br>XR2 : GAACCATTACACCAAGTGAAGGACG | N/A | 1<br>2<br>3 | 95 : 3.5<br>95 : 1.5<br>95 : 1.5 | 61 : 2<br>62 : 1<br>62 : 1 | 72 : 3.5<br>72 : 3<br>72 : 5 | 100 | 50 ng + | 3.0 mM | 200 | 24 ng BSA |
| Sac I | SF2 : AAGATGGAGAAGGAAGGATGACTGC<br>SR5 : GAGAGCAAGCCGATGTGGTAGAATG | 5935<br>6496 | 1<br>2<br>3 | 98 : 2.5<br>98 : 1.5<br>98 : 1.5 | 59 : 1.5<br>60 : 1<br>60 : 1 | 72 : 1<br>72 : 1<br>72 : 5 | 100 | 50 ng + | 2.0 mM | 200 | 0.1% Triton-X<br>24 ng BSA |
| IZ/Bgl I | IZF5 : CCAGGACACAGATGGAGGCTAATGT<br>IZR5 : TTGGAGTTGGGTGCATGAGGC | 7570<br>9183 | 1<br>2<br>3 | 95 : 3.5<br>95 : 1.5<br>95 : 1.5 | 61 : 2<br>62 : 1<br>62 : 1 | 72 : 3<br>72 : 3<br>72 : 5 | 50 | 500 ng | 1.5 mM | 400 | 1 microgram GP32<br>24 ng BSA<br>0.1% Triton-X |
| Pseudozeta/ Zeta | PZF1 : GGGTGGGCAGAAGGGGAGAC<br>PZR7 : CAGATGCTGGGACTTTAGAGGCTGT | 11866<br>1339/<br>12996 | 1<br>2<br>3 | 98 : 2.5<br>98 : 1.5<br>98 : 1.5 | 61 : 1<br>62 : 1<br>62 : 1 | 72 : 2<br>72 : 2<br>72 : 5 | 100 | 50-500 ng | 1.0 mM | 200 | 24 ng BSA<br>0.1% Triton-X |
| Acc I | AF1 : TTTTAATAGAGACACAGTTTCACCATG<br>AR1 : CTTGATTGTTTCTTGATCTGG | 17092<br>17917 | 1<br>2<br>3 | 98 : 2.5<br>98 : 2<br>98 : 2 | 54 : 2<br>55 : 1.5<br>55 : 1.5 | 72 : 3<br>72 : 2<br>72 : 5 | 100 | 50 ng + | 3.0 mM | 200 | 0.1% Triton-X<br>24 ng BSA |
| Rsa I | RsaF3 : GACCTCCCACCCTCCCCCTC<br>RsaR1: GTCTCCACCCTCCACCCGCCACTC | 20835<br>21652/<br>25453 | 1<br>2<br>3 | 98 : 2.5<br>98 : 1.5<br>98 : 1.5 | 60 : 2<br>61 : 1<br>61 : 1 | 72 : 1.5<br>72 : 1.5<br>72 : 5 | 100 | 50 ng + | 1.0 mM | 200 | 0.1% Triton-X<br>24 ng BSA |
| alpha-Pst I | aPF : GGTTGAATGCTCCAGCCGGTTCCAGCTATTGCT<br>aPR : GAGTCTGTGGGGACCAGAAGAGT | 24727<br>21863/<br>25664 | 1<br>2<br>3 | 95 : 3.5<br>95 : 1.5<br>95 : 1.5 | 59 : 2<br>60 : 1<br>60 : 1 | 72 : 3<br>72 : 3<br>72 : 5 | 100 | 50 ng + | 2.0 mM | 200 | 24 ng BSA |
| theta-Pst I | P99F1 : ACGGTGTCTTCGTTATTTCATCAGG<br>P99R4 : TCAGTGCTTCATTTCTTTTTGTGGC | 27438<br>28325 | 1<br>2<br>3 | 95 : 3.5<br>95 : 1<br>95 : 1 | 59 : 3<br>60 : 2<br>60 : 2 | 72 : 4<br>72 : 4<br>72 : 5 | 100 | 50 ng + | 3.0 mM | 200 | 24 ng BSA |

"Position in cluster" is that relative to the Cap site of the ζ2 gene (D. R. Higgs, personal communication). Stage 1 of PCR is constituted of 2 cycles, stage 2 of 27 cycles and stage 3 of 1 cycle. All reactions use 1U of Vent polymerase. N/A = not applicable (*Xba I* RFLP is positioned 5' to the α-gene cluster; refer to Chapter 1).

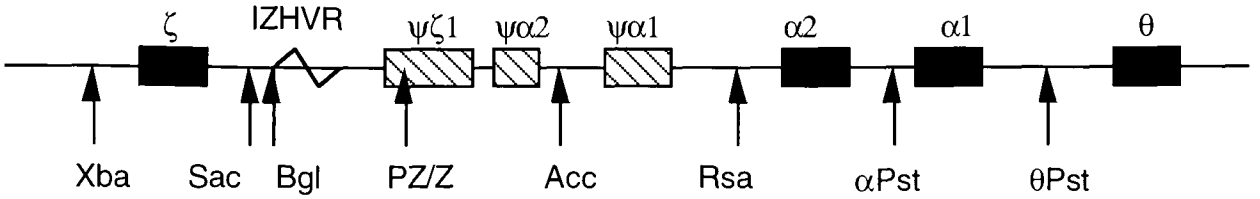
product). When haplotyping by Southern blotting, it is possible to determine linkage between IZHVR, *Bgl I* and *Sac I* sites, although by PCR, only linkage between *Bgl I* and IZHVR is determined. This does not usually represent a significant disadvantage of the method.

The PZ/Z and IZHVR sites are typed according to the size of the products. The other sites are typed following the appropriate restriction enzyme digest. 10-15 µl of each amplified product are incubated at 37°C with an excess of enzyme (~20U), according to the manufacturer's instructions. The excess ensured that the inhibitory effect of DMSO upon the restriction enzyme was overcome, although it may be advisable to initially include controls for non-digestion. Digestion products can then be easily visualised on agarose gels of the appropriate concentration as indicated in Table 2.3 and shown in Figures 2.2 & 2.3. Most digestion patterns are simple to interpret (although it is often helpful to run uncut and cut aliquots of each sample next to each other on the gel) and more so than Southern blots.

The α Pst PCR digestion product is slightly more complicated because it contains two *Pst I* sites: the RFLP site included in the haplotype, and also a rare polymorphism that was noted in the original Southern blotting study of α-haplotypes [Higgs *et al.*, 1986]. The rare *Pst I* site is always (+) in the non-African populations studied to date, but is absent in some Africans on the IIc haplotype. Therefore, its absence can be noted by the uncut 937 bp PCR fragment (Table 2.3). There are a number of other rare polymorphisms within the α-globin complex, but these are not included within the PCR fragments generated for haplotyping in this protocol.

Table 2.3: α-Haplotyping Sites

A)

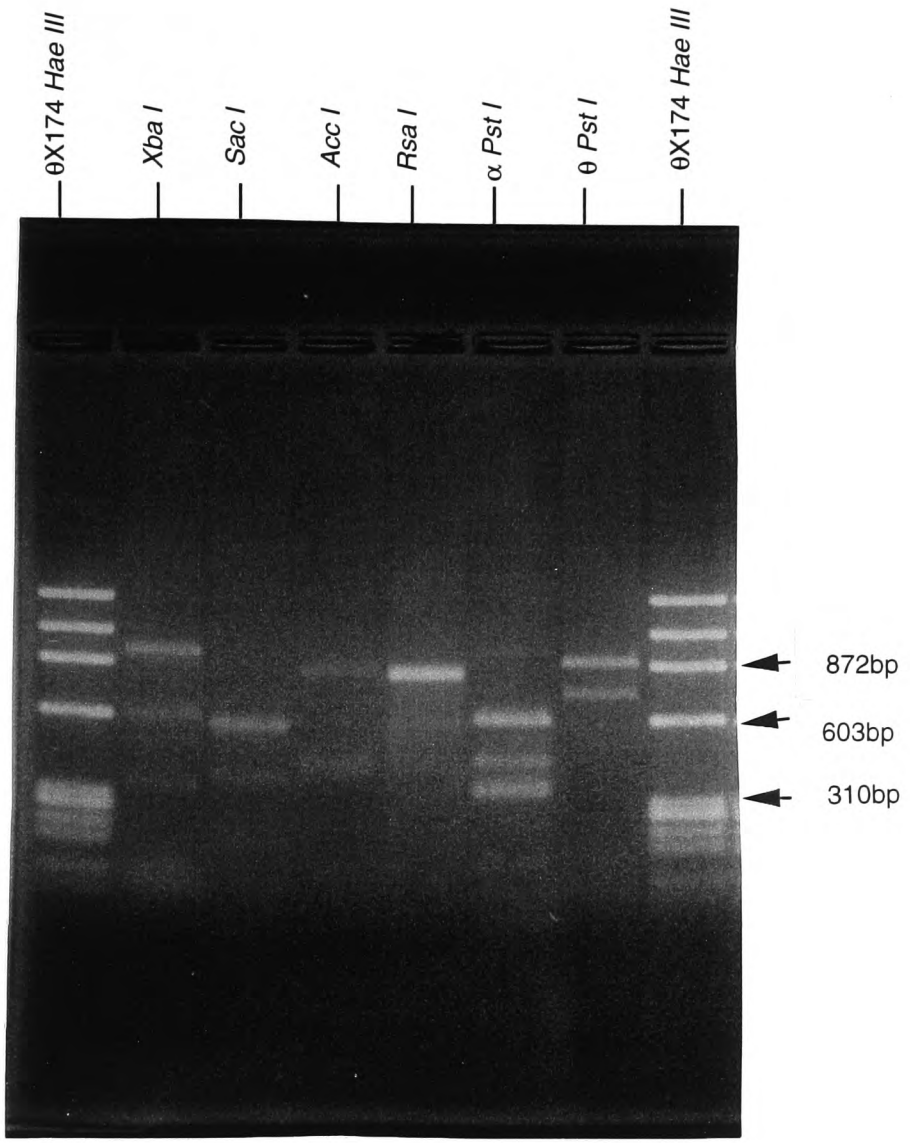



B)

| Site | Digestion Products (bp) | | Gel |
|---|---|---|---|
| | + | - | (% agarose) |
| *Xba I* | 586<br>324 | 910 | 1.5 |
| *Sac I* | 367<br>194 | 561 | 1.5 |
| *Bgl I*/IZ | M ~1600 | S ~ 900<br>M ~ 1800<br>L ~ 2600 | 1.8 |
| PZ/Z | N/A | PZ: ~ 989<br>Z: 620 | 1.5 |
| *Acc I* | 431<br>395 | 826 | 1.5 |
| *Rsa I* | 460<br>357 | 817 | 1.5 |
| α *Pst I* | 436<br>343<br>159 | 595<br>343 | 2 |
| θ *Pst I* | 716<br>172 | 888 | 1.5 |

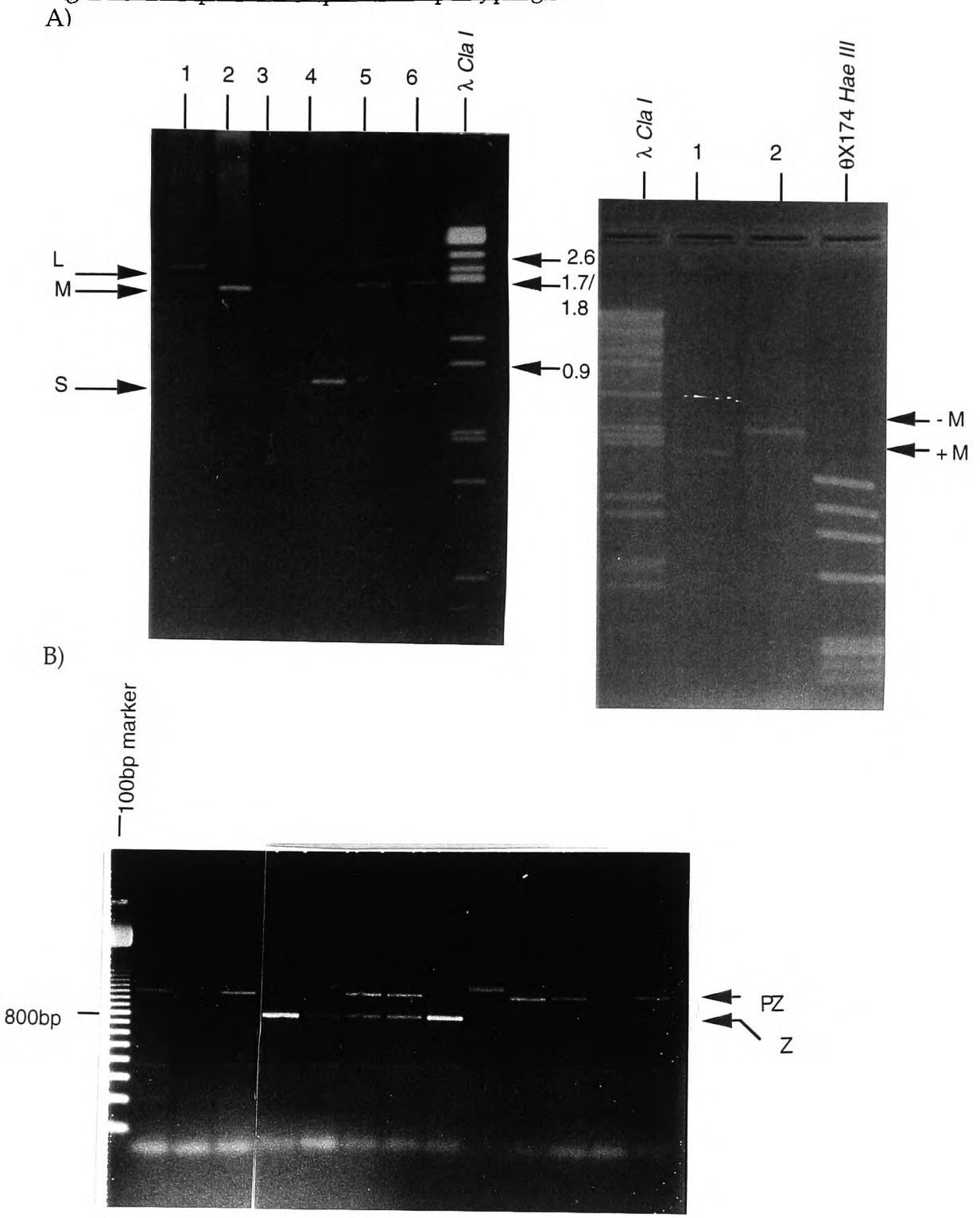
A) α-Globin gene cluster showing positions of α-haplotyping sites. B) Enzyme digestion products and sizes of PCR products. Gel concentrations refer to standard agarose when gels are run in 1x AGB (see Appendix 1). In most cases, the (-) alleles are the same size as the untreated PCR product; the α Pst region contains a rare *Pst I* polymorphism in addition to the RFLP which is detected as a 937 bp fragment (494 + 343 bp).

Figure 2.2: α-Haplotyping Digests




Digests to type non-repetitive regions of the α-haplotype (on 2% agarose gel). Status of haplotype sites in this sample are: +/- *Xba I*, +/- *Sac I*, +/- *Acc I*, -/- *Rsa I*, +/- α *Pst I* and +/- θ *Pst I*. Refer to Table 2.3 for sizes of digested products.

Figure 2.3: Repetitive Sequence Haplotyping Sites




A) IZHVR S, M, L size classes (sizes on right refer to λ Cla I marker) and *Bgl I* digestion of M IZHVR product (right) - sample 1 is +M, sample 2 is -M. B) PZ/Z polymorphism. All gels are 1.5% agarose.

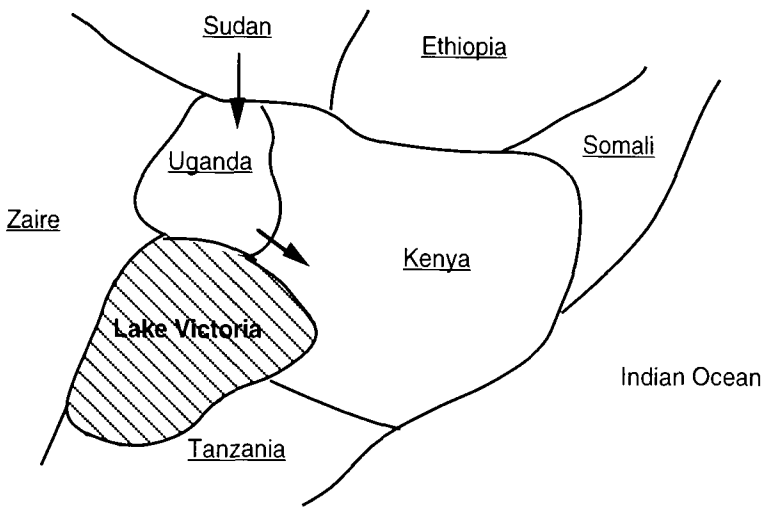
## A Study of African Diversity: α-Haplotypes in Kenya

This study was performed firstly to illustrate the use of the PCR protocol detailed in the first part of this chapter and to compare its efficiency with the method of Southern blotting. In addition, African diversity is poorly represented at most genetic loci [Excoffier *et al.*, 1987]: data are often biased towards one population group of interest, and therefore not very informative, and gathered for medical rather than evolutionary studies. Genetic trees drawn to determine world population affinities require more than just one African population, although this has not been common practice; inclusion of greater information on African mtDNA diversity has been noted to alter the branch lengths of these trees [Poloni *et al.*, 1995]. This study will help generate more information about the intracontinental diversity at the α-globin gene locus.

### α-Haplotype Diversity of Luo speakers from Kenya

The diverse range of languages in Africa can be classified into four major families: Niger-Kordofanian, Nilo-Saharan, Afroasiatic and Khoisan [Greenberg, 1963]. In East Africa, all four language families are spoken, whereas West Africa is more linguistically homogenous, with Niger-Congo (a subdivision of Niger-Kordofanian) being the most common language. To complement the linguistic data, genetic differences between East and West Africa are observed at a number of loci [Excoffier *et al.*, 1987]. Luo is a Nilo-Saharan language spoken in East Africa by people who originated in Sudan, but over the past 450 years have spread to Uganda, Kenya and Tanzania [Cohen and AthienoOdhiambo, 1989]. Luo speakers are believed to have entered Western Kenya around the north of Lake Victoria from Uganda ~350 years ago (Figure 2.4).

Figure 2.4: East Africa Showing Migration Route of Kenyan Luos




Arrows show direction of the Luo migration from Sudan into Uganda and onto Kenya, around the north of Lake Victoria in East Africa.

A sample of 72 Luo-speaking unrelated individuals was α-haplotyped by PCR. α-haplotyping was performed according to the protocol outlined earlier in this chapter. A number of sites were also typed by Southern blotting and the results compared with those obtained by PCR. It was found that correlation between the two methods was exact for the RFLP sites, with incomplete digestion of PCR products not being evident in any samples tested. Results obtained from PCR typing were generally easier to interpret, due to less complicated digestion patterns and by overcoming any problems associated with unspecific probe binding and high background signal on filters. Problems with Southern blotting were noted, however, associated with uneven hybridisation of the probe to different sized bands of the IZHVR locus and consequent errors in typing: primarily S/M heterozygotes typed as M/M homozygotes. These errors have been reported previously [Armour *et al.*, 1992; Chakraborty and Jin, 1992; Steinberger *et al.*, 1993] and occur because the intensity of the hybridisation signal is proportional to the number of tandem repeats.

Genotypes[1] were determined by digestion with *Bgl II*, Southern blotting and probing with ζ (refer to Appendix 1; raw haplotype and genotype data can be found in Appendix 2).

The unambiguous Luo haplotype composition (based on homozygotes and single-site heterozygotes) is shown in Table 2.4 with data from three other African populations which have been studied at this locus. The predominant haplotypes in the Luo, as in the Bantu, are the IIIf and IIg, and the presence of group VII haplotypes confirms the Luo haplotype repertoire as fairly "typical" of African populations studied so far [Higgs *et al.*, 1986; Martinson *et al.*, 1995; Ramsay and Jenkins, 1988]. However, it is immediately obvious that there is only a small number of resolved chromosomes: only 32 from a total 144 (22%). This represents a relatively low level of successful haplotype assignment, although this is not uncommon for African populations, as is shown in Table 2.5.

In European populations, many haplotypes can be assigned because one haplotype, the Ia, is predominant (50% of chromosomes), so that in heterozygotes for Ia and another haplotype, the latter can be inferred. In African populations, a large number of different haplotypes are often present [Martinson *et al.*, 1995] and although some haplotypes will be more common than others, there will be far fewer samples which are homozygotes or single-site heterozygotes than in a European population. Thus, African populations, by virtue of their diversity, present problems for haplotype studies. Previous studies have partly overcome this problem by utilising genealogical information [Martinson *et al.*, 1995] or by studying only nuclear families [Ramsay and Jenkins, 1988]. However, in the absence of such information, only the more common haplotypes will be assigned and rarer haplotypes will not be

[1]Genotypes were determined by Michelle Rugless in this laboratory.

Table 2.4: α-Haplotypes in Africa

| Haplotype | | % | | | |
|---|---|---|---|---|---|
| | | Luo | Senegal | Bantu | San |
| Ia | + + − M PZ + + − − | 3.1 | 11.4 | 5.6 | |
| Ib | − + − M PZ + + − − | | 1.9 | | |
| Ic | + − − M PZ + + − − | | 1.9 | | |
| Ie | − − + M PZ + + − − | | | 2.8 | |
| If | − − − M PZ + + − − | | | 2.8 | |
| IIc | + + − M PZ + − − − | 9.4 | 17.6 | 2.8 | 10.8 |
| IIc* | + + − M PZ + − * − | | | 5.6 | |
| IId | − − − M PZ + − − − | | 1.4 | | |
| IIe | + − − S PZ + − − − | | 2.9 | | |
| IIg | + − − M PZ + − − − | 15.6 | 4.3 | 11.1 | 56.8 |
| IIj | − + − M PZ + − − − | 12.5 | 4.3 | | 2.7 |
| IIk | − − − M PZ + − − − | | 1.9 | | |
| III | − + − S PZ + − − − | | 0.5 | | |
| IIm | + + − S PZ + − − − | | 1 | | |
| IIIa | − − + M Z − − − − | | 1 | 2.8 | |
| IIIb | + − + M Z − − − − | | 1.4 | 2.8 | 5.4 |
| IIId | + − − M Z − − − − | | 6.2 | | |
| IIIe | − + − M Z − − − − | 3.1 | 1 | | |
| IIIf | − − − M Z − − − − | 31.3 | 8.1 | 16.7 | |
| IIIi | − + + M Z − − − − | | 1.4 | | |
| IVa | + − − S PZ + − + + | | | 2.8 | |
| IVd | + + − M PZ + − + + | | | 5.6 | |
| Vc | + − − S PZ + − − + | | | 5.6 | 10.8 |
| Ve | + + − M PZ + − − + | | | 5.6 | |
| Vf | + − − M PZ + − − + | | | 2.8 | 10.8 |
| VIb | + + − L PZ − − − − | | 1 | | |
| VId | + − − S PZ − − − − | | 0.5 | | |
| VIe | + − − M PZ − − − − | | | | 2.7 |
| VIf | − + − M PZ − − − − | | 1 | | |
| VIIa | − + − M Z + − − − | 3.1 | 2.9 | | |
| VIIb | − + + M Z + − − − | | 0.5 | | |
| VIIc | + − − M Z + − − − | 3.1 | 7.6 | 2.8 | |
| VIId | + − + M Z + − − − | | 0.5 | | |
| VIIe | − − − M Z + − − − | 6.3 | 5.2 | | |
| VIIf | + + − M Z + − − − | | 1.9 | | |
| VIIg | + + − S Z + − − − | 6.3 | 1 | 5.6 | |
| VIIi | − + − S Z + − − − | 6.3 | 6.2 | 11.1 | |
| VIIIc | + + − M PZ + − + − | | 1.9 | | |
| VIIId | + − − M PZ + − + − | | 1 | | |
| IXb | + + − M Z + + − − | | 1 | | |
| XIa | + + − M PZ + + − + | | | 2.8 | |
| XIIa | + − − M Z + − + − | | | 2.8 | |
| No. chromosomes classified | | 32 | 210 | 36 | 37 |
| No. unclassified | | 108 | 180 | - | - |
| incomplete | | 4 | 10 | - | - |
| total chromosomes | | 144 | 400 | 36 | 37 |

detected. In determining population relationships, it is necessary to know as much about the haplotype distributions as possible, including the rare haplotypes.

---

Table 2.5: Proportion of Successfully Assigned Chromosomes in Different Populations

| Population | Assigned Chromosomes | No. Chromosomes Studied |
|---|---|---|
| Australian Aborigines | 98% | 212 (4) |
| Britain | 82% | 44 (12) |
| Mediterranean | 68% | 78 (16) |
| Asian Indian | 71% | 74 (12) |
| Saudi Arabia | 71% | 38 (24) |
| Papua New Guinea | 40% | 64 (4) |
| Amazonian Indian | 40% | 99 (8) |
| Island Melanesia | 56% | 40 (22) |
| Jamaica | 32% | 38 (0) |
| Nigeria | 7% | 38 (10) |

Data from [Higgs *et al.*, 1986; Roberts-Thomson *et al.*, in press; Zago *et al.*, 1995]. Number of incomplete chromosomes are shown in parentheses.

---

There are a number of possible approaches to determining more about a population haplotype distribution. Single chromosomes could be studied by the use of an asymmetric or single-molecule PCR technique [Ruano *et al.*, 1990], although such approaches are technically demanding and were not attempted in this study. Another way is to estimate haplotype frequencies by the use of a maximum-likelihood (ML) algorithm. A program was used here which utilises an expectation-maximisation (EM) algorithm to produce estimates of haplotype frequencies, under the assumption of Hardy-Weinberg proportions [Excoffier and Slatkin, in press] .

This method builds upon the basic principle of making a preliminary list of the haplotypes in a sample by examining the unambiguous individuals and then examining the remaining individuals for the occurrence of any of the previously defined haplotypes. The complementary haplotype can be added to the list of recognised haplotypes until the phase information for all individuals is resolved or

identified as unresolvable. The maximum-likelihood method, however, places the problem within an EM algorithm, making the process an iterative one and starts with arbitrary values for haplotype frequencies. Successive sets of haplotype frequencies are computed until convergence is reached; that is, when the change in haplotype frequency in consecutive iterations is less than a small specified value (epsilon).

The program was run with two different values of epsilon ($e = 10^{-7}$ and $10^{-10}$) and the best solution (with the highest likelihood value) was retained and is shown in Table 2.6. Only haplotypes predicted to have polymorphic frequencies ($\geq 1\%$) are shown. The solution agrees with the results based simply on unambiguous chromosomes on the two most frequent haplotypes (IIIf and IIg) but disagrees with the frequencies of the other haplotypes. The program resolved 97% of chromosomes, but 50% (19 from 37) of these assigned haplotypes are novel, that is, they have not been seen in any other population to date. Ten of these novel haplotypes are new combinations of the 5' and 3' ends of the haplotype, but nine are completely new haplotype groups (3' ends; refer to Chapter 1). The total frequency of these novel haplotype groups is 13% of the entire haplotype repertoire, which is exceptionally high.

Although populations (particularly African populations) do show novel groups, these have always been relatively rare [Martinson *et al.*, 1995; Ramsay and Jenkins, 1988], and do not contribute significantly to the observed haplotype diversity. It is possible that the Kenyan Luo represent a special case with very high frequencies of novel 3' ends, but there is no reason to suspect that they do. All ML solutions, and not just that shown in Table 2.6, demonstrated the same high level of contribution from these haplotypes (11-13% of the total assigned haplotypes). Two novel 3' haplotypes, PZ - - - + and PZ - - + +, appeared at consistently higher frequencies

Table 2.6: Maximum Likelihood Estimation of Luo α-Haplotype Repertoire

| Haplotype | | % Max Likelihood |
|---|---|---|
| Ia | + + − M PZ + + − − | 7 |
| Ib | − + − M PZ + + − − | 1 |
| Ic | + − − M PZ + + − − | 1 |
| IIc | + + − M PZ + − − − | 5.3 |
| IIe | + − − S PZ + − − − | 3.6 |
| IIg | + − − M PZ + − − − | 10.5 |
| IIj | − + − M PZ + − − − | 6.7 |
| IIh* | − + − S PZ + − − − | 1.5 |
| IIIe | − + − M Z − − − − | 1 |
| IIIf | − − − M Z − − − − | 16.3 |
| IIIj* | + + − M Z − − − − | 1 |
| IVe* | + + − S PZ + − + + | 1 |
| IVf* | + − − M PZ + − + + | 1 |
| Vc | + − − S PZ + − − + | 1.9 |
| Ve | + + − M PZ + − − + | 1.7 |
| VId | + − − S PZ − − − − | 1.2 |
| VIg* | + + − S PZ − − − − | 1 |
| VIh* | − − − M PZ − − − − | 4 |
| VIi* | − + − S PZ − − − − | 1 |
| VIIa | − + − M Z + − − − | 1.5 |
| VIIc | + − − M Z + − − − | 1.5 |
| VIIe | − − − M Z + − − − | 2.8 |
| VIIg | + + − S Z + − − − | 2.3 |
| VIIi | − + − S Z + − − − | 8.9 |
| VIIj* | + + + S Z + − − − | 2.1 |
| VIIk* | − − − S Z + − − − | 1 |
| VIIId | + − − M PZ + − + − | 1 |
| XIIb* | + − − S Z + − + − | 1.6 |
| NOVEL 1 | + + − M PZ + + + + | 2 |
| NOVEL 2 | + + − M PZ − + + + | 1 |
| NOVEL 3A | + + − M PZ − − − + | 1 |
| NOVEL 3B | + − − S PZ − − − + | 1.2 |
| NOVEL 4 | + − − S PZ − − + + | 3.9 |
| NOVEL 5 | + − − M PZ + − + − | 1 |
| NOVEL 6 | − − − M PZ − − + − | 1 |
| NOVEL 7 | + − + S Z − − + − | 1 |
| NOVEL 8 | + − − S Z − − − + | 1 |

Solution with greatest associated likelihood derived using ML program ($e=10^{-7}$) described in text. Haplotype frequency is shown as percentage. The total number of fully typed individuals in this solution is 70. Haplotypes marked with an asterisk are novel.

than the other novel 3' ends, and it is possible that these are genuinely contained within the Luo repertoire, but such a judgement would have to be based on other evidence.

The problem with using ML programs may primarily be that the program has no biological or other criteria on which to base haplotype estimations. Although the ML program was seen to perform well in a previous study of α-globin haplotypes in Africa [Martinson *et al.*, 1995], this may be a consequence of the nature of the haplotype repertoire in the Senegal population, rather than the efficiency of the program. Much is already known about the frequencies of haplotypes in world populations, and this information can be used to judge which haplotypes are most likely within any one multiply heterozygous individual. For instance, it seems unlikely that the haplotype PZ + + + + , suggested by the ML program, is present in the Luo. This group has not been seen in any other African or non-African population, and if the unambiguous 3' ends of the haplotypes in the data set are considered, it seems likely that this 3' end is not present in the Luo either.

The unambiguous haplotype groups (ignoring 5' motifs) in the Luo are shown in Table 2.7. 83 (59%) chromosomes could be resolved in this way. The frequencies of the haplotypes groups I, II, III and VII predicted from full unambiguous haplotypes are very similar to the predictions for the same groups here. However, five other groups are revealed: IV, V, VI, VIII and one novel (PZ - - - +), which was suggested by the ML method. Thus, the Luo have quite a variable haplotype repertoire, with all eight major haplotype groups represented. This may indicate why the ML method produced so many novel haplotypes; combining a rarer group IV haplotype with a more common type I, for example, would produce the following multiple 3' end heterozygote: PZ +/+ +/- +/- +/-. This individual might be resolved as PZ + + + + (a novel 3' end) and PZ + - - - (II). The type II is a "genuine" 3' end seen within the

data set but the other haplotype is novel and a type consistently suggested by the ML program.

---

Table 2.7: Luo Unambiguous Haplotype Groups

| | 3' End | Frequency | Frequency (from unambiguous chrms) |
|---|---|---|---|
| I | PZ + + - - | 3.9 | 3.1 |
| II | PZ + - - - | 33.8 | 37.5 |
| III | Z - - - - | 26 | 34.4 |
| IV | PZ + - + + | 2.6 | - |
| V | PZ + - - + | 3.9 | - |
| VI | PZ - - - - | 7.8 | - |
| VII | Z + - - - | 27.3 | - |
| VIII | PZ + - + - | 1.3 | - |
| Novel | PZ - - - + | 1.3 | - |

Percentage frequencies are shown. Chrms = chromosomes.

---

Given the 3' ends unambiguously present in the data set (Table 2.7), the above example would be assigned by inspection as a IV/I heterozygote, since both haplotype groups are seen in this and other populations. The ML program does not output data referring to specific individuals in the data set so it is unclear how often the program would assign such a heterozygote as a combination between a novel and a known haplotype, but from the ML solutions obtained and the frequencies in Table 2.6, it would seem that the program is underestimating the frequencies of group IV haplotypes, for example, and overestimating novel haplotypes.

The reason that the ML program gave better results (novel 3' haplotypes represented only a total 1.9% of the suggested repertoire) with the study of Senegalese Mandenka α-haplotypes [Martinson *et al.*, 1995] is likely to be related to the fact that type IV and V haplotypes are absent in that population and thus complex 3' heterozygotes ( e.g.. PZ +/+ +/- +/- +/-) are not present, whereas a greater number of unambiguous chromosomes are. The identification of type IV and V haplotypes in

the Luo means that the Mandenka population is the only African population studied so far to not have group IV and/or V haplotypes.

This can be illustrated further by considering the individual site frequencies from different African populations (Table 2.8). Although haplotypes are the most powerful way to relate populations to one another, if there are problems in assigning full haplotypes, consideration of site frequencies can at least use all the data. Given some knowledge about haplotypes in populations, it may then possible to infer a certain amount about a specific population's haplotype repertoire from its site frequencies. For example, most African populations show relatively low frequencies of *Bgl I* (+), related to the low frequency of haplotypes such as the IIIa which carry this mutation, and they also show high frequencies of *Acc I* (+) related to the high frequencies of haplotype groups I, II and VII [Higgs *et al.*, 1986; Martinson *et al.*, 1995; Ramsay and Jenkins, 1988].

The frequencies of the α and θ *Pst I* sites effectively reflect the frequencies of haplotypes IV, V and VIII in these populations. It can be seen that the Senegalese Mandenka show a relatively low frequency of α *Pst I* (+); the haplotype repertoire indicates that these are associated with group VIII and not group IV haplotypes. This site frequency is significantly lower than in the Luo, Gambian and Zambian populations ($\chi^2$ analysis, Table 2.9). At the θ *Pst I* site, the Mandenka have significantly lower frequencies than all other African groups, confirming that group IV and V haplotypes are not present in the Mandenka, but they probably are in the other populations.

**α-Haplotypes and African Prehistory**

African populations share a longer and probably more complicated history than any other continental group; the oldest hominoid fossils are found in Africa. It is from

Table 2.8: Individual Site Frequencies of African Populations

| Site | Frequency of (+) allele | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Kenya | Senegal | San | SA Blacks | Gambia | Zambia | Nigeria | Jamaica |
| Xba I | 0.55 (144) | 0.59 (398) | 0.97(39) | 0.62(40) | 0.444(36) | 0.500(38) | 0.526(38) | 0.579(38) |
| Sac I | 0.44 (144) | 0.55 (394) | 0.15(40) | 0.44(39) | 0.412(34) | 0.526(38) | 0.421(38) | 0.526(38) |
| Bgl I | 0.03 (144) | 0.07 (396) | 0.05(40) | 0.07(40) | 0.083(36) | 0.105(38) | 0.026(38) | 0.079(38) |
| IZHVR-S | 0.33(140) | 0.25(398) | 0.10(40) | 0.26(39) | 0.438(36) | 0.158(38) | 0.184(38) | 0.108(37) |
| IZHVR-M | 0.67(140) | 0.74 (398) | 0.90(40) | 0.74(39) | 0.562(36) | 0.816(38) | 0.816(38) | 0.892(37) |
| IZHVR-L | - | 0.01(398) | - | - | - | - | - | - |
| PZ | 0.59 (144) | 0.50 (390) | 0.95(40) | 0.55(40) | 0.441(34) | 0.579(38) | 0.553(38) | 0.500(38) |
| Acc I | 0.67(144) | 0.74(386) | 0.87(40) | 0.72(39) | 0.750(36) | 0.684(38) | 0.853(34) | 0.684(38) |
| Rsa I | 0.12 (144) | 0.13(384) | 0.02(40) | 0.18(39) | 0.139(36) | 0.211(39) | 0.289(38) | 0.263(38) |
| Alpha Pst I | 0.12 (144) | 0.05(388) | 0.00(40) | 0.10(40) | 0.167(36) | 0.158(38) | 0.079(38) | 0.026(38) |
| Theta Pst I | 0.17(144) | 0.01(394) | 0.20(40) | 0.21(40) | 0.111(36) | 0.132(38) | 0.132(38) | 0.079(38) |
| Source | This study | Martinson et al [1995] | Ramsay and Jenkins [1988] | | D.R.Higgs (personal communication) | | | |

Numbers of chromosomes are shown in parentheses. Frequencies of (+) RFLP alleles are shown. Frequencies of length allele "PZ" and all three IZHVR alleles are shown.

## Table 2.9: $\chi$-Analysis of $\alpha$-Haplotyping Site Information in African Populations

| Site | Kenya and Senegal | | Kenya and Bantu | | Kenya and San | | Kenya and Gambia | | Kenya and Zambia | | Kenya and Nigeria | | Kenya and Jamaica | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* |
| Xba | 0.745 | 0.3879 | 0.743 | 0.3886 | 24.124 | <0.0001* | 1.254 | 0.2628 | 0.286 | 0.5929 | 0.06 | 0.8061 | 0.112 | 0.7378 |
| Sac | 5.668 | 0.0173* | 0.02 | 0.8878 | 11.04 | 0.0009* | 0.074 | 0.7853 | 0.956 | 0.3282 | 0.033 | 0.8556 | 0.956 | 0.3282 |
| Bgl | 3.491 | 0.0617 | 1.907 | 0.1672 | 0.49 | 0.4839 | 2.378 | 0.123 | 4.709 | 0.03* | 0.0002 | 0.9609 | 2.129 | 0.1446 |
| IZ | 33.505 | <0.0001* | 19.983 | <0.0001* | 8.932 | 0.0028* | 1.685 | 0.1943 | 4.21 | 0.0402* | 2.979 | 0.0844 | 7.018 | 0.0081* |
| PZ | 3.437 | 0.0638 | 0.209 | 0.6478 | 18.278 | <0.0001* | 2.484 | 0.115 | 0.016 | 0.8996 | 0.175 | 0.6755 | 1.001 | 0.3172 |
| Acc | 2.467 | 0.1163 | 0.158 | 0.6907 | 6.625 | 0.0101* | 0.924 | 0.3364 | 0.042 | 0.8379 | 4.564 | 0.0326* | 0.042 | 0.8379 |
| Rsa | 0.203 | 0.6521 | 1.016 | 0.313 | 3.071 | 0.0797 | 0.117 | 0.7328 | 1.973 | 0.1602 | 6.787 | 0.0092* | 8.776 | 0.0031* |
| aP | 51.494 | <0.0001* | 1.078 | 0.2991 | 7.667 | 0.0056* | 0.611 | 0.4345 | 0.432 | 0.5109 | 0.47 | 0.4929 | 2.839 | 0.092 |
| tP | 6.737 | 0.0094* | 1.79 | 0.1809 | 1.79 | 0.1809 | 0.677 | 0.4107 | 0.276 | 0.5991 | 0.276 | 0.5991 | 1.1831 | 0.176 |

| Site | Senegal and San | | Senegal and Bantu | | Senegal and Gambia | | Senegal and Zambia | | Senegal and Nigeria | | Senegal and Jamaica | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* | $\chi^2$ | *p value* |
| Xba | 22.329 | <0.0001* | 0.18 | 0.6715 | 2.886 | 0.0894 | 1.167 | 0.28 | 0.588 | 0.4433 | 0.019 | 0.8904 |
| Sac | 23.347 | <0.0001* | 1.885 | 0.1697 | 2.434 | 0.1187 | 0.084 | 0.7724 | 2.347 | 0.1255 | 0.084 | 0.7724 |
| Bgl | 0.243 | 0.622 | 0.01 | 0.9198 | 0.079 | 0.7787 | 0.785 | 0.3755 | 1.096 | 0.2952 | 0.035 | 0.8506 |
| IZ | 4.593 | 0.0321* | 2 | 0.1573 | 6.291 | 0.0121* | 1.643 | 0.1999 | 0.842 | 0.3588 | 3.813 | 0.0509* |
| PZ | 29.593 | <0.0001* | 0.363 | 0.5469 | 0.433 | 0.5106 | 0.863 | 0.3528 | 0.384 | 0.5356 | 0 | - |
| Acc | 3.508 | 0.0611 | 0.097 | 0.7555 | 0.014 | 0.9054 | 0.573 | 0.4492 | 2.091 | 0.1482 | 0.573 | 0.4492 |
| Rsa | 3.789 | 0.0516* | 0.0737 | 0.3905 | 0.022 | 0.8826 | 1.68 | 0.195 | 7.093 | 0.0077* | 9.502 | 0.0021* |
| aP | 2.05 | 0.1522 | 1.857 | 0.173 | 8.225 | 0.0041* | 7.434 | 0.0064* | 0.635 | 0.4255 | 0.397 | 0.5287 |
| tP | 46.681 | <0.0001* | 46.681 | <0.0001* | 18.414 | <0.0001* | 25.05 | <0.0001* | 25.05 | <0.0001* | 10.289 | 0.0013* |

this continent that two major hominoid migrations are thought to have occurred: of *Homo erectus* (1-2 mya) to Asia and Europe, and then of anatomically modern humans, *H. sapiens sapiens* [Stringer and Andrews, 1988]. Information about human migrations and gene flow within Africa and between African and non-African populations, is undoubtedly contained in the within-Africa diversity, and may help address the question of how many migrations have occurred between the different continental regions [Lahr and Foley, 1994; Nei and Roychoudhury, 1993]. Due to the vastness of the African continent, peoples have often been subjected to isolation, and within sub-Saharan Africa, ecological fragmentation has occurred over time so that considerable genetic diversification via founder effects, genetic drift and possibly differential selective pressures has resulted [Excoffier *et al.*, 1987].

The ancestral sub-Saharan population may have originated in East Africa; from here, movement to South Africa and strong founder effects produced the Khoisan (there is evidence for these founder effects from α-haplotypes, [Ramsay and Jenkins, 1988]). The ancestral population also moved West and differentiated progressively [Excoffier *et al.*, 1987]; it is likely that waves of settlement to the West have occurred since ~9000 years ago, from a region centred around the Nile (possibly near present-day Khartoum). The heterogeneous Nilo-Saharans may represent the scattered remnants of this ancient East to West Sudanic peopling zone [Ki-Zerbo, 1990]. Thus, the Nilotic (Nilo-Saharan) Luo are part of an old group that would be expected to have affinities with almost all other African groups. Such affinities are suggested from the α-haplotype composition of the Luo, although they seem closest to the Bantu (from full haplotype data; Table 2.4) and to the Gambians (from site frequency data; Tables 2.8 & 2.9).

During their history, Luo peoples have most likely had contacts with Afroasiatic (Kushitic) people in the Nubian stretch of the Nile Valley, as indicated by some

transfer of language from the Kushitics to the Nilotics [Mokhtar, 1990; Oliver, 1991]. The Bantu expansion, which began ~2000-3000 years ago in Cameroon or Nigeria, and perhaps related to the use of agriculture, pushed the San and Khoi into further isolation in the South and East, but Nilotic East Africa formed part of this push and may explain the affinities seen between the South African Bantu and the Luo. This concurs with studies of sickle cell ($\beta^S$) chromosomes which show that the Luo in Kenya possess the CAR or Bantu type of β-globin haplotype [Nagel and Ranney, 1990; Ojwang *et al.*, 1987].

It is interesting that the absence of group IV and V haplotypes in the Senegalese Mandenka differentiates them not only from the Luo, but also from the other African groups studied, including their geographical neighbour, Gambia. It is therefore possible that present-day West African populations are relatively heterogeneous, with some, such as the Gambians and Nigerians, having clearly closer links with East Africa (site-frequency data; Table 2.8). The Mandenka may have closer links with other groups, possibly Caucasian populations which also have low frequencies of group IV and V haplotypes [Higgs *et al.*, 1986]. In a study of mtDNA RFLPs, three contemporary Senegalese populations (Mandenka, Wolof and Pular) were found to contain a number of haplotypes related to those found in non-African populations, with evidence of gene flow with other African groups [Chen *et al.*, 1995]. This is likely to be related to the cultural and economic interactions with Saharan and North African populations, since the first sub-Saharan cities arose in West Africa consisted of Mande-speakers (Mandenka) and were related to the establishment of trading routes across the Sahara.

**African Diversity and the Mechanisms producing Diversity in Populations**

Diversity at the α-globin locus is high in African populations. This diversity is clearly useful for studying African population affinities, even with the complex and

ancient history of the peoples from this continent. This is encouraging since it suggests that the types of population analyses that have been possible in the Pacific (refer to Chapter 1) may also be possible in other geographical regions given sufficient amounts of data on variation. It is also evident that there are differences between African populations, not only in haplotype repertoire, but also in amount of haplotype diversity [Martinson *et al.*, 1995]. The Senegalese Mandenka show particularly high levels of sample-size independent haplotype diversity compared to other African, European, Polynesian and Asian populations. The South African Bantu were found to have comparable levels of diversity, but given their heterogeneous origins, the levels of diversity may be inflated. The Luo show significantly less haplotype diversity than the Mandenka. This was demonstrated by the use of a bootstrap approach in which a smaller sample, equivalent to the set of 32 unambiguous Luo haplotypes, was generated randomly from the Mandenka population [Martinson *et al.*, 1995]. This procedure was repeated one thousand times to produce an estimate of the expected number of haplotypes that would be seen in the Mandenka population (Table 2.10):

Table 2.10: Comparison between Diversity in the Luo and Mandenka Populations

| | Luo | Senegal | |
|---|---|---|---|
| No. Chromosomes | 32 | 32 | t= 36.2397; p<0.0005 |
| No Haplotypes | 11 | 15.837 | |

It may be that the Mandenka in Senegal are very diverse, even for African populations. Neutral theory predicts that diversity in a population is due to a balance between new mutation and loss of alleles by genetic drift [Kimura, 1983](Chapter 1). Assuming that all new mutations are unique (the infinite-alleles model), diversity ($\theta$) is related to the effective (or breeding) population ($N_e$) and mutation rate ($\mu$) according to equation 2.1:

$$\theta = 4N_e\mu \qquad (2.1)$$

If population sizes are large or mutation rate is high, levels of haplotype diversity will be high. There is no evidence at present to suggest that the mutation rate is higher in the Mandenka than in other African populations, so it is more likely that the high levels of diversity are caused by demographic effects. As suggested earlier, the history of living in large settlements by the Mandenka peoples, possible contact with Caucasian/North African peoples possessing different haplotypes, and also gene flow with other Sub-Saharan Africans [Chen *et al.*, 1995; Graven *et al.*, 1995; Poloni *et al.*, 1995] may all have influenced diversity in this population.

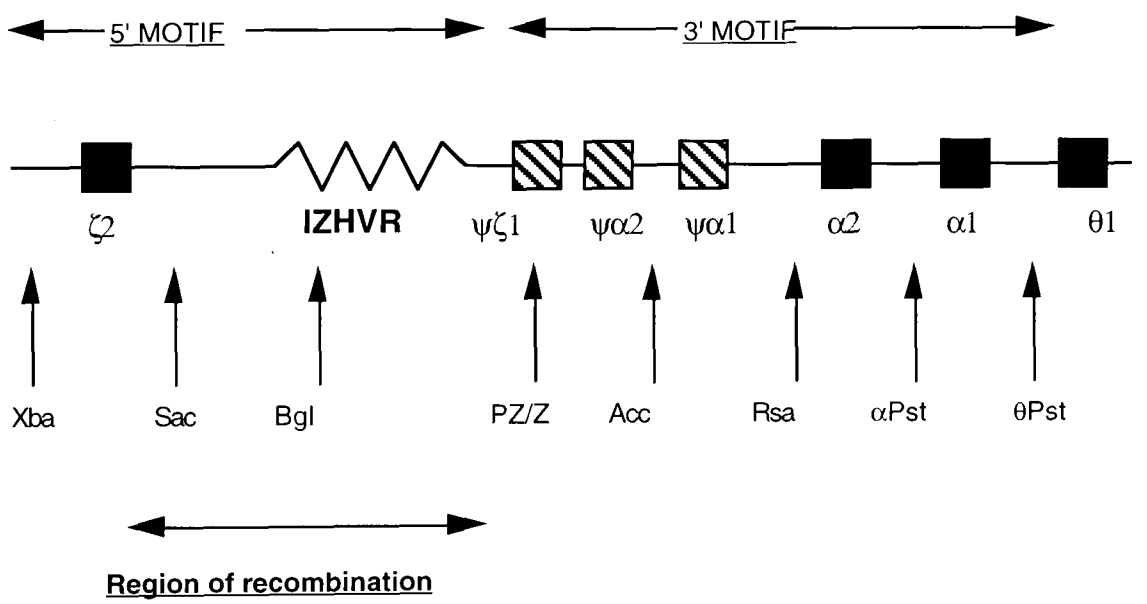
Selection, particularly malarial selection (which is experienced in sub-Saharan Africa) for α-thalassaemia [Flint *et al.*, 1986] is another possible influence on haplotype diversity; if the α-thalassaemia deletion occurred on just a few haplotypes, then these might come to dominate a population's repertoire by the action of selection (Chapter 1). The frequency of α-thalassaemia ($-\alpha^{3.7}$ type) in the Luos is quite high (19% gene frequency) presumably due to selection, but these deletions are present on at least seven different haplotypes (as is true also of α-thalassaemia deletions in the Mandenka; J J Martinson, personal communication) and therefore it seems unlikely that selection is strongly biasing the haplotype distribution in the Luo, and thereby lowering diversity.

The haplotype diversity in these African, and in non-African, populations seems to be produced predominantly by a mechanism of recombination between the 5' and 3' ends of the haplotype (Figure 2.5). The African populations are particularly difficult to haplotype unambiguously precisely because it is difficult to link the 5' and 3' ends. In the Luo, it was seen that over half of the 3' ends could be unambiguously

assigned whereas only 22% full haplotypes could be identified. Generally, the 3' portion of the haplotype seems more stable: only twelve or thirteen of the possible thirty-two ($2^5$) 3' ends have been found, whereas all eight ($2^3$) 5' RFLP combinations have been found, as have many of the twenty-four (8x3) 5' combinations of RFLPs and IZHVR. The great variety of novel 3' ends suggested by the ML program for the Luo probably does not reflect the true picture, since only one novel haplotype group was unambiguously present. This 3' end "stability" suggests that the arbitrary classification of haplotype groups from the 3' end sites (Chapter 1) may actually represent a mechanistically meaningful way in which to describe α-globin gene cluster variation.

---

Figure 2.5: Region of Preferential Recombination in the α-Haplotype




---

It is possible for haplotypes to be formed by a number of mechanisms, including point mutation and recombination. In order to describe the haplotype variety seen in African populations, mutations would have to recur often, so that it is recombination which seems to be more likely to produce the observed haplotype diversity. Within populations, the same 5' end is often found associated with a number of different

haplotype groups. For example, Asian and European (but not African) populations have a high frequency of the IIa haplotype ( - + - L PZ + - - -). The 5' end, - + - L, is also found on a number of other groups to form rarer European haplotypes such as the IVc ( - + - L PZ + - + +) and the VIa ( - + - L PZ - - - -) [Higgs *et al.*, 1986]. The group IV haplotypes in Africa and the Pacific, however, are associated with 5' ends that are more common in these populations (e.g. the IVa: + - - S PZ + - + +). In the Australian Aborigine population, ~75% of haplotypes are IIIa ( - - + M Z - - - -) [Roberts-Thomson *et al.*, in press]. A number of novel haplotypes are also found, but these are mostly characterised by having the 5' end associated with the IIIa, - - + M, in association with a number of different 3' ends. For example, the 5' end of the novel haplotype IIh ( - - + M PZ + - - -) has most likely come from the IIIa, so that this haplotype is formed by recombination between two different haplotype groups.

The infinite-alleles model (IAM) assumes unique mutation, but frequent recombination would lead to an excess number of haplotypes, since motifs produced by point mutation are being constantly shuffled. Within the Senegalese Mandenka population, the haplotype distribution was found to be in agreement with the predicted haplotype distribution based on the IAM [Martinson *et al.*, 1995]. This does not mean, however, that recombination at some position between the 5' and 3' ends of the haplotype does not have a role to play in producing haplotypes, but that perhaps it is occurring at too low a rate to be detected by the IAM test. The fact that full α-haplotypes exist indicates that the cluster does not contain a recombination hotspot, as at β [Chakravarti *et al.*, 1986], but it is still possible that recombination occurs non-randomly in relatively "cold regions" of the genome.

Within populations it does seem that recombination is the most important mechanism of producing haplotype diversity. The more stable nature of the 3' end, and the association of 5' ends with a variety of these 3' ends is indicative of a

tendency for recombination to occur within certain regions of the 30 kb gene cluster; from consideration of the 5' and 3' ends of the haplotype, the region of preferential recombination appears to be in the vicinity of the IZHVR, although it is difficult to pinpoint the region any further on the basis of population haplotype data. Interestingly, many of the endpoints of illegitimate gene deletions occur between the ζ genes [Nicholls *et al.*, 1987], again suggesting a predisposition of this region for recombination events, perhaps due to a predisposing chromatin structure. Repeated sequences have been implicated in a number of genomic recombination hotspots, for example [Chakravarti *et al.*, 1986; Steinmetz *et al.*, 1986](Chapter 1), but it is debatable whether they are *directly* responsible for the recombinogenic activity of a region. Nevertheless, the presence of the IZHVR within a region of comparatively higher recombination may be significant.

As a consequence of the 5' and 3' end recombination of the α-haplotype, it is likely that the evolutionary history of a full haplotype will be complicated and dependent on population context. The apparently less internally recombinogenic 3' end suggests that haplotype groups represent a more valid structure from an evolutionary point of view and are potentially easier to investigate at the sequence level. This finer resolution of the α-haplotype is precisely what the PCR haplotype method offers: an opportunity of studying the fine-level structure of the haplotype.

# Chapter 3

## Sequence Variation at the α-Globin Gene Cluster

### The Method of Direct Sequencing the α-Globin Complex

Sequence variation at the α-globin locus was determined by direct analysis of PCR products from the haplotyping protocol detailed in Chapter 2. There are a number of possible ways of characterising mutations within PCR products, many of which have the advantages of being relatively inexpensive and rapid. These indirect methods give information about the presence of mutations but do not always indicate the nature and location of mutations [Grompe, 1993]. For this study, the precise nature and location of mutations are required, so the method of choice was sequencing.

Subcloning into sequencing vectors is time-consuming and therefore unrealistic for processing a large number of samples. Also, errors made by the polymerase during the PCR lead to overestimation and false identification of polymorphism, as noted in a study of sequence variation at the β-globin locus [Fullerton, 1995]. Although Vent polymerase, used in the PCR haplotyping protocol (Chapter 2) has a higher fidelity than Taq, the identification of even rare polymerase errors would be unacceptable. Sequencing of the whole population of PCR products in direct sequencing effectively eliminates the problem; only an error occurring at the very early stages of a PCR could reach sufficient frequency to affect the sequence obtained, although such an error is unlikely at this stage.

Direct sequencing protocols utilise the chain-termination method [Sanger *et al.*, 1977]. This involves *in vitro* synthesis of DNA strands by a polymerase, initiated at one site where an oligonucleotide primer anneals to the template. The method works

best with a single-stranded (ss) template [Sambrook *et al.*, 1989]. The polymerisation reaction is terminated by incorporation of nucleotide analogues that will not support continued DNA elongation. These analogues are usually 2', 3'-dideoxynucleoside 5'-triphosphates (ddNTPs) which lack the 3' OH group necessary for chain elongation. With the proper mixture of dNTPs and ddNTPs, the polymerase reaction is terminated in a fraction of the population of chains and at each site along the template, giving a nested set of fragments.

In order to ensure that this method is amenable to a survey of variation within human populations, there are two requirements to be met: the preparation of ss template must be rapid and effective, and sequencing must be of high-quality for confident detection of sequence polymorphism.

**Design of Sequence Primers**

Sequencing primers were designed in accordance with the requirements for PCR primers (Chapter 2) and are listed in Appendix 3. One primer was designed for every 250-300 bp of sequence to be studied. This represents the amount easily readable on the sequencing gels used (Appendix 1, A1.9).

**Methods to produce single-stranded template**

Single-stranded (ss) template can be produced in a number of ways. PCRs were performed with one kinased primer (Appendix 1, A1.3.1 & A1.7.2) and λ-exonuclease digestion was used to produce ss DNA (the enzyme attacks the 5' terminal phosphate group of the labelled strand)[Higuchi and Ochman, 1989]. This method is relatively time-consuming, however, and did not always result in high

yields of template, especially following the necessary purification steps of ss DNA prior to use in sequencing reactions.

The use of streptavidin-coated magnetic beads (Dynabeads; Dynal, UK) represented the most rapid and effective method, by incorporating the production of ss DNA and its purification into one protocol and by ensuring minimal loss of template during these stages. PCRs were performed using one primer end-labelled with biotin (Oswel) and the products were immobilised on the beads in the presence of a 2M salt solution (Appendix 1: A1.7.3). Following removal of the unbiotinylated strand by treatment with 0.1M NaOH, the ss template, still attached to the beads, was used directly in sequencing reactions (Appendix 1, A1.7.3).

**The Achievement of High-Quality Sequencing**

PCR Sequencing

PCR sequencing is an attractive alternative to other methods of direct sequencing because it requires very little template (~100 ng), can use double-stranded templates, and is potentially more efficient when sequencing GC-rich templates because of the use of high temperature reactions. In this study, Vent (exo-), a derivative of Vent polymerase with 3'-5' exonuclease activity removed, was used in a commercially available kit (CircumVent, New England Biolabs). Both ss DNA (Dynabead templates) and double-stranded DNA templates (gene-cleaned PCR products; Appendix 1, A1.5.2) were used; the latter were sequenced with two protocols: $^{35}S$ dATP incorporation and using a $^{32}P$ end-labelled sequencing primer (Appendix 1, A1.3.1). $^{35}S$ incorporation has the advantage of being more "user-friendly" and the reactions could be stored for longer, although the method does require use of more

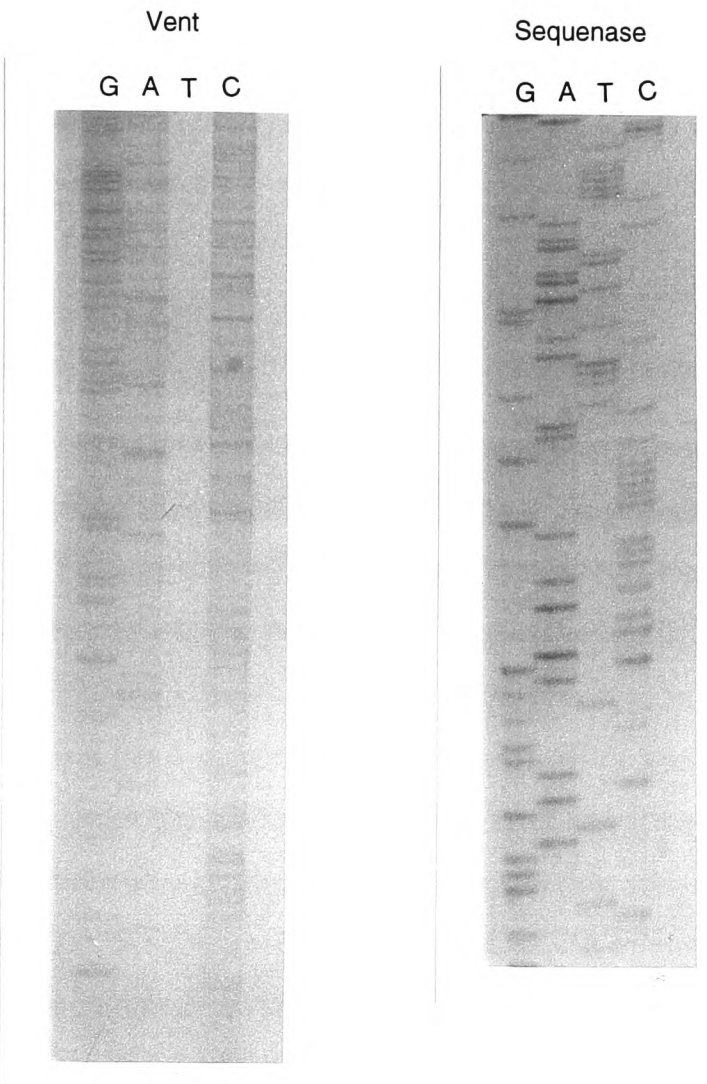
$^{35}S$ than other protocols (20μCi per reaction versus 5μCi for Sequenase sequencing- see next section).

The quality of sequence generated by this method was generally inferior to that obtained by the use of Sequenase (Figure 3.1); as with standard PCRs, optimisation of cycling conditions is required for each sequencing primer used in the PCR sequencing protocol. Since a large number of sequencing primers were used, it was unrealistic to optimise all the reactions. Therefore, this method was primarily used for confirmation of sequence identity of new PCR products, rather than for mutation detection.

Sequencing using the "Sequenase "Enzyme

This method, combined with ss DNA template production using Dynabeads, resulted in the highest quality sequencing. Sequenase (United States Biochemical) is a modified form of bacteriophage T7 DNA polymerase with the 3'-5' exonuclease activity removed. ~1μg of DNA is required and insufficient template can result in reduced labelling close to the primer, loss of sequence information and very weak sequence signals. The PCRs designed for haplotyping were therefore also optimised for yield, as described in Chapter 2, which required the use of 50 pmoles of each PCR primer in the amplification reaction. This amount of primer could potentially interfere with binding of PCR product to the Dynabeads or with the sequencing reaction. Therefore, in initial experiments, CLB-6 columns were used to remove the primers at the end of the PCR (Appendix 1, A1.7.1). However, it was found that the quality of sequence was not improved over that obtained with reactions not employing this step, and given that this step is time-consuming when processing large numbers of samples, it was subsequently omitted.

Figure 3.1: Sequencing using Sequenase Enzyme versus PCR Sequencing




Both sequences are of the Acc region using the primer A.SF1 (Appendix 3). Sequence obtained with Vent(exo-) is the result of an unoptimised reaction using conditions recommended by the manufacturer (20 cycles of 20 seconds at each of 95°C, 55°C, and 72°C); improvement is possible by increasing denaturation time (to 30 seconds) and alteration of annealing temperature. However, some problems of background signal and uneven band intensity (as seen here) remain, making detection of polymorphisms unfeasible with this method. The Sequenase sequence was obtained without modification to recommended protocol (USB).

The Sequenase protocol was used according to manufacturer's instructions with a number of minor alterations. In order to slow the reaction down and ensure that sequence close to the primer was not lost, the extension of the Sequenase enzyme (labelling step) was performed on ice for three minutes, which was especially important if the amount of template was limiting. To increase the strength of the sequencing signal without compromising sequence close to the primer, additional dNTPs were routinely added to the termination mixes (extending mix, USB).

*Alterations to Sequenase Protocol: Problems of GC-rich Template*

A number of other modifications of protocol were attempted in response to occasional difficulties due to the GC-rich nature of the α-globin complex. There are two types of problem associated with sequencing GC-rich templates. Firstly, access to the template by sequencing primer and progression of the polymerase along the template can be inhibited by secondary structure, resulting in sequence failures and pauses (visible as a band present in all four termination lanes of one sample on a sequencing gel). Secondly, GC-rich templates may not always be fully denatured during electrophoresis, resulting in bands appearing closer together than they normally would (forming a compression). In this study, compressions and sequence pausing were encountered; secondary structure was also likely to be responsible for blurred sequencing observed following poly (A) tracts at the end of *Alu* repeats.

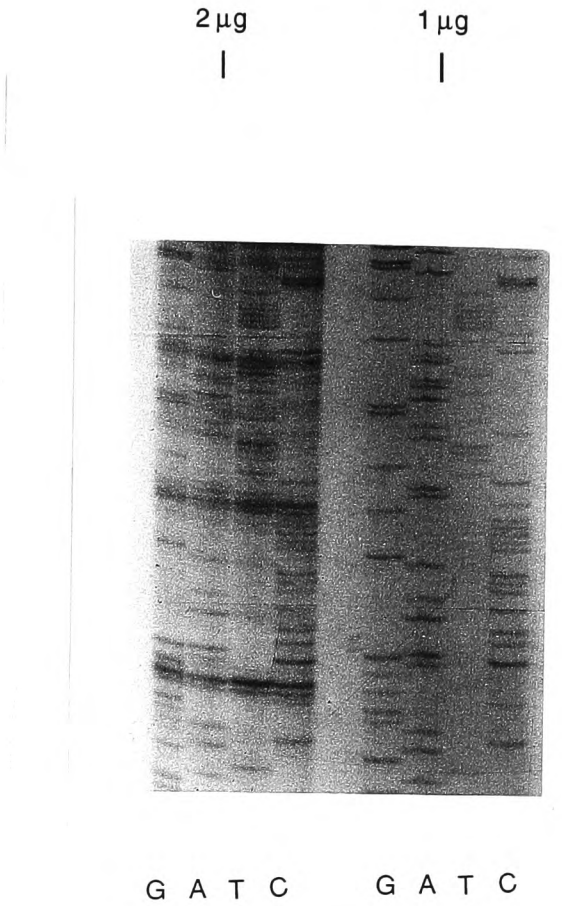
A number of additional steps and modifications to protocol were tested, including the use of single-strand DNA binding protein (SSB protein, USB), use of a high temperature "stringency wash" of Dynabeads and annealed template prior to the labelling step [Thomas *et al.*, 1994], and performing extending reactions at 70°C (instead of 37-42°C) for 10 mins with Taq polymerase [Beck *et al.*, 1993]. All these methods failed to improve sequence quality sufficiently in difficult regions.

Substituting the analogue dITP for dGTP, which does not form such stable secondary structure (I-C bonds are formed by only two hydrogen bonds, unlike G-C bonds), and PCR sequencing using high temperatures also resulted in poor sequence quality.

The approach finally adopted was relatively simple; it was found that increasing the amount of template used in a sequencing reaction increased sequence signal but also tended to increase the likelihood of compressions and background signal (Figure 3.2). In fact, the boundary between having enough template to achieve high-quality sequencing and too much template so as to induce frequent compression problems in the sequence, was found to be quite narrow. Therefore, for regions prone to compressions, the estimated amount of template used in the reactions was kept close to 1μg, or reduced (to a minimum of ~0.5μg) to produce fainter but clearer sequence. Any persistent problems of sequence ambiguity were resolved by sequencing the complementary strand, although this rarely proved necessary. Only one strand was routinely sequenced since this study is not of new, unknown sequence. Results could therefore be checked against the published sequences to check for any errors; singleton polymorphisms, occurring in non-problematic regions, were distinguished from errors by repeating the sequencing reactions.

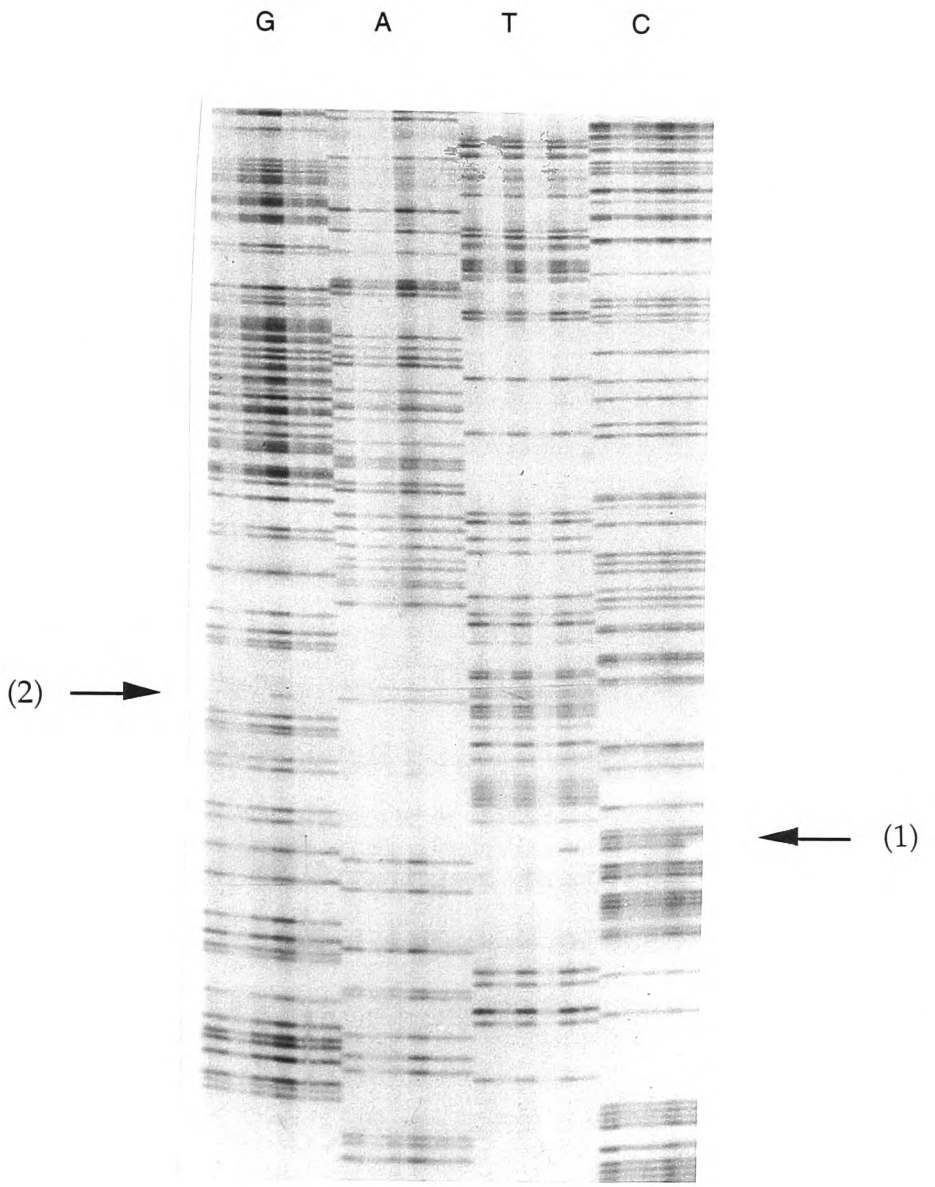
Sequences were run on denaturing polyacrylamide gels as described in Appendix 1. Six to twelve samples were loaded on one gel, with the same termination reaction from different samples run together (Figure 3.3). This enabled the rapid identification of polymorphisms without the need to read sequences for all samples. Despite the fact that diploid sequences were determined, polymorphisms could be identified within any one individual without ambiguity and their heterozygote or homozygote state determined (Figure 3.3).

Figure 3.2: Effect of Excess Template on Sequencing Quality




Sequence of the Acc region obtained using A.SF1 sequencing primer (Appendix 3) and using Sequenase enzyme (USB). Halving the amount of template placed in sequencing reaction (1 µg rather than 2 µg) reduces compressions.

Figure 3.3: Identification of Polymorphisms by Sequencing




Sequence of α Pst region (sequencing primer αPSF2; Appendix 3). Two polymorphic sites are present: site (1) is present homozygously in one individual, evidenced from the missing band in the termination lane corresponding to the other samples. Site (2) is present heterozygously in one individual since a band is present at the same position in two termination lanes.

## Sequence Variation of the Unique Sequence Regions of the α-Globin Gene Cluster

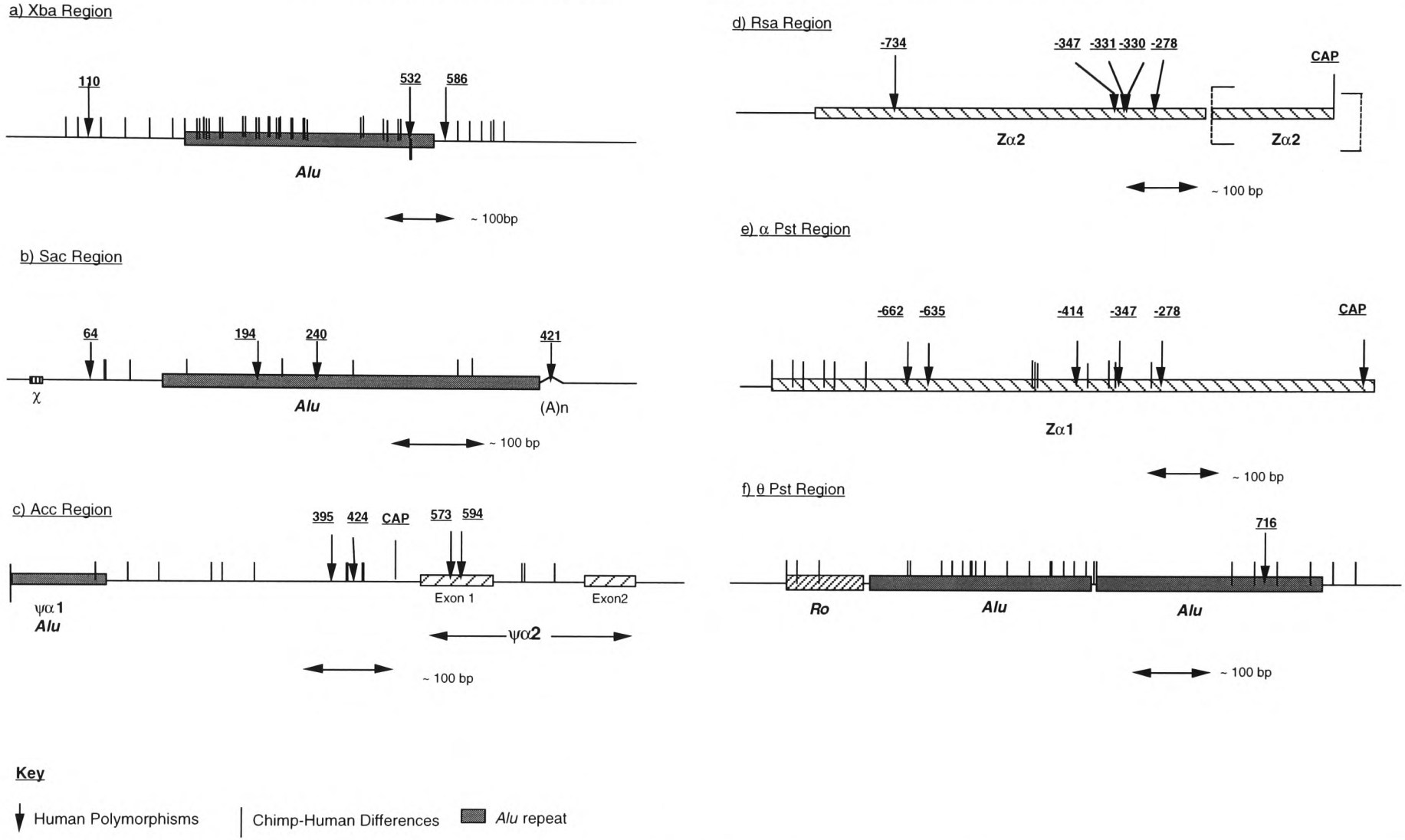
### Identification of Sequence Polymorphism

In this chapter, only those regions of the α-haplotype which do not contain repetitive sequences are considered. These include the Xba, Sac, Acc, Rsa, α Pst and θ Pst regions (Figure 3.4); each region is named after the PCR amplification containing a specific RFLP site (Chapter 2), and thus each is of a different size: from 561 bp (Sac region) to 910 bp (Xba region) (Table 2.3; Chapter 2), and contains within its sequence the RFLP site (which will be referred to in italics). These six regions were sequenced in a sample of 19 human individuals chosen to represent common haplotypes from as many different populations as possible (haplotypes had been determined in previous Southern blotting surveys; D R Higgs, personal communication; J J Martinson, personal communication; T Jenkins, personal communication; [Roberts-Thomson *et al.*, in press]). The samples sequenced, and their haplotype status and geographical origin are shown in Table 3.1. Individuals heterozygous for different haplotypes were chosen to maximise the number of haplotypes studied. Overall, some haplotypes are represented more frequently within the sample than others; this is because the frequencies of different haplotypes vary in world populations. For example, there are seven Ia chromosomes, from a number of geographical locales, versus three group VII chromosomes; the Ia is a very common haplotype, present in almost all world populations, whereas group VII haplotypes are confined primarily to Africa [Higgs *et al.*, 1986].

In addition, two chimpanzee[1] chromosomes were sequenced to determine the ancestral states of human polymorphisms. The human and chimpanzee sequences

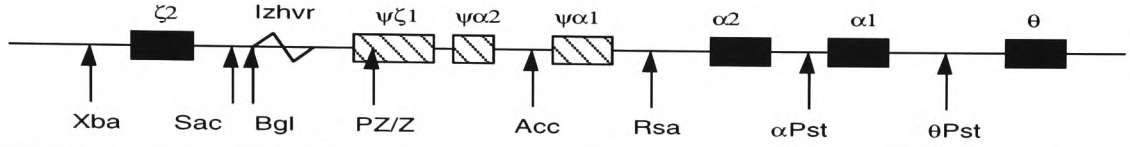
[1]The chimpanzee was from a colony derived from a group captured in Sierra Leone, and its DNA was a gift from R.Bontrop, Primate Centre, Rijswijk, The Netherlands.

Figure 3.4: Regions Sequenced Showing Positions of Polymorphisms and Interspecific Sequence Differences




Human polymorphic sites are numbered relative to the α-genes' Cap sites in the Rsa (brackets indicate region not sequenced but containing the α2 gene Cap site) and α Pst regions; in all other regions numbering refers to the position relative to 5' end of the amplified region.

Table 3.1: Polymorphic Sites Within the Regions Sequenced in the α-Globin Cluster




| Xba | | | Sac | | | | Acc | | | | Rsa | | | | | Alpha Pst | | | | | Theta Pst | Sample | Haplotype | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 110 | 532 | 586 | 64 | 194 | 240 | 421 | 395 | 424 | 573 | 594 | 734 | 347 | 331 | 330 | 278 | 662 | 635 | 414 | 347 | 278 | 716 | | | |
| G | C | A | C | C | C | 15 | T | T | G | G | C | G | C | G | C | C | A | G | G | C | T | | | |
| - | - | - | - | -/G | -/T | - | - | - | - | - | -/T | - | - | - | - | - | - | -/C | -/A | - | -/C | W Samoa | Ia/IVa | ++-M PZ++--/+--S PZ+-++ |
| - | - | -/G | - | - | -/T | - | - | - | - | - | -/T | - | - | - | - | - | - | -/C | -/A | - | -/C | Britain | Ia/IVc | ++-M PZ++--/-+-L PZ+-++ |
| - | - | -/G | - | - | -/T | - | -/C | - | - | - | -/T | - | - | - | T/T | - | - | - | - | -/T | - | Britain | Ia/VIa | ++-M PZ++--/-+-L PZ---- |
| - | - | -/G | - | - | -/T | 14/15 | - | - | - | - | -/T | -/A | -/G | -/A | - | - | - | - | - | - | - | S Arabia | Ia/IIa | ++-M PZ++--/-+-L PZ+--- |
| - | - | -/G | - | - | -/T | - | - | -/G | - | - | -/T | -/A | -/G | -/A | - | - | - | - | - | - | - | Iceland | Ia/IIa | ++-M PZ++--/-+-L PZ+--- |
| - | - | - | - | - | T/T | 14/15 | - | - | - | - | -/T | -/A | -/G | -/A | - | - | - | -/- | - | - | - | Senegal | Ia/IIc | ++-M PZ++--/++-M PZ+--- |
| -/A | - | - | - | - | - | 14/15 | -/C | - | - | - | - | A/A | - | - | - | - | - | - | - | - | - | Senegal | IIc/VIb | ++-M PZ+---/++-L PZ---- |
| - | - | - | - | -/G | -/T | 14/15 | - | - | - | A/A | - | A/A | - | - | - | - | - | - | - | - | - | Philippine | IIc/IIe | ++-M PZ+---/+--S PZ+--- |
| - | - | -/G | - | -/G | - | - | -/C | - | - | - | -/T | A/A | - | - | - | - | - | - | - | - | - | Austr Ab | Ia/IInovel | ++-M PZ++--/---M* PZ+--- |
| - | - | - | - | G/G | - | - | -/C | - | - | - | - | -/A | - | - | - | - | - | - | -/A | - | - | Senegal | IIe/IIId | +--S PZ+---/+--M Z---- |
| - | - | -/G | - | -/G | - | - | -/C | - | - | - | - | -/A | - | - | - | - | - | - | - | - | - | Brunei | IIb/IIIa | ++-L PZ+---/--+M Z---- |
| - | - | G/G | - | G/G | - | - | C/C | - | - | - | - | A/A | - | - | - | - | - | - | - | - | - | Senegal | IIIa/IIIf | --+M Z----/---M Z---- |
| - | - | -/G | - | G/G | - | - | C/C | - | - | - | - | A/A | - | - | - | - | - | - | - | - | - | Iceland | IIIa/IIIb | --+M Z----/+-+M Z---- |
| - | - | - | - | -/G | -/T | - | - | - | - | - | - | -/A | -/G | -/A | - | T/T | - | - | A/A | - | - | Senegal | VIIc/IIc | +--M Z+---/+++-M PZ+--- |
| - | -/G | -/G | -/G | -/G | - | - | - | - | - | - | - | -/A | -/G | -/A | - | - | - | -/C | -/A | - | -/C | Bantu | IVa/VIIi | +--S PZ+-++/-+-S Z+--- |
| - | - | - | - | G/G | - | 14/15 | - | - | -/A | - | - | - | - | - | - | - | - | - | -/A | - | -/C | San | Vc/IIg | +--S PZ+--+/+--M PZ+--- |
| - | - | -/G | - | G/G | - | - | -/C | - | - | - | - | -/A | -/G | -/A | - | - | -/G | - | -/A | - | -/C | Bantu | Vc/IIIf | +--S PZ+--+/---M Z---- |
| - | - | G/G | - | G/G | - | - | - | - | - | - | - | -/A | -/G | -/A | - | - | - | -/C | A/A | - | C/C | N Guinea H | Vd/IVb | ---S PZ+--+/---S PZ+-++ |
| - | - | - | - | -/G | -/T | - | -/C | - | - | A/A | - | -/A | - | - | - | - | - | - | - | - | -/C | Ponape | Vc/VInovel | +--S PZ+--+/++-S PZ---- |
| G/G | C/C | G/G | C/C | G/G | T/T | 14/14 | T/T | T/T | G/G | G/G | C/C | n/a | n/a | n/a | n/a | C/C | A/A | G/G | G/G | C/C | T/T | Chimp | II/II | --?? PZ+---/--?? PZ+--- |

Polymorphic sites in the six regions sequenced for 19 heterozygous human samples. Polymorphic sites in the Xba, Sac, Acc and θ Pst regions are named according to region and numbered according to position within the PCR fragment (from 5' end). Polymorphic sites in the Rsa and α Pst regions are named according to region and numbered according to position 5' to the Cap site of the α-genes. RFLP sites are shaded. Consensus sites are shown in dark border, according to the known sequence (D. R. Higgs, pers comm). Bases identical to the consensus are indicated by -. Chimp sequence at each human polymorphic site is also shown; Rsa sequence for the chimp is incomplete with sequence information available only for the RFLP site. M* indicates unusually sized M allele (intermediate between M and L classes; refer to Chapter 3) in Australian Aborigine sample.

can be found in Appendix 4. Amplification using human primers proved possible because the chimpanzee α-globin gene complex is similar in structure to the human complex [Zimmer *et al.*, 1980]. Use of the Rsa region primers, 5' to the α2 gene, resulted in two amplification products, consistent with the observation that the predominant number of α-genes in the chimpanzee is three, not two as in humans [Zimmer *et al.*, 1980], and that it is the α2 gene that has been duplicated in the chimpanzee [Liebhaber and Begley, 1983]. In order to obtain the sequence at the *Rsa I* RFLP site, the band similar in size to that obtained exclusively in humans was cut from agarose, purified and sequenced (Appendix 1, A1.5.3).

Analysis of the six unique-sequence regions resulted in 4,940 bp of sequence information per chromosome (total ~190 kb, 38 human chromosomes). Twenty-two polymorphic sites were identified, inclusive of the six known RFLP sites within the regions sequenced (Table 3.1; RFLP sites shaded). Of these polymorphic sites, twenty-one are nucleotide substitutions, and one is an insertion/deletion polymorphism. The latter is present in the poly (A) tail of an *Alu* repeat within the Sac region, with either 14 or 15 As found in succession (Figure 3.4b). Similar polymorphisms have been reported at other *Alu* repeats in the human genome [Economou *et al.*, 1990; Orita *et al.*, 1990], and are likely to be caused by a replication-slippage mechanism [Levinson and Gutman, 1987].

Approximately 70% of the nucleotide substitutions were transitions; that is, purine to purine or pyrimidine to pyrimidine base pair changes. From studies of neutrally evolving pseudogene sequences, it has been estimated that transitions typically represent ~60% of substitutions [Li and Graur, 1991]; transitions are more frequent than transversions partly because CpG dinucleotides, which are common in the α-globin cluster [Perutz, 1990], can behave as mutation hotspots [Coulondre *et al.*, 1978]. Cytosines occurring in the dinucleotide are often methylated in human and

vertebrate DNA to give 5-methylcytosine; spontaneous deamination of 5-methylcytosine occurs over an evolutionary time-scale to generate thymine, and an excess of C to T transitions. If the transitions present at CpG dinucleotides are discounted in the α-globin sequenced regions, the proportion of transitions falls to ~57% (seven transitions; Table 3.1), which is close to the corrected proportion (without CpG) for pseudogene sequences of ~54% [Li and Graur, 1991].

Necessarily, all (+) alleles at the RFLP sites contained within the regions studied have the same sequence. The chimpanzee possesses (-) alleles at all RFLP sites apart from *Acc I* (Table 3.1) and so the (+) alleles at the *Xba I, Sac I, Rsa I, α Pst I and θ Pst I* sites are likely to have arisen in humans since divergence from chimpanzees (~ 5 myr ago, see Chapter 1); the combination of sites in the chimpanzee indicates that the ancestral haplotype group is probably the II. Two of the (+) alleles in humans, at the *Rsa I* and *θ Pst I* sites, have been produced by a transition at a CpG dinucleotide (Table 3.1). However, given that these (+) alleles are found on a limited number of haplotype backgrounds (see Chapters 1 and 2), they most likely represent single mutations, rather than multiple mutations at a CpG hotspot. All the (-) RFLP alleles possess the same sequence in both the chimp and human samples, regardless of geographical origin (Table 3.1), even though a site could potentially be lost by numerous mutations within the recognition sequence.

Excluding the six RFLP sites, nine of the remaining sixteen polymorphisms identified here are singletons (i.e. found only once in the sample), with six of these singletons being found in African samples. This conforms with the higher level of polymorphism described in African populations from α-haplotypes (see Chapter 2). Seven polymorphisms are common, being found on more than one chromosome in the sample; one of these, site Acc: 594, has an allele only present in two homozygous individuals from the Philippines and Micronesia, suggesting that it has a limited

geographical distribution. It is possible that some or all of the singleton polymorphisms identified here may actually be population or regionally specific, although the sampling strategy employed would not reveal this. Excluding Acc: 594, the remaining six polymorphisms are found on many haplotype backgrounds and from a variety of geographical locales. These common polymorphisms, however, are not spread over all the regions studied: two polymorphisms were found within the Sac region (Sac: 240 and Sac: 421), three in the Rsa region (Rsa: 347, Rsa: 331, Rsa: 330) and one in the αPst region (αPst: 347). Other regions, by comparison, show no common polymorphism other than the RFLP or no extra polymorphism at all, such as in the Xba and θ Pst regions, respectively.

**Sequence Polymorphism and Divergence at the α-Gene Cluster**

Regional Differences in Sequence Polymorphism

This survey of sequence variation at the α-haplotype investigated regions which are less than 1 kb in length and which were already known to contain one commonly polymorphic site (the RFLP); the number of other polymorphic sites likely to be present depends on the variability of the α-globin cluster. This variability, or polymorphism, is dependent upon the underlying mutation rate of the locus, assuming neutrality and the infinite-alleles model ($\theta = 4N\mu$; Equation 2.1, Chapter 2). It is possible to estimate the variability of a locus on the basis of both restriction enzyme and sequence data. Since the sequence data presented here are not random in terms of the particular regions of the α-globin gene cluster studied and the samples sequenced, both types of data were used in the estimation of diversity.

Assuming neutrality, sequence diversity (θ) in a population can be estimated from restriction enzyme data from Equations 3.1 and 3.2:

$$\theta_{R1} = (k_4 + k_6)/[(8m_4 + 12m_6) \ln n] \quad \textbf{(3.1)}$$

where m = total number of cleavage sites for all restriction enzymes (4 and 6 bp cutters), k = number of polymorphic restriction sites (4 and 6 bp cutters), and n = number of chromosomes [Ewens *et al.*, 1981].

$$\theta_{R2} = (k_4 + k_6)/AR \quad \textbf{(3.2)}$$

where $s_r$ = number of polymorphic restriction enzyme sites in the sample, and

$$A = \sum_{i=1}^{n-1} (1/i) \qquad \text{and} \qquad R = \sum_{i=1}^{s} m_i r_i$$

where s = number of enzymes used, $r_i$ = nucleotides in the recognition sequence of the enzyme, $m_i$ = number of restriction sites and n is as previously defined [Nei and Tajima, 1981].

Both calculations were performed in order to estimate sequence diversity at the α-cluster, using the data obtained by restriction enzyme digestion of the Kenyan Luo in the previous chapter, and the α-haplotype data of the South African Bantu [Ramsay and Jenkins, 1988] and Senegalese Mandenka [Martinson *et al.*, 1995]. The use of these data assumes that the only RFLPs present with the enzymes used are those already defined by the α-haplotype. Given the extensive survey of RFLP sites originally performed to define α-haplotypes [Higgs *et al.*, 1986], this assumption is most probably valid. Only the 6-cutter enzymes *Sac I, Acc I* and *Pst I* and the 4-cutter *Rsa I* were considered; *Bgl I* is an 11-cutter and thus cannot be used in the above equations, and *Xba I* was not included because the haplotyping site is positioned

some distance upstream of the α-cluster, where sequence information, and therefore the *Xba I* digestion pattern, is not presently available. The results are shown in Table 3.2.

---

Table 3.2: Using Restriction Enzyme Data to Calculate Diversity at the α-Globin Gene Cluster

| Population | N | $(k_4 + k_6)$ | $\theta_{R1}$ | $\theta_{R2}$ |
|---|---|---|---|---|
| Bantu | 36 | 6 | 0.0022 | 0.0038 |
| Mandenka | 210 | 4 | 0.0009 | 0.0018 |
| Luo | 32 | 5 | 0.0019 | 0.0033 |

Calculations were performed using the following total number of restriction sites for each enzyme, over a ~ 30 kb region from the Cap site of the ζ2 gene to the θ gene: *Sac I*, 10; *Rsa I*, 28; *Pst I*, 30; *Acc I*, 4. N = number of chromosomes. For determination of $(k_4 + k_6)$, refer to Table 2.4 (Chapter 2).

---

Estimation of sequence polymorphism ($4N\mu$) is also possible by considering the number of segregating (polymorphic) sites in the sample from Equation 3.3:

$$L = \theta = p_n / A \qquad \textbf{(3.3)}$$

where $p_n = s_n/m_t$, $s_n$ = number of polymorphic sites per sequence, $m_t$ = total number of nucleotides examined, and A is as previously defined. L can be considered as a measure of heterozygosity [Watterson, 1975]. This calculation was performed using the data in Table 3.1; that is, of samples from a number of different populations (Table 3.3).

Table 3.3: Diversity Calculated from Sequence Data

| Region | $m_t$ (bp) | $s_n$ | L | D (%) |
|---|---|---|---|---|
| Xba | 910 | 3 | 0.00075 | 6 |
| Sac | 561 | 3 | 0.00128 | 2.14 |
| Acc | 826 | 4 | 0.00115 | 1.6 |
| Rsa | 817 | 5 | 0.00146 | N/A |
| α Pst | 938 | 5 | 0.00127 | 2.3 |
| θ Pst | 888 | 1 | 0.00027 | 3.4 |
| | | Average | 0.00103 | 3.1 |

N/A= information not available. D = divergence between human and chimp sequences.

Averaged over all six regions studied, including only point substitutions (i.e. 21 sites), L is equal to 0.00103 (Table 3.3). This is a little lower in value than the estimates based upon RFLP data (overall average of 0.0023; Table 3.2) but not substantially so. The higher estimate of $\theta_R$ from RFLP data may be due to the use of African populations which have higher levels of variation (Chapter 2). The average value of $\theta_{R2}$ (0.003, Table 3.2) corresponds fairly well with a similar estimate from β-globin ($\theta_{R2}$, average 0.002) [Kazazian *et al.*, 1983], and L is not substantially different from estimates obtained from β-globin sequences (0.0014) [Fullerton *et al.*, 1994], and from a variety of cDNA and genomic sequences (0.00076) [Li and Sadler, 1991]. Thus, overall, sequence diversity at the α-globin gene cluster is comparable to other regions of the genome, with approximately 1 in 700 bp varying between any two randomly chosen sequences in the sample set. It is not surprising, therefore, that few extra polymorphisms were detected within the ~800-900 bp windows sequenced in this study. The more surprising aspect of the sequencing data is that certain regions contain more variation than might be expected on the basis of the sequence diversity estimations.

From values of L calculated for each region (Table 3.3), it is clear that levels of polymorphism vary across the α-globin gene complex. As indicated previously, there are just three regions which contain commonly polymorphic sites (Sac, Rsa and α Pst) (Table 3.1). The Acc region has three polymorphisms other than the RFLP (Table 3.1), giving a relatively high value of L; this calculation, however, does not incorporate frequencies of the individual polymorphisms and the extra polymorphisms in the Acc region were not as common in the sample as those in the Sac, Rsa and α Pst regions. The Xba and θ Pst regions have low values of L, corresponding to 1 in ~1-3 kb varying between two randomly chosen sequences.

Chimpanzee-Human Sequence Divergence

In order to explain these regional differences in polymorphism, it is helpful to compare the human and chimpanzee sequences. Neutral theory predicts that levels of within-species polymorphism should be correlated with levels of interspecific divergence, since both the rate of fixation of nucleotide substitutions and the level of neutral polymorphic variation are functions of the same neutral mutation rate [Kimura, 1983](Chapter 1). The positions of the chimpanzee-human sequence differences (both nucleotide substitutions and insertion/deletions) are shown in Figure 3.4 (a-f). Aligned chimpanzee and human sequences are listed in Appendix 4.

The overall divergence of human and chimpanzee, for the regions studied, is 3.1% (Table 3.3), which corresponds to a mutation rate of $3.1 \times 10^{-9}$ substitutions per site per year, assuming a 5 myr divergence time. This is presumably a neutral mutation rate since the regions studied contain primarily non-coding sequence (Figure 3.4, a-f). By comparison, the β-globin locus has a divergence of 1.6% (estimated from the 5' flanking region of the β-globin gene) [Savatier *et al.*, 1985], corresponding to $1.6 \times 10^{-9}$ substitutions per site per year. The synonymous substitution rates in α- and β-globin

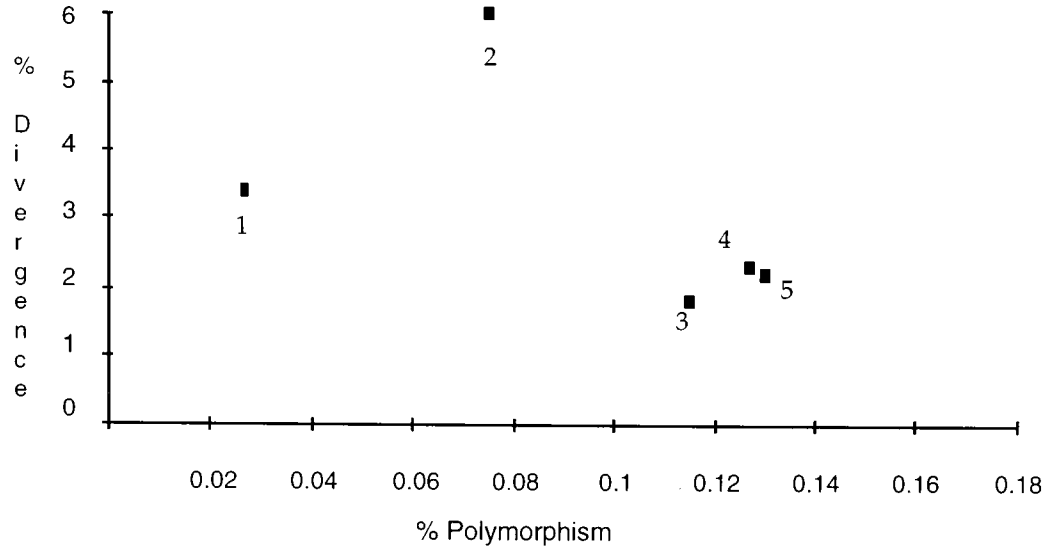
gene sequences, based on comparisons of human and rodent genes, are $5.14 \times 10^{-9}$ and $3.05 \times 10^{-9}$ respectively; these values are not too dissimilar to those based on the human-chimpanzee comparisons above, and represent intermediate levels of mutation compared to other loci in the genome [Li and Graur, 1991]. The reason for a slightly higher mutation rate in the α-globin cluster, compared to β, is not clear, although a similar observation has been made previously [Hardison and Gelinas, 1986]. It has been suggested that mutation rate corresponds to the GC-content of genes and their flanking regions; GC-rich regions of the genome replicate early and have corresponding low mutation rates [Wolfe *et al.*, 1989]. However, if this were the explanation for the differences between the two globin gene clusters, the later-replicating β-globin gene cluster should have the higher mutation rate.

The average estimate of the α-globin mutation rate is, however, elevated by the presence of very high divergence (6%) in the Xba region and relatively high divergence in the θ Pst region (3.4%). Discounting these two regions, the average divergence at the α-gene cluster becomes 2.0%, which is not substantially higher than the approximations based on the β-globin gene cluster [Savatier *et al.*, 1985] and also for non-coding DNA in general [Sawada *et al.*, 1985]. Curiously, the highly divergent Xba and θ Pst regions show low levels of polymorphism within humans, compared to the other regions studied (Table 3.3), so that the predicted correlation between polymorphism and divergence from neutral theory is not present (Figure 3.5).

**Evolution of *Alu* Repeats in the α-Globin Gene Cluster**

The Xba and θ Pst regions both contain *Alu* repeats: the former contains one full-length *Alu* fused at its 3' end with a partial *Alu* sequence (Figure 3.4a), and the latter region with very closely positioned duplicated full-length repeats (Figure 3.4e). *Alu* repeats consist of a right and left half which are related in sequence [Ullu and

Figure 3.5: Chart Showing Relationship between Divergence and Polymorphism




1=θ Pst region; 2= Xba; 3= Acc; 4=α Pst; 5= Sac. There is no significant correlation between polymorphism and divergence (Fisher's r to z, correlation = - 0.544, p=0.3884).

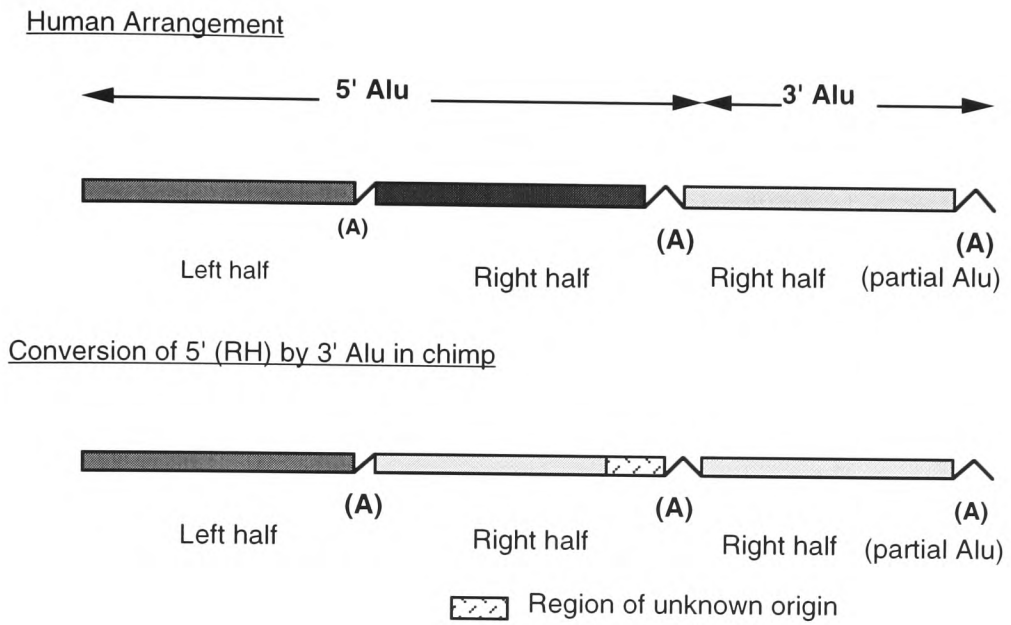
Tschudi, 1984]; the partial *Alu* in the Xba region has only the right half, possibly representing the remains of an *in situ* deletion of the left half of a full *Alu*. Comparison with a consensus human *Alu* sequence [Britten *et al.*, 1988] reveals that the Xba and θ Pst *Alus* belong to a common subfamily (Group II [Britten *et al.*, 1988]; equivalent to the S-type family [Jurka and Smith, 1988]) which is believed to have inserted in the primate lineage some time before the chimpanzee-human divergence. The age of these *Alus*, as well as the sequence identity flanking the *Alu* repeats in chimpanzees and humans, confirms that the chimpanzee Xba and θ Pst *Alus* are orthologous to those in humans; thus, the reason for the high divergence in these repeats is not due to inadvertent amplification of a different *Alu* repeat in the chimpanzee.

*Alu* sequences evolve at a similar rate to other non-coding DNA and have been reported to show human-chimpanzee divergences of between 1.4% and 2.2% ([Britten *et al.*, 1988; Sawada *et al.*, 1985], discounting CpG dinucleotides which mutate rapidly in *Alu* sequences [Britten *et al.*, 1988]. Inspection of the Xba and θ Pst *Alus* reveals that mutations at CpG dinucleotides have increased observed human-chimpanzee divergence; with these mutations removed, divergence decreases substantially. However, in two sections of the *Alu* repeats, the right half of 5' Xba *Alu* and the left half of the 5' θ Pst *Alu*, the reduced divergences are still high, and beyond the reported range [Britten *et al.*, 1988; Sawada *et al.*, 1985](Table 3.4).

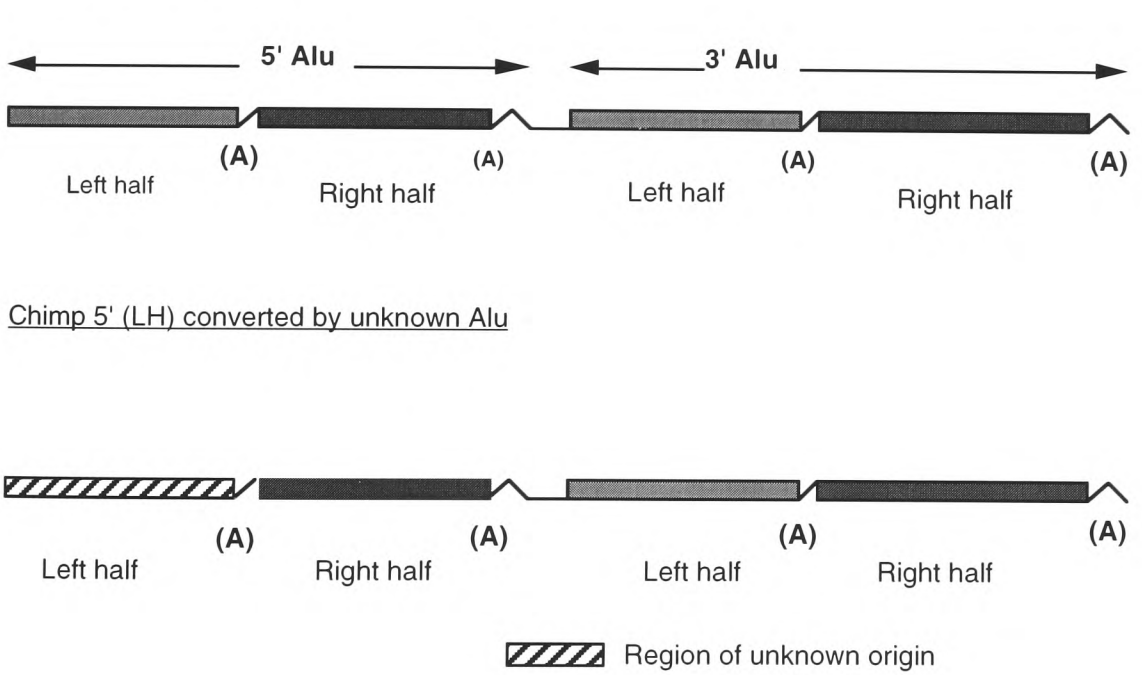
The explanation for these localised regions of high divergence is gene conversion, evident from consideration of the *Alu* sequences. The chimpanzee Xba (RH) *Alu* has converted almost completely to the sequence of the fused, partial 3' *Alu* (Figure 3.6A); the θ Pst 5' *Alu* sequence also appears to have undergone conversion, although in this case, conversion has occurred with a source repeat other than the adjacent 3' *Alu* (Figure 3.6B). The altered sequence of the 5' *Alu* is, however, similar

Figure 3.6: Conversion Events at α-Globin *Alu* Repeats

**A)**




**B)**




A) Xba region *Alu* repeats; B) θ Pst region *Alu* repeats

to other reported *Alu* sequences, including those contained within the α-globin gene cluster [Britten *et al.*, 1988; Sawada *et al.*, 1985]; similarly, a small region of the converted Xba (RH) *Alu* also appears to have undergone exchange with an *Alu* repeat other than those adjacent to it. These gene conversion events alter the recipient sequence to one which is rather different to that present at apparently orthologous positions, thereby explaining the unusually high observed divergences.

---

Table 3.4: Regional Divergence of *Alu* Repeats

| Region | *Alu* Repeat | Divergence (%) + CpG | -CpG |
|---|---|---|---|
| Xba | 5' Alu (LH) | 13.3 | 4.8 |
| | 5' Alu (RH) | 3.5 | 1.4 |
| | 3' Alu | 2.5 | 1.9 |
| θ Pst | 5' Alu (LH) | 5.8 | 5.0 |
| | 5' Alu (RH) | 5.3 | 1.8 |
| | 3' Alu (LH) | 1.4 | 0 |
| | 3' Alu (RH) | 1.9 | 1.3 |

(RH) = right half of *Alu* repeat, (LH) = left half. The 5' and 3' *Alu* repeats are depicted in Figure 3.4 a & e.

---

Gene conversion has been reported at another α-globin *Alu* locus, positioned between the ψα1 and ψα2 genes [Sawada *et al.*, 1985], and has also been observed in a duplicated *Alu* repeat site in the β-globin cluster [Maeda *et al.*, 1983]. In fact, *Alu* repeats have been suggested to be highly recombinogenic in other studies [Rudiger *et al.*, 1995], mainly because they are often found near the breakpoints of deletions and rearrangements, such as many of those that occur in the LDL receptor gene [Hobbs *et al.*, 1990], cytochrome P450 hydroxylase gene [Steen *et al.*, 1995], sex chromosomes [Rouyer *et al.*, 1987] and α-globin gene cluster [Nicholls *et al.*, 1987]. It is not clear, however, whether all *Alu* repeats behave in this way; it is possible that the events that have occurred at the Xba and θ Pst *Alu* repeats may be unusual [Cullen *et al.*, 1995; Elbein, 1992; Sawada and Schmid, 1986; Sawada *et al.*, 1985]. The

two other regions sequenced in this study which contain *Alu* repeats, Acc (containing a partial *Alu* sequence; Figure 3.4 c) and Sac (Figure 3.4 b) do not show evidence of similar large correction events to other *Alu* repeats. Nevertheless, the comparatively low level of polymorphism in the human Xba and θ Pst *Alus* (Table 3.4) may be due to frequent conversion between allelic copies of the repeats; conversely, the Sac *Alu* repeat shows comparatively high levels of human polymorphism and, as will become clear, this too could be the result of conversion.

*Alu* sequences are often polymorphic at the sequence level, as indicated by a survey of a large number of repeats in the human genome [Orita *et al.*, 1990]. The Sac region contains a number of polymorphic sites, although the common sites are contained within the *Alu* sequence itself (RFLP, S: 240 and S: 421; Figure 3.4b). Since these polymorphisms are widespread on haplotype backgrounds (Table 3.1), they are likely to be useful in terms of the ultimate aim of relating haplotypes to one another. Thus, a further 84 chromosomes, of varying haplotypes and geographical origins, were characterised(refer to Chapter 5 for further details of these samples). From a total of human 122 chromosomes (38+84), sequence haplotypes formed by S: 240 and S: 421 were determined (Table 3.5).

Table 3.5: Frequencies of Sequence Haplotypes in Sac Region

| | Sac: 240 | Sac: 421 | Number of Chromosomes |
|---|---|---|---|
| Humans | T | $A_{15}$ | 25 |
| | T | $A_{14}$ | 2 |
| | C | $A_{15}$ | 82 |
| | C | $A_{14}$ | 13 |
| Chimp | T | $A_{14}$ | 2 |
| | | | 124 |

The chimpanzee T-$A_{14}$ sequence haplotype is rare in humans; the three predominant human sequence haplotypes are likely to have been produced by mutation and recombination, perhaps from the chimpanzee sequence, assuming it is ancestral. The presence of all four possible sequence haplotypes suggests that the closely positioned sites, Sac: 240 and Sac: 421 are not tightly linked; further, the RFLP and Sac: 240 sites are only ~50 bp apart (Table 3.1) but the human-associated allele C at Sac: 240 is associated with both (+) and (-) alleles. This indicates the importance of recombination (perhaps localised gene conversion events) within the Sac region in humans.

It is interesting that the 5' end of the haplotype, containing the *Sac I* region, appears to recombine with the 3' end of the haplotype in populations (Chapter 2). It is possible, rather than the 5' end of the haplotype recombining as a unit (i.e. all sites as a linked unit, from *Xba I* to the IZHVR) with the 3' end, that the 5' end sites recombine separately. This possibility would imply that the *Sac I* RFLP and other polymorphic sites could be swapped from one chromosome to another by gene conversion, which is exactly the suggestion from the sequence haplotypes shown in Table 3.5. It is also possible that if the Sac *Alu* is highly recombinogenic, it could have a role to play in the localisation of other recombination events around this region of the α-haplotype, also accounting for the observed recombination between 5' and 3' ends.

A χ sequence, which is recombinogenic in *E.coli* [Smith and Stahl, 1984], is present at the 5' end of the Sac region (Figure 3.4b) and possibly has a role to play in the recombination in this region. Perhaps, the presence of both *Alu* repeats and other recombinogenic signals, and minisatellite sequences in the 5' end of the haplotype induces a localised "hotspot" (although they do not do so in the mammalian HLA complex, [Cullen *et al.*, 1995]). The role for *Alu* sequences requires further

investigation, however, before their involvement in such activity can be confirmed. Although *Alu* sequences do result in deletions in the α-globin gene cluster, with many of the 5' breakpoints positioned at the 5' end of the haplotype (between the ζ genes, [Nicholls *et al.*, 1987]), it is not clear that *Alu*-mediated recombination could occur at sufficient frequency to play a dominant role in production of haplotype diversity (i.e. 5' to 3' end recombination).

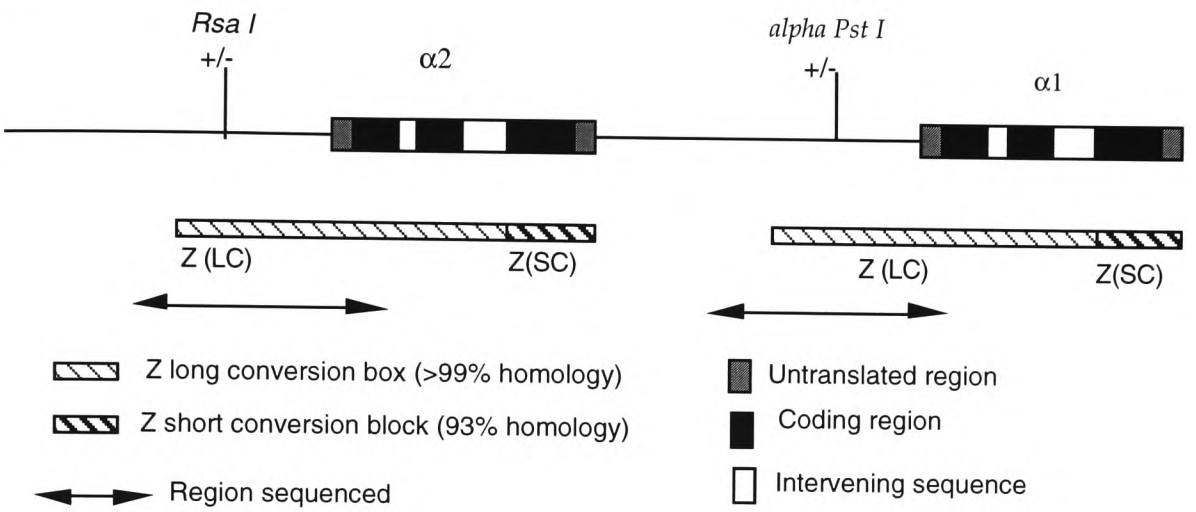
**The α–Gene Z Homology Boxes**

The α-genes are contained within highly homologous duplication blocks: the X, Y and Z boxes [Lauer *et al.*, 1980]; the Rsa and α Pst regions contain the Z box sequences flanking, 5', the α genes (Zα2 and Zα1) (Figures 3.4 d, e & 3.7a). Since these regions are homologous, the sequences and positions of the human polymorphic sites in these regions can be aligned (Figure 3.7b). This reveals the presence of a number of different types of sequence difference between Zα2 and Zα1. Some sites are polymorphic on one Z box, but not on the other; these include both RFLPs, and site -662 on Zα1, and sites -632 and -331 to -330 on Zα2 (Figure 3.7b).

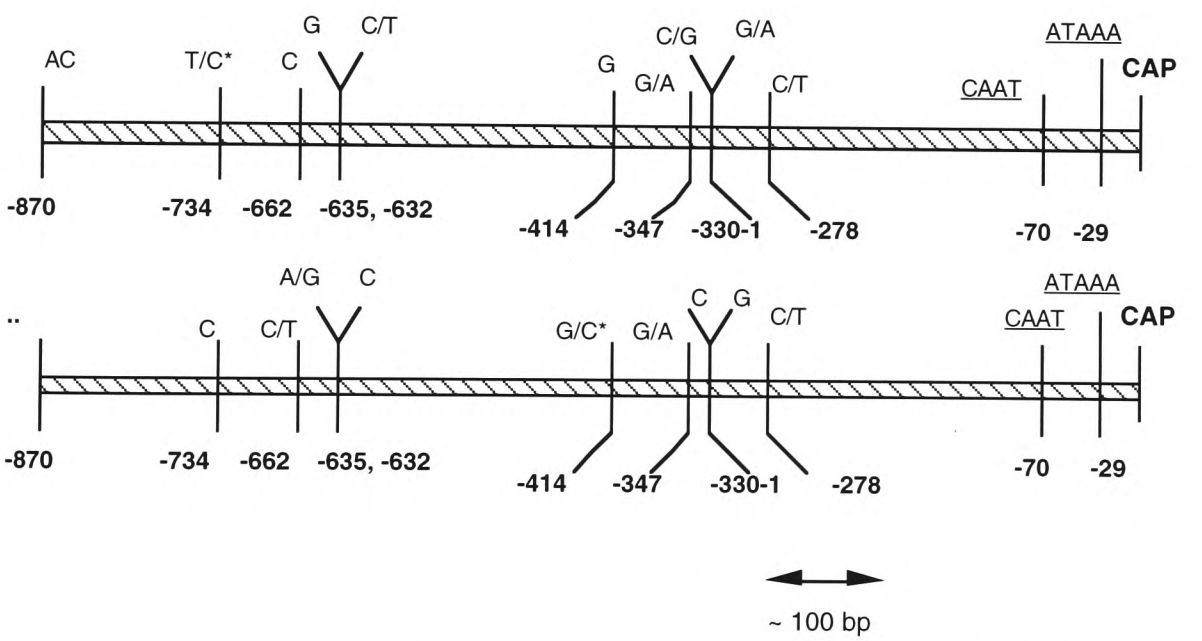
Other sites appear to have been transferred between Zα2 and Zα1 by gene conversion. Site -635 has the sequence A on Zα1 and G on Zα2, and thus was previously reported as a 'fixed' difference between the two Z boxes [Michelson and Orkin, 1983]. However, on one of 38 chromosomes (Bantu Vc/IIIf; Table 3.1), Zα1 has the sequence G, characteristic of Zα2 (Figure 3.7b). Site -278 shows a rare C to T polymorphism, but the rare allele is found within one individual on both Zα2 and Zα1 (British Ia/VIa, Table 3.1) as two combinations (Figure 3.8).

Figure 3.7: Sequenced Regions of the Human α-Gene Z boxes and Positions of Polymorphic Sites

A)

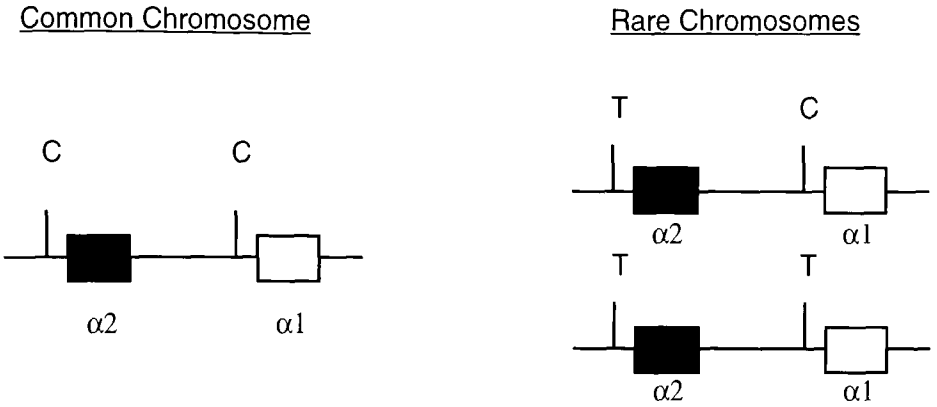



B)




A) The Z box homology block consists of two regions (short and long conversion units) which differ in degree of homology. They are separated by a 7 bp deletion/insertion in IVSII which may affect conversion efficiency [Michelson and Orkin, 1983]. The sequenced regions contain part of Z (LC). B) Sequences of Rsa (Zα2, top) and α Pst (Zα1, bottom) regions aligned, with human polymorphisms named in position relative to the α-gene Cap site. RFLP sites are indicated with an asterisk and dots refer to a deletion. Promoter sequences are underlined.

Figure 3.8: Arrangements of Alleles at Site -278




It is likely that following a mutation on one Z box, the rare allele has been transferred to the other Z box, on the same chromosome, by conversion. This particular example is reminiscent of a previous report of a putative gene conversion event between the α-genes. In an individual expressing an unusually high level of HbI (codon 16, Lys to Glu), the mutation was found *in cis* on both linked α2 and α1 genes [Liebhaber *et al.*, 1984]. Site -347 also appears to have been converted between Zα2 and Zα1, but unlike site -278, it is commonly polymorphic on both Z boxes, with the sequence at this site being G or A (Figure 3.7b; Table 3.1).

Since this polymorphism is common, a further 84 chromosomes were characterised at site -347 (these additional chromosomes are the same as those used to characterise the sites S: 240 and S: 461 in the Sac region above). All four possible sequence haplotypes were found within the total 122 chromosomes sampled (Table 3.6).

Given the presence of these four sequence haplotypes, it seems that both intra- and interchromosomal conversion events are occurring, presumably following misalignment between α2 and α1 during meiosis- such as might occur prior to the formation of an α-gene deletion (Figure 1.8; Chapter 1). For instance, assuming G-G

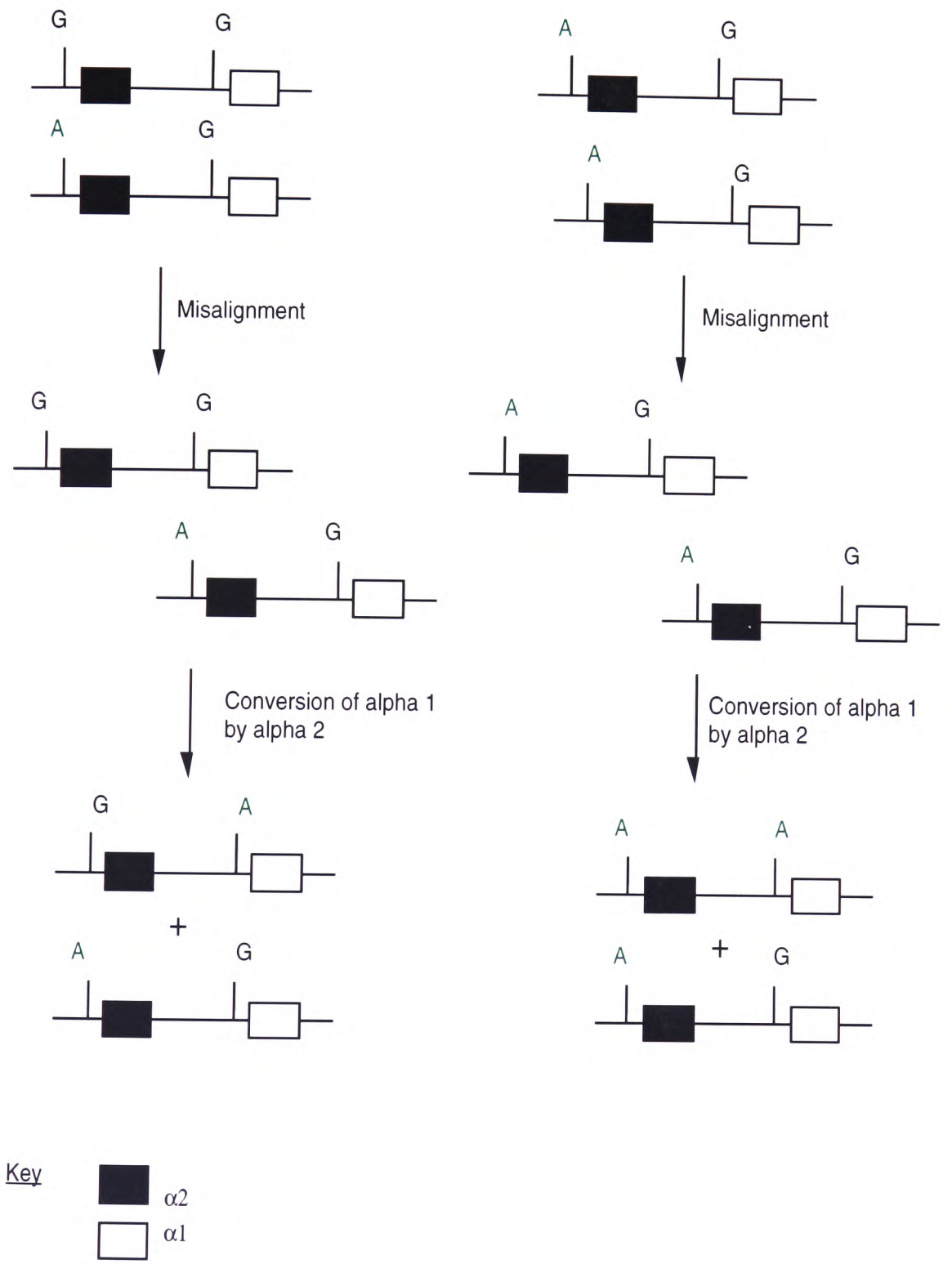
to be the ancestral sequence haplotype - following duplication, the two genes would have the same sequence (and the chimp sequence on Zα1 is G; Table 3.1)- the G to A mutation may have arisen firstly on Zα2 since the A-G sequence haplotype is most common (Table 3.6). The allele A, could then have been transferred by interchromosomal conversion onto Zα1 to form the G-A sequence haplotype, and by intrachromosomal conversion to form the A-A sequence haplotype (Figure 3.9). It is not possible to determine whether inter- or intrachromsomal conversion is more frequent on the basis of the data presented because the relative frequencies of sequence haplotypes from population surveys may reflect demography rather than molecular mechanisms. Nevertheless, the fact that α-gene deletions occur on different haplotype backgrounds in different populations suggests that *de novo* mutation and therefore crossover is relatively frequent [Higgs *et al.*, 1989]. Duplicated genes in yeast demonstrate a more frequent rate of intrachromosomal conversion relative to crossover and interchromosomal conversion [Petes *et al.*, 1991], and if the same is true for the α-globin gene complex, then conversion, and possibly intrachromosomal conversion, may be frequent indeed.

Table 3.6: Zα2 and Zα1 Sequence Haplotypes (Site -347)

| α2 | α1 | Number of chromosomes |
|---|---|---|
| A | A | 4 |
| G | G | 37 |
| A | G | 61 |
| G | A | 20 |
| | | 122 |

Since it is not possible to formally prove gene conversion in mammals, it is possible that the apparently converting sites (-635, -347 and -278) represent hotspots for mutation, such as might feasibly occur at CpG dinucleotides. However, these sites are not contained within CpG dinucleotides, and there is a large body of evidence

Figure 3.9: Schematic Representation of Likely Gene Conversion Events Between α-Genes



suggesting that similar conversion events have occurred at other globin sequences: between duplicated human ζ loci [Hill *et al.*, 1985], horse α-globin genes [Clegg, 1987b] and between the duplicated foetal globin gene loci, $^{G}\gamma$ and $^{A}\gamma$ in the β-globin cluster [Jeffreys, 1979; Powers *et al.*, 1984; Slightom *et al.*, 1980; Smithies and Powers, 1986; Stoeckert *et al.*, 1984].

It is interesting that in hominoid γ genes, directional gene conversion has been observed, with preferential conversion of the downstream $\gamma^2$ by the upstream $\gamma^1$ gene [Fitch *et al.*, 1990]. Likewise, in human α-globin genes, α2 appears to be converting α1 more frequently; Zα2 is more polymorphic than Zα1 (Tables 3.1 & 3.3) and it is the recipient sequence which would have less opportunity to accumulate mutations and for those mutations to become common before being converted again to the donor sequence. Also, many of the observed gene conversion events, such as that at site -278 (Figure 3.8), are explicable in terms of α2 to α1 conversion. This directionality is related to transcriptional differences between the duplicated genes; the preferential donors, $\gamma^1$ and α2 are both dominantly expressed [Fitch *et al.*, 1990; [Higgs *et al.*, 1989; Orkin and Goff, 1981], with α2 mRNA predominating, throughout development, over α1 by a ratio of 3:1 despite identical translational profiles; this predominance is reflected in the fact that most non-deletional α-thalassaemias occur in the α2 gene [Higgs *et al.*, 1989]. The transcriptional gradient present at duplicated α-gene loci has been recently demonstrated in sheep, where multiple copies of the α genes are down-regulated in a gradient from 5' to 3' [Ristaldi *et al.*, 1995]. The cause of the gradient is proximity of promoters to the α-gene controlling element, HS-40, which lies upstream of the α-cluster [Vyas *et al.*, 1992], and acts as an enhancer [Sharpe *et al.*, 1992].

Transcription and recombination are linked processes due to the properties of transcriptionally active chromatin: free from nucleosomes, and sensitive to DNase

and dsbs [Liu *et al.*, 1995; Nicholas *et al.*, 1989; Ponticelli and Smith, 1992; Thomas and Rothstein, 1989]. DNase-hypersensitive sites are found 5' to α2 and α1 genes, associated with promoter sequences, which are present within 100 bp of the Cap site [Nicholls *et al.*, 1987](Figure 3.7b). If these promoter sequences behave as initiation sites for recombination, facilitated by α-gene sequence homology, the transcriptionally more active α2 gene may be less susceptible to dsbs through being more frequently associated with the protein complexes of transcription. Conversely, the α1 gene promoter sequences may be more accessible to the recombination proteins and suffer dsbs, and consequently behave as the preferred recipient of information [Nicolas *et al.*, 1989]. It is of course possible that conversion does occur from α1 to α2 but that selection on the dominant α2 gene ensures that any mutations introduced into α2 as a result of being converted, do not persist. Again, it is difficult to distinguish between these different mechanism on the basis of population data.

Assuming a role for the α-gene promoter sequences in the initiation of recombination, comparison with yeast would suggest the possibility of gene conversion tracts extending from the promoter sequences (<100 bp from the Cap site; Figure 3.7b). Such a tract could explain why a number of site differences between Zα2 and Zα1, situated some distance from the promoter, have not undergone conversion. Inefficient conversion is unlikely to be the explanation, however, for all sites and particularly for those situated between other sites which are known to convert; for example, the dinucleotide polymorphism, sites 331-330, has not undergone conversion between Zα2 and Zα1 but is positioned between converting -347 and -278. It is possible that conversion tracts do not extend continuously along the Z box or even that initiation occurs at more than one site, with only a few bases or small patches converted at any one time (see Michelson and Orkin, [1983] and Zangenberg *et al.*, [1995] for evidence of this type of conversion pattern in humans). It is likely that a number of different types of event occur,

including more extensive conversion responsible for the sequence homogenisation of the α-genes, and smaller events which swap individual polymorphisms. The probability of a particular polymorphism being converted between Z boxes is also likely to be determined not only by its position relative to an initiation site, but by its age. Polymorphic sites -331 to -330 may have arisen recently (despite being relatively common; refer to Chapter 5) so that the opportunity has not yet arisen for conversion. Similarly, the *Rsa I* and *α Pst I* RFLP (+) alleles have not undergone conversion, and are likely to be young mutations, a possibility supported by the fact that these alleles are found on only a limited number of haplotype backgrounds [Higgs *et al.*, 1986]. The lack of involvement of these RFLPs is probably the reason why the frequent conversions that have occurred between α2 and α1 are not detectable at the RFLP haplotype level.

The fact that the RFLP sites are not apparently involved in the conversion events, and that sequencing has demonstrated that (-) alleles all have the same sequence tends to confirm the RFLP haplotype as a valid structure with which to summarise sequence variation in the α-globin gene cluster. However, it is clear that the definition of a recombination hotspot depends on the level of resolution used; the α-genes, for instance, may be considered hotspots for conversion, facilitated by the presence of extensive homology around and within the duplicated genes, although overall, the α-globin gene cluster has a "normal" level of recombination. This highlights the importance of investigating recombination in small regions of the genome.

Localised recombination has important consequences for the observed sequence variation within humans and for patterns of divergence between human and chimpanzee. Firstly, it can homogenise sequences, as at the anciently duplicated α-genes [Liebhaber and Begley, 1983; Zimmer *et al.*, 1980] which produce identical

polypeptides in humans, and at the human Xba and θ Pst *Alus*. The reasons why some genes, such as the α-genes, remain identical after duplication, while others diverge (the θ and ψα genes) are not clear [Hughes, 1994], although the rate of gene conversion is likely to be an important factor. Secondly, gene conversion can increase the number of polymorphic sites, such as within the Z box sequences and Sac *Alu* repeat. The increase effected by conversion is indirect; recombination can uncouple neutral and selected sites (within the α-gene coding sequence, for example) thereby overcoming the influences of genetic hitch-hiking (see Chapters 1& 6).

Before the consequences of recombination for the determination of α-haplotype relationships can be considered, a question is raised about the regions of the α-haplotype which contain repetitive sequences, since such sequences are known to be recombinogenic. It seems entirely possible that these regions might also exhibit similarly localised "hotspot" activity, and if so have further implications for haplotype structure and evolution.

# Chapter 4

## Variation at the Hypervariable Regions of the α-Haplotype

There are seven hypervariable regions (HVRs) in and flanking the α-globin gene cluster: the 5' HVR [Jarman and Higgs, 1988], 3' HVR [Higgs *et al.*, 1981; Jarman *et al.*, 1986], IZHVR [Goodbourn *et al.*, 1983; Goodbourn *et al.*, 1984; Higgs *et al.*, 1981] and the HVRs contained in the first and second introns of the zeta (ζ2/ζ1) and pseudozeta (ψζ1) genes (IVS1.HVR and IVS2.HVR) [Proudfoot *et al.*, 1982].The α-globin HVRs are related in sequence to each other, and to an HVR 5' to the insulin gene [Bell *et al.*, 1982](Figure 4.1).

---

Figure 4.1: Sequence Relationships between the Repeat Units of the α-HVRs and Insulin HVR

```
                                                    GNGGGNACAG        Core
                                                      |||   |||
GGGGAGCATTCAGGAGGCCTTCCCGGAGGTAGGGTGGTGGGAAGAAGGGG-TCAGCGT  5' HVR
                                                      |||   |||
                                                  RGGAGGG-ACAGTG   ζ HVR
                                                      |||   |||
                                                  TGTGGGG-ACAGGGG  Insulin
                                                      |||   |||
                                        TGTGAGGGTGCCCGGGACGGCTTGTGGGGCACAG-GT  IZHVR
                                                      |||   |||
                                                  ACACGGGGGGAACAGCG  3' HVR
```

A consensus or core sequence is common to all these HVRs. R= G or A.

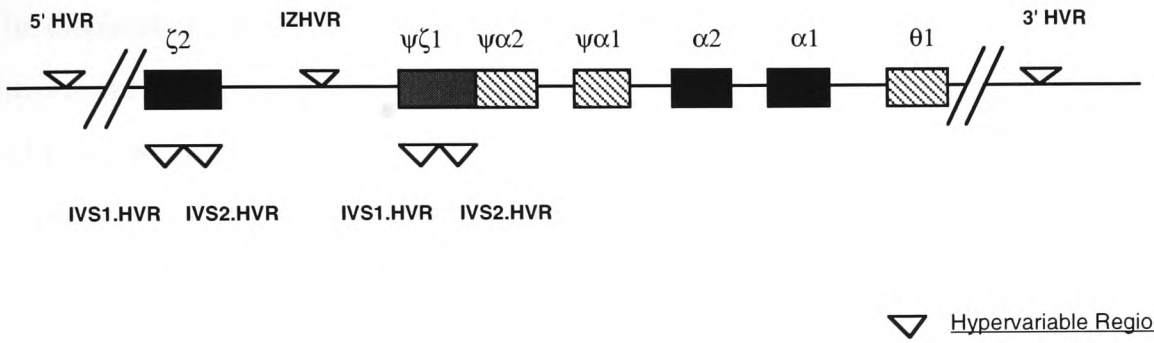
---

The highly polymorphic 5' and 3' HVRs (Figure 4.2) are not included in the α-haplotype [Higgs *et al.*, 1986], although they do exhibit allelic relationships with particular haplotypes [Martinson *et al.*, 1994]. The 3' HVR consists of a tandem array of a 17-21 bp repeat sequence; in family studies, a *de novo* mutation was detected in 1 in 250 meioses [Jarman *et al.*, 1986], indicating a mutation rate comparable to non-α globin HVR loci [Jeffreys *et al.*, 1988]. The 5' HVR consists of a 57 bp repeat unit, and has a lower mutation rate than the 3' HVR; no *de novo* mutations have been found and the HVR is stable in bacteria [Jarman and Higgs, 1988]. There is evidence that

both minisatellite and microsatellite loci with the lowest levels of inter-repeat sequence variability have greater allelic variability or instability [Jeffreys *et al.*, 1985a; Eichler *et al.*, 1994]. This is consistent with the observed differences between the 3' and 5' HVR, since the more polymorphic 3' HVR has comparatively little sequence variation between repeat units within an array [Jarman and Higgs, 1988].

---

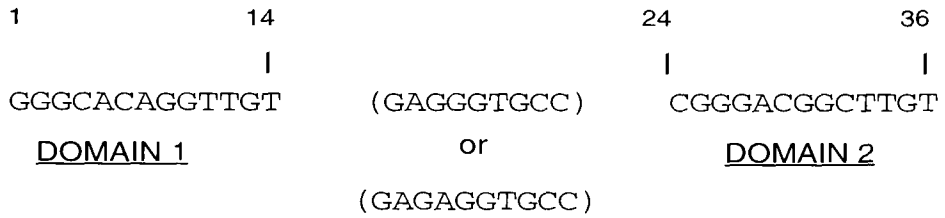
Figure 4.2: Hypervariable Sequences within the α-Globin Gene Cluster




---

The IZHVR (Figure 4.2), which is less polymorphic than the 5' and 3' HVRs, is included in the α-haplotype. It was originally defined by restriction enzyme analysis to have just three alleles, termed small, medium and large (S, M, L) [Higgs *et al.*, 1981]. The tandem array is formed from a 36 bp (or 37 bp) repeated sequence which consists of two 14 bp domains, related in sequence to each other and to the 14 bp repeat unit of the IVS1.HVR in the ζ genes [Goodbourn *et al.*, 1983], separated by an unrelated 9-10 bp segment (Figure 4.3). The number of repeats in the IZHVR array ranges from about 12-58 [Goodbourn *et al.*, 1983; Goodbourn *et al.*, 1984], but, as indicated by the S, M, L classification, the distribution of these alleles is trimodal (unlike the more continuous distributions of 5' or 3' HVR alleles). The tri-allelic classification of the IZHVR was determined at a low level of resolution. Better resolution, using different enzymes and Southern blotting [Goodbourn *et al.*, 1984], reveals extra size variation within the M and L size classes. This additional variation has not, however, been systematically characterised.

Figure 4.3: Structure of the IZHVR Repeat Sequence



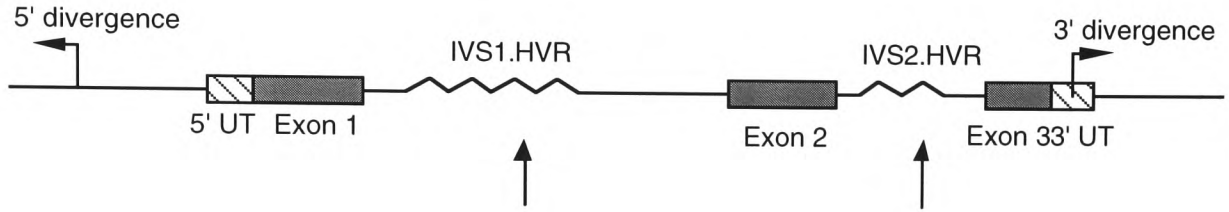

The duplicated ζ genes have a similar overall structure to the α-genes, but their two introns are much larger due to the presence of repeat sequences. In intron 1 there is a 14 bp repeat array (IVS1.HVR), and within intron 2 there is a pentanucleotide repeat or microsatellite (IVS2.HVR) (Figures 4.2 & 4.4). The downstream ζ gene is polymorphic in human populations, being present either as a pseudogene (ψζ1) [Proudfoot *et al.*, 1982] or as a gene with an apparently functional sequence (ζ1) [Hill *et al.*, 1985]. The intron repeats of the ψζ1 gene have expanded in size: the IVS1.HVR consist of 39 repeats in ψζ1, compared to either 16 or 12 repeats within the ζ2 gene [Hill *et al.*, 1985; Proudfoot *et al.*, 1982] (Figure 4.4). The IVSI.HVR of the converted ζ1 gene consists of 16 repeats [Hill *et al.*, 1985], and it is the consequent size difference between ψζ1 and ζ1 that forms the basis of the PCR haplotyping strategy at the PZ/Z site (Chapter 2; and also the original haplotyping protocol, Appendix 1). As at the IZHVR, additional size variation attributable to differences in numbers of tandem repeats, has been noted at the ψζ1 IVS1.HVR in Southern blotting studies [Chapman *et al.*, 1986; Goodbourn *et al.*, 1983].

## Characterisation of Variation at the IZHVR Locus

### Size Variation

The full α-haplotype, determined by Southern blotting, defines the IZHVR according to the simplified S, M, L classification [Higgs *et al.*, 1986](Chapter 1). PCR

Figure 4.4: Comparison of the Homologous Regions of the ψζ1 and ζ1/ζ2 Genes



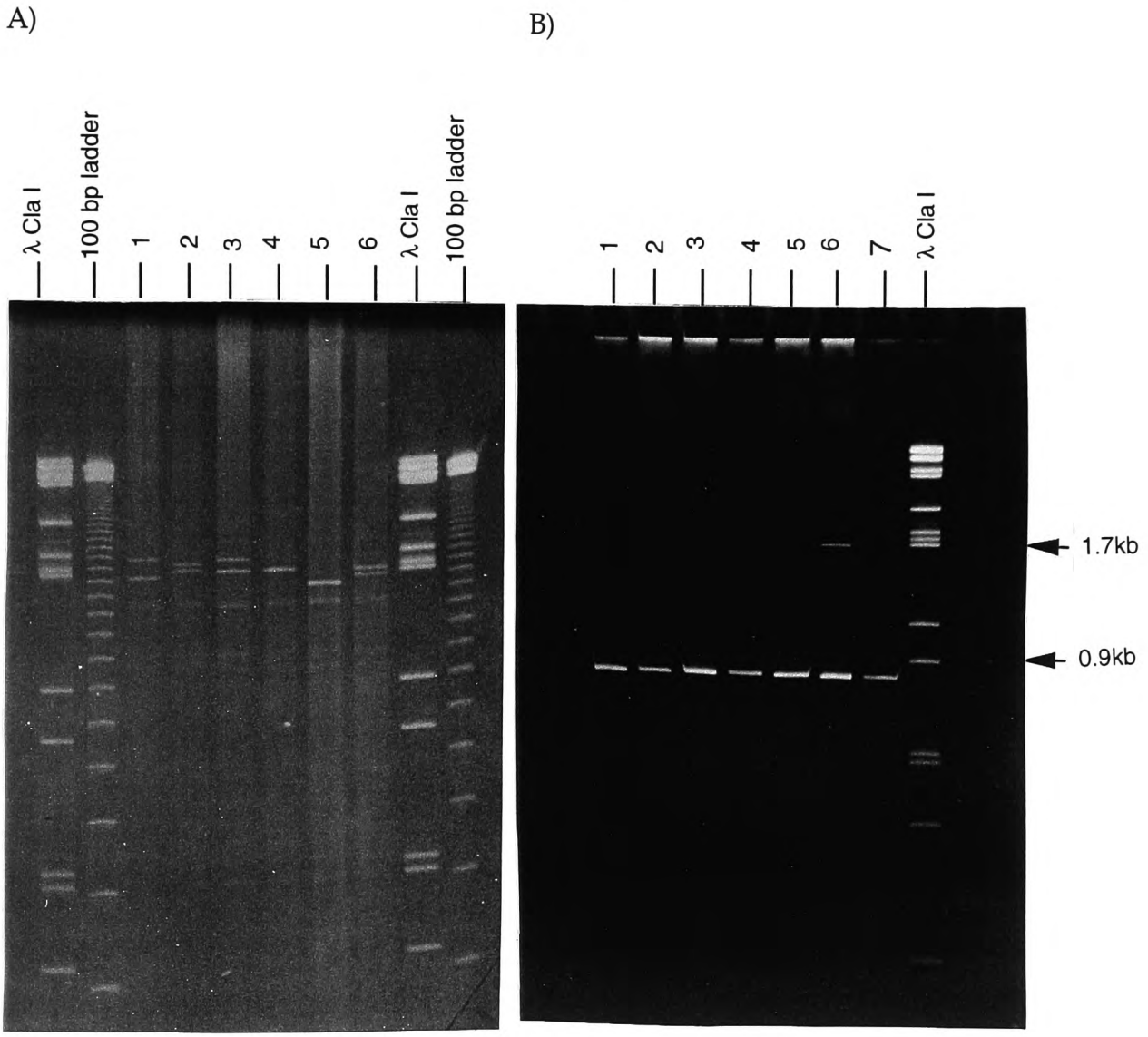

| | IVS1.HVR | IVS2.HVR |
|---|---|---|
| ζ2 | 12/16 | S |
| ζ1 | 16 | L |
| ψζ1 | ~ 39 | LL |

The ζ2/ζ1 and ψζ1 genes are homologous over a region that extends from the 5' flanking region through to the 3' untranslated region; beyond these regions, the sequences diverge [Proudfoot *et al.*, 1982]. Within the first and second introns of the genes are two repetitive sequences: IVS1.HVR (a VNTR) and IVS2.HVR (a microsatellite). The sizes of these repeat arrays differ between the different genes, as shown (S = small, L = large and LL indicates an allele larger in size than an L).

haplotyping provides the opportunity to classify IZHVR alleles according to the S, M, L classification (Chapter 2), or to characterise the reported extra allelic variation [Goodbourn *et al.*, 1984] in greater detail. The ability to characterise variation at this locus fully is important because it is not clear whether two alleles of apparently identical size (according to blotting) are equivalent, and this a question that needs to be addressed when comparing haplotypes in order to determine their mutational relationships.

The size of the amplified product is in a range comparable to the fragments generated in high resolution Southern blotting studies (Chapter 2) [Goodbourn *et al.*, 1984]; in the latter, alleles differing by two or three repeats of 36 bp were resolved. Southern blotting, however, inevitably results in bands which are more diffuse than PCR bands, and there are associated problems of inconsistent hybridisation with alleles of different sizes, as indicated in Chapter 2. Resolution of amplified IZHVR alleles is better, being considerably enhanced by the possibility of analysing alleles on a variety of gels. The IZHVR was amplified using primers described in Chapter 2, and differences between alleles corresponding to one repeat (36 bp) were identified for the S and M size class alleles, but not L alleles (which are prohibitively large in size). Initial estimation of allelic variation was made on agarose gels (Appendix 1: A.4.1) which included reference alleles (i.e. sized on a previous gel). Finer resolution of S and M alleles was then possible by the use of hybrid agarose/polyacrylamide gels (Figure 4.5) (Appendix 1: A1.4.2); these gels can be made with low concentrations of polyacrylamide since the presence of agarose strengthens the gel matrix (without affecting the resolving power of the gel) [Arteiler and Porteous, 1992].

Figure 4.5: IZHVR Allelic Size Variation




A) M alleles on hybrid 3% acrylamide/ 1% agarose gel; B) S alleles (sample 6 is an M/S heterozygote) on 4% acrylamide/ 1% agarose gel.

A survey of IZHVR variation was conducted in a total sample of 726 unrelated chromosomes: 198 from a number of world populations, 230 Senegalese Mandenka, 158 Kenyan Luo and 142 Icelandic chromosomes. This survey confirmed variation within the allelic size classes noted in previous studies [Goodbourn *et al.*, 1984]. In total, ten types of M allele and three types of S allele were found (although one allele predominates in frequency), and although L alleles are too large to be subtyped accurately, two variant sized alleles were found which differed enough in size from the "common" L allele (estimated to be ~59-60 repeats) to be distinguishable (Table 4.1). Seven of the M suballeles found here, and variation in size of L alleles, have been observed previously [Goodbourn *et al.*, 1984](Table 4.1); subtypes of the S allele have not been previously reported.

Table 4.1: IZHVR Repeat Number and Estimated Size of PCR Sub-Alleles

| No. Repeats (N) | Approximate amplified size (36 bp. x N)+ 443 bp. |
|---|---|
| 9 | 767 |
| 11 | 839 |
| 12 | 875 |
| 38 | 1,811 |
| 39* | 1,847 |
| 40* | 1,883 |
| 41* | 1,919 |
| 42* | 1,955 |
| 43* | 1,991 |
| 44* | 2,027 |
| 46* | 2,099 |
| 47 | 2,135 |
| ~55 | 2423 |
| ~59 | 2567 |
| ~64 | 2747 |

* Size range of [Goodbourn *et al.*, 1984]. Amplified flanking sequence, 5' and 3', totals 443 bp. Arrays of 9-12 repeats are small alleles, 38-46 are medium alleles and 55-64 are large alleles. Alleles are subsequently referred to as size class and number of repeats; e.g. an allele of a size consistent with 44 repeats would be termed an $M_{44}$ allele.

Despite relatively small sample sizes, there is evident world population differentiation in IZHVR allele frequency distributions (Table 4.2), with similarities in allele frequency between related populations. For example, from α-haplotype data, Micronesian populations are believed to have recent Southeast Asian affinities [O'Shaughnessy *et al.*, 1990], and possess $M_{41}$, $M_{40}$ and $S_{12}$ sub-class alleles in common with Taiwanese, Philippine and Brunei populations. From α-haplotype data, Polynesians show evidence of genetic input from both Southeast Asia and Melanesia [O'Shaughnessy *et al.*, 1990]; in support of this, Polynesians possess $M_{40}$ alleles (a Southeast Asian, but not Melanesian, allele), and $S_{11}$ alleles (which appear to be Melanesian-specific in this survey).

The prevalence of the two most common M alleles, $M_{40}$ and $M_{41}$ (Table 4.2), can be explained by the haplotype composition of the surveyed populations: the predominant haplotypes from many of these populations are the Ia and IIIa. To illustrate this haplotypic component to population differentiation, further samples of Ia and IIIa haplotypes were subtyped (Table 4.3). This demonstrates the association between Ia and IIIa haplotypes with $M_{40}$ and $M_{41}$ alleles, respectively. The other predominant M allele, $M_{42}$, was found in this survey in Sri Lankans on IIIa haplotypes (Table 4.3), New Guinea highland and African populations; if this allele is related in these populations, it may represent the remnant of ancient relationships (an ancient eastern migration from Africa may have populated New Guinea and left traces in parts of Asia [Lahr and Foley, 1994]).

Since the sample of world populations included only chromosomes of known haplotype, the range of IZHVR alleles observed may not be representative of the full range of allelic sizes that exists within a random sample (especially given associations between some haplotypes and allele size). For this reason, the

Table 4.2: IZHVR Allelic Variation in World Populations

| Allele | Number of Chromosomes | | | | | | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Senegal | Zambia | Jamaica | Amaz. Indians | Taiwan | Philippines | Brunei | Micronesia | Melanesia | PNG (H) | PNG (C) | Polynesia | Hong Kong | Sri Lanka | Iceland | Britain | |
| L59 | - | - | - | - | 3 | - | - | - | - | - | - | - | - | - | - | - | 3 |
| M43 | - | - | - | - | - | - | - | - | - | 1 | - | - | - | - | - | - | 1 |
| M42 | 11 | - | 2 | - | - | - | - | - | - | 2 | - | - | - | 4 | - | - | 19 |
| M41 | - | 2 | - | 6 | 2 | 1 | 2 | 8 | 6 | 9 | 6 | - | - | - | 10 | 1 | 47 |
| M40 | 3 | - | 4 | 4 | 17 | 3 | 4 | 14 | - | - | - | 6 | 2 | 2 | 6 | 3 | 64 |
| S12 | - | - | - | - | 22 | 2 | 2 | 8 | 5 | - | - | - | 2 | 2 | - | - | 43 |
| S11 | - | - | - | - | - | - | - | - | 1 | 2 | 4 | 4 | - | - | - | - | 11 |
| Total | 14 | 2 | 6 | 10 | 44 | 6 | 8 | 30 | 12 | 14 | 10 | 10 | 4 | 8 | 16 | 4 | 198 |

Note: samples include Senegalese Mandenka, Amazonian Indians as detailed in Zago *et al.*, [1995], New Guinea Coastal (C) peoples and Highlanders (H), Melanesians from Vanuatu, Aboriginal Australians from Arnhem Land, French Polynesian, Western Samoan, Mauke and Mitiaro islanders, aboriginal Taiwanese, Micronesians from the Majuro, Ponape, Kiribati and Guam Islands, and Hong Kong Chinese.

Table 4.3: Association of Ia and IIIa Haplotypes with IZHVR Alleles in World Populations

| Allele | Number of Chromosomes (Ia) (+ + - M PZ + + - -) | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| | Melanesia | Australia | Micronesia | SE Asia | Sri Lanka | Europe | PNG (H + C) | |
| M42 | | | | | 4 | | | 4 |
| M41 | 6 | 8 | 4 | | | 3 | 6 | 27 |
| M40 | | | | 2 | | | | 2 |
| | | | | | | | | 33 |

| Allele | Number of chromosomes (IIIa) (- - + M Z - - - -) | | | | | Total |
|---|---|---|---|---|---|---|
| | Africa | Polynesia | Micronesia | SE Asia | Europe | |
| M42 | 2 | - | - | - | - | 2 |
| M41 | - | - | - | 2 | 1 | 3 |
| M40 | 8 | 6 | 16 | 3 | 3 | 36 |
| | | | | | | 41 |

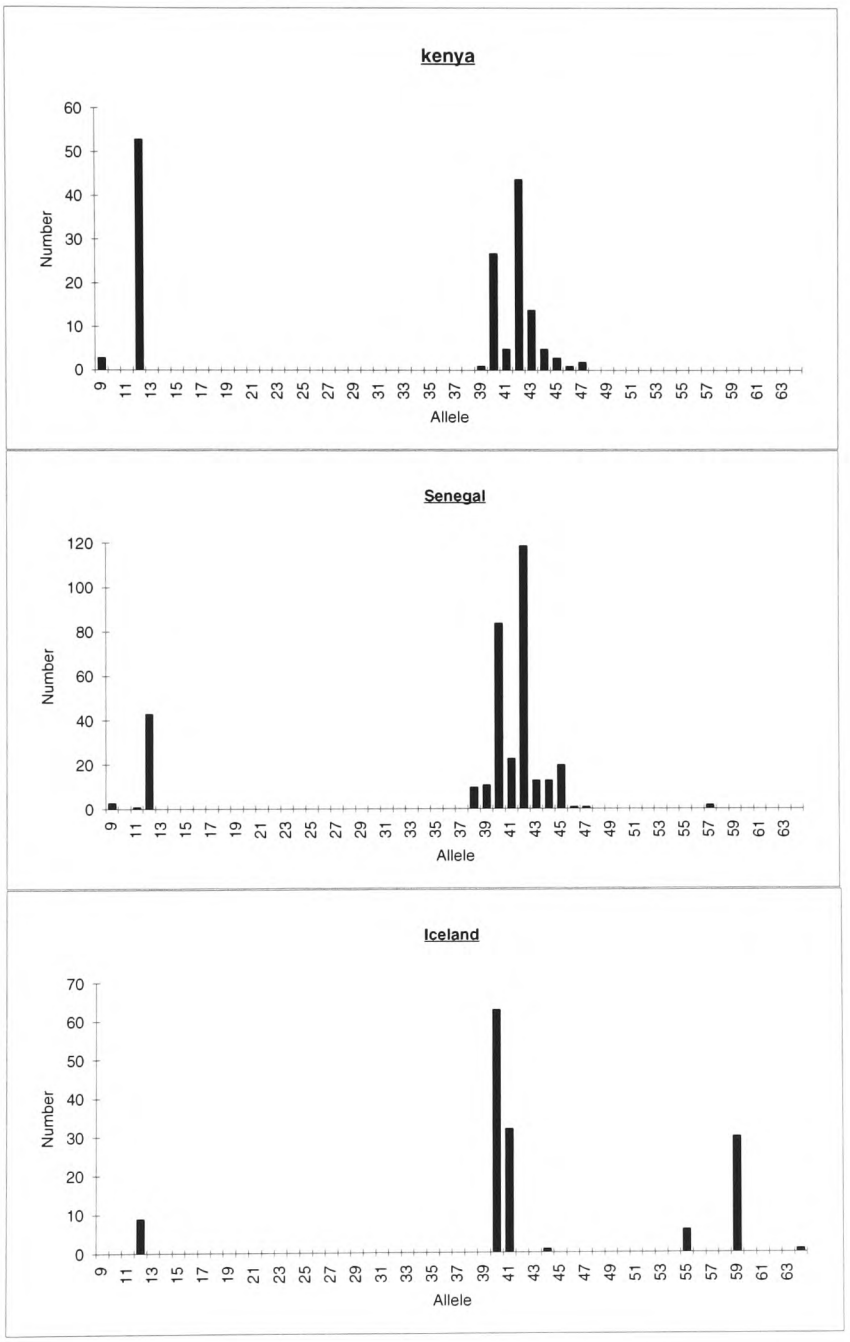
Mandenka, Luo and Icelandic populations were studied in detail (Table 4.4). This revealed a greater number of alleles and further population differentiation.

The African allele frequency distributions are similar in showing quite extensive, quasi-continuous variation within the M class (Figure 4.6). An L allele is found in the Mandenka; this size class is very rare in Africa [Higgs *et al.*, 1986], although it is relatively common in non-African populations (in Iceland, the L allele class has an overall frequency of 26%; Table 4.4). The L-class alleles may therefore represent a mutation that occurred outside of Africa; the Senegalese L allele, in being a different size to the Icelandic L alleles, may represent either a *de novo* mutation or gene flow from a non-African population carrying the allele. This highlights the need to ultimately look at HVRs at the sequence level, as well as at size variation.

Table 4.4: Allelic Variation in Three Populations

| | | Frequency (%) | |
|---|---|---|---|
| Allele | Kenya | Senegal | Iceland |
| L~64 | - | - | 0.7 |
| L~59 | - | - | 21.1 |
| L~57 | - | 0.6 | - |
| L~55 | - | - | 4.2 |
| M47 | 1.3 | 0.3 | - |
| M46 | 0.6 | 0.3 | - |
| M45 | 1.9 | 5.8 | - |
| M44 | 3.2 | 3.8 | 0.7 |
| M43 | 8.9 | 3.8 | - |
| M42 | 27.8 | 34.6 | - |
| M41 | 3.2 | 6.7 | 22.5 |
| M40 | 17.1 | 24.4 | 44.4 |
| M39 | 0.6 | 3.2 | - |
| M38 | - | 2.9 | - |
| S12 | 33.5 | 12.5 | 6.3 |
| S11 | - | 0.3 | - |
| S9 | 1.9 | 0.9 | - |
| No. Chromosomes | 158 | 344 | 142 |

Figure 4.6: IZHVR Allele Distributions in Three Populations




Note: alleles in number of repeats.

The relative diversities of these three populations were compared; diversity, or heterozygosity (H), corrected for sample size, was calculated by Equation 4.1:

$$H = 2n\,(1-\sum_{i=1}^{m} x_i^2)/(2n-1) \qquad \textbf{(4.1)}$$

where n= sample size, i= allele and m= number of alleles [Nei, 1987]. The variance of H is calculated by Equation 4.2 [Nei, 1987]:

$$V(H) = 2\,/2n\,(2n-1)\,\{2\,(2n-2)\,[\Sigma x_i^3-(\Sigma x_i^2)^2] + \Sigma x_i^2 - (\Sigma x_i^2)^2\} \qquad \textbf{(4.2)}$$

Heterozygosity ranges from 0.70 to 0.79 in these populations (Table 4.5); if only size class had been recorded (i.e. the locus studied at the level of resolution used in α-haplotyping by Southern blotting), H would be 0.441 for the Luo. This is close to the maximum possible value of H of 0.5, since only two size classes (S and M) were found in this population, and indicates the greater information that the locus can provide when subtyped. H is not significantly different between the two African populations, but the Icelandic population shows significantly lower diversity compared to the African populations (Table 4.5). The lack of significant difference in diversity between the Mandenka and Luo contrasts with the comparison made from their α-haplotype repertoires, which indicated that the Luo were less diverse (Chapter 2). Presumably the mutation rate of the HVR locus is higher compared to the full haplotype, making its diversity less subject to the demographic effects which could lead to lower haplotype diversity (see Chapter 2).

The lower Icelandic diversity probably does reflect demographic events, however, since the alternative explanation, a lower mutation rate, seems unlikely. That HVR heterozygosity can be sensitive to demographic events is demonstrated in Polynesia, where a sudden reduction in population size, or bottleneck event, is believed to have

occurred and this has reduced heterozygosity of a number of HVR loci [Flint *et al.*, 1989]. Iceland, compared to African populations, may have had a smaller effective population size and over a shorter time (if we consider African populations to be oldest; refer to Chapter 6). However, since L alleles cannot be subtyped in the same way as M and S alleles, the possibility remains that diversity has been underestimated in the Icelandic population.

---

Table 4.5: Heterozygosity at the IZHVR within Three Populations

| Population | No. Chrms | No. Alleles | H | t | |
|---|---|---|---|---|---|
| | | | | Senegal | Kenya |
| Senegal | 230 | 14 | 0.785(0.0089) | - | 0.7792 |
| Kenya | 158 | 11 | 0.773(0.0126) | 0.7792 | - |
| Iceland | 142 | 7 | 0.704(0.017) | 4.22* | 3.621* |

Significant values of the normal deviate, t are indicated with an asterisk. * Variance of heterozygosity (H) is shown in parentheses.

---

**Sequence Variation at the IZHVR**

Internal Restriction Enzyme Mapping

The principle of this indirect method to characterise sequence variation is to label DNA at one end and partially digest with a restriction enzyme to produce an overlapping series of molecules with a common labelled terminus [Smith and Birnstiel, 1976]. Applied to amplified HVRs, information about the internal structure of the array can be obtained [Jeffreys *et al.*, 1990; Martinson *et al.*, 1994]. The method was used to analyse S alleles which were not of a prohibitively large size for examining the restriction enzyme products on acrylamide gels, and for which it was easy to obtain plenty of template for digestion. PCR primers were $^{32}$P-labelled with polynucleotide kinase (Appendix 1; A1.3.1) and PCRs performed with one primer labelled.

From the published sequence of the M allele [Goodbourn *et al.*, 1983], it is evident that most of the sequence polymorphisms present between repeats within an array cannot be typed by restriction enzymes. One polymorphic dinucleotide region of the repeat unit, however, has *BsaJ1* and *BstN1* sites (*Sma I* could be used as an alternative enzyme to *BstN1*, since it produces a complementary digestion pattern [Goodbourn *et al.*, 1983]) (Figure 4.7). Conditions for partial digestion, so that each partial digestion band is represented, were determined empirically (Appendix 1; A1.6.1). Four individuals, homozygous for one of three differently sized S alleles, were amplified using a labelled IZ.R5 primer and partially digested with *Bsa J1* and *BstN1* to obtain information from the 3' end of the HVR array (Figure 4.8; Table 4.6). The $S_{12}$ alleles were found to be similar in map structure, with only one difference in digestion pattern between the alleles in the *BstN1* digest (Table 4.6). The $S_{11}$ alleles from the New Guinea highlands, although one repeat smaller than the $S_{12}$, have a similar map structure to the Melanesian $S_{12}$, suggesting that these alleles are probably related (Table 4.6).

---

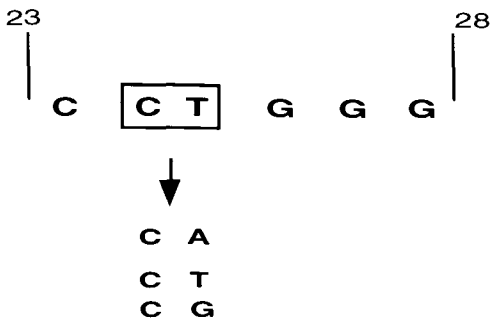
Table 4.6: Internal Mapping Codes for Eight S-Class IZHVR Alleles (First Six Repeats)

| Allele | Population | Code |
|---|---|---|
| $S_{12}/S_{12}$ | Micronesia | abbbba |
| $S_{12}/S_{12}$ | Melanesia | abbaba |
| $S_{11}/S_{11}$ | New Guinea H | abbaba |
| $S_9*/S_9*$ | Senegal | abcbbb |

Note: * 37 bp, not 36 bp, repeats; a= both enzyme sites are +, b= *Bsa J1* is (+) and *Bst N1* is (-), and c = both enzymes sites are (-). Note that *Bsa J1* cannot be (-) if *Bst N1* is (+) (see Figure 4.7). H= highlands.

---

Figure 4.7: Polymorphic Position that can be Detected by Restriction Enzyme Analysis

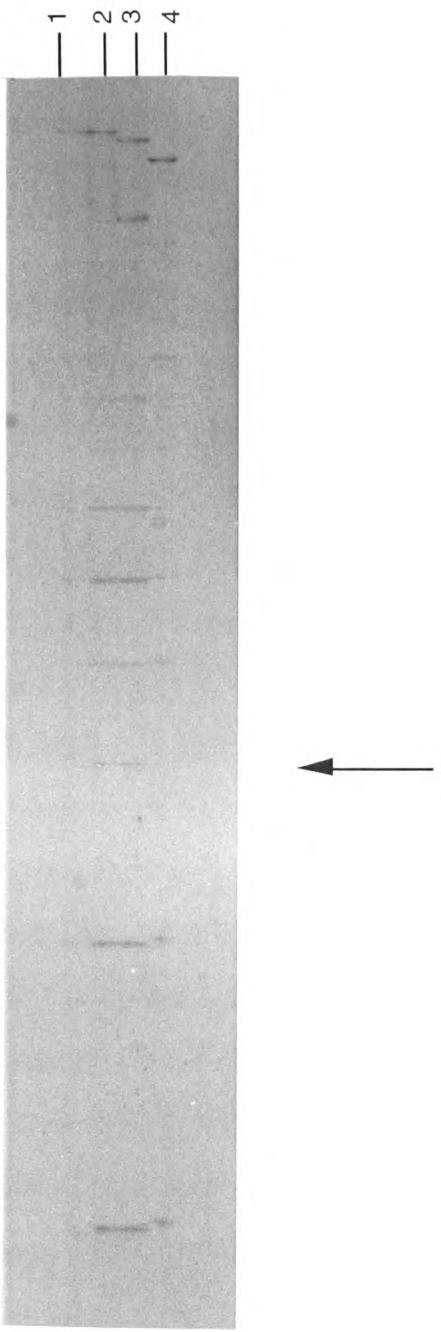



Structure Of Published M sequence

| Repeat number | *Sma I* | *BsaJ1* | *BstN1* |
|---|---|---|---|
| 1 | - | + | + |
| 2 | - | + | + |
| 3 | - | + | + |
| 4 | - | + | + |
| 5 | - | + | + |
| 6 | - | + | + |
| 7 | + | + | - |
| 8 | + | + | - |
| 9 | + | + | - |
| 10 | + | + | - |
| 11 | - | + | + |
| 12 | - | + | + |
| 13 | + | + | - |
| 14 | + | + | - |
| 15 | + | + | - |
| 16 | + | + | - |
| 17 | + | + | - |
| 18 | + | + | - |
| 19 | + | + | - |
| 20 | + | + | - |
| 21 | - | + | + |
| 22 | + | + | - |
| 23 | + | + | - |
| 24 | - | + | + |
| 25 | + | + | - |
| 26 | - | + | + |
| 27 | + | + | - |
| 28 | + | + | - |
| 29 | - | + | + |
| 30 | - | + | + |
| 31 | + | + | - |
| 32 | + | + | - |

In the sequence of the M allelic array [Goodbourn *et al.*, 1983], positions 24 and 25 in a 36 bp repeat unit, are polymorphic. Within the array, there are three possible sequences at these positions: C T, C A, C G. The sequences can be typed by *Sma I*, *Bst N1* and *Bsa J1*.

Figure 4.8: Restriction Enzyme Mapping of Small Size Class IZHVR Alleles




Analysis of repeats at the 3' end of four *Bsa J1* partially digested alleles, on a 3% denaturing polyacrylamide gel. Allele 4 consists of 37 rather than 36 bp repeats. An example of allelic sequence differences is indicated by arrow: samples 1 and 4 have no *BsaJ1* site at repeat number 3.

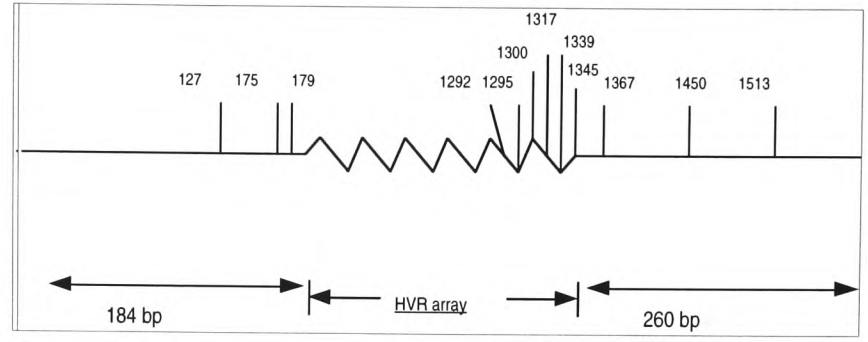
The African $S_9$ allele is quite different in size and in map structure; from the 3' end of the array, it has 37 bp, not 36 bp, repeats and the map of this sample is therefore out-of-register with the other alleles (Figure 4.8), and sequence differences are evident on the basis of both enzyme digests (Tables 4.6; Figure 4.8).

Direct Sequencing Analysis of IZHVR Alleles

Single-stranded templating and direct sequencing of the amplified IZHVR alleles were performed as described in Chapter 3, with sequencing primers described in Appendix 3. Note that repetitive regions often constitute problems for sequencing, primarily due to secondary structure, and therefore the procedures for optimising sequencing results described in Chapter 3 (such as the use of extending mix and careful control of the amount of input DNA) are particularly relevant here.

The regions flanking the HVR, both 5' and 3' (total 444 bp), as well as a few repeats into the HVR array, were sequenced for 18 M-class and 6 S-class alleles (Table 4.7). The position of the α-haplotyping *Bgl I* polymorphism in the IZHVR array has not previously been identified, and although most of the sequenced M alleles are from type IIIa haplotypes (- - + M Z - - - -), the *Bgl I* recognition sequence was not found within the flanking 3' or 5' sequence. This is contrary to the suggested position of the site 5' to the IZHVR [Higgs *et al.*, 1986], but is consistent with the *Bgl I* digestion products of the IZHVR (Table 2.3, Chapter 2), which indicate that the site is contained within the HVR array ~400 bp from the 5' or 3' end of the PCR product. This is beyond the region sequenced in this study and therefore the mutation causing the *Bgl I* (+) allele was not identified. Inspection of the IZHVR sequence indicates that possible *Bgl I* (+) sites are present (with a single mutation) [Goodbourn *et al.*, 1983] within the repeat sequence; this implies that the *Bgl I* (+) site could be present in different repeats in the array in different alleles and therefore be due to

Table 4.7: Sequence Variation Flanking and Contained within Initial Repeats of the IZHVR




| Sample | Allele Class | Haplotype | Polymorphic Position | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 127 | 179 | 1292 | 1295 | 1300 | 1317 | 1339 | 1345 | 1367 | 1450 | 1513-17 |
| | | | G | C | T | T | A | G | A | C | A | T | (G)=5 |
| Austra Aborig | M/M | IIIa/IIIa | CC | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. |
| Micronesia | M/M | IIIa/IIIa | CC | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. |
| Sri Lanka | M/M | IIIa/IIIa | CC | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. |
| Austra Aborig | M/M | IIIa/IInovel | CC | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. |
| W.Samoa | M/M | Ia/Ia | .. | .. | .. | .. | GG | .. | .. | .. | .. | .. | .. |
| Micronesia | S/S | IIf/IIf | .. | .. | CC | CC | .. | .. | TT | GG | CC | CC | 4 |
| New Guinea H | S/S | IVa/Vb | .. | TT | | | .. | AA | TT | GG | CC | CC | .. |
| Iceland | M/M | IIIa/IIIb | CC | .. | | | | | | .. | .. | .. | .. |
| Micronesia | M/M | IIIa/IIIa | CC | .. | | | | | | .. | .. | .. | .. |
| Senegal | M/M | IIIa/IIIf | CC | .. | | | | | | .. | .. | .. | .. |
| Bantu | S/S | IVa/VII | .. | TT | | | | | | | | | |
| Senegal | M/M | IIc/VIIc | .. | .. | | | | | | | | | |

= Sequence information not available

Sequence polymorphisms are numbered according to position in IZHVR sequence given in Appendix A4.3 (from Goodbourn *et al.*, [1983]), and given as the sequence of the 5'-3' strand, although positions 1292 to 1517 were typed using a primer (IZ.SR1; Appendix 3) which gives the sequence of the 3'-5' strand. Positions 127 and 179 were typed using IZ.SF1 (Appendix 3).

multiple mutation. However, no differences in size between *Bgl I* (+) fragments of digested PCR products (Chapter 2) have been detected, and this evidence, coupled with the limited haplotype distribution of the (+) allele [Higgs *et al.*, 1986], indicates that the site is probably in the same repeat unit in all (+) alleles. Thus, the *Bgl I* (+) allele on different haplotypes can be considered equivalent, as are the other RFLP sites contained within the α-haplotype (Chapter 3).

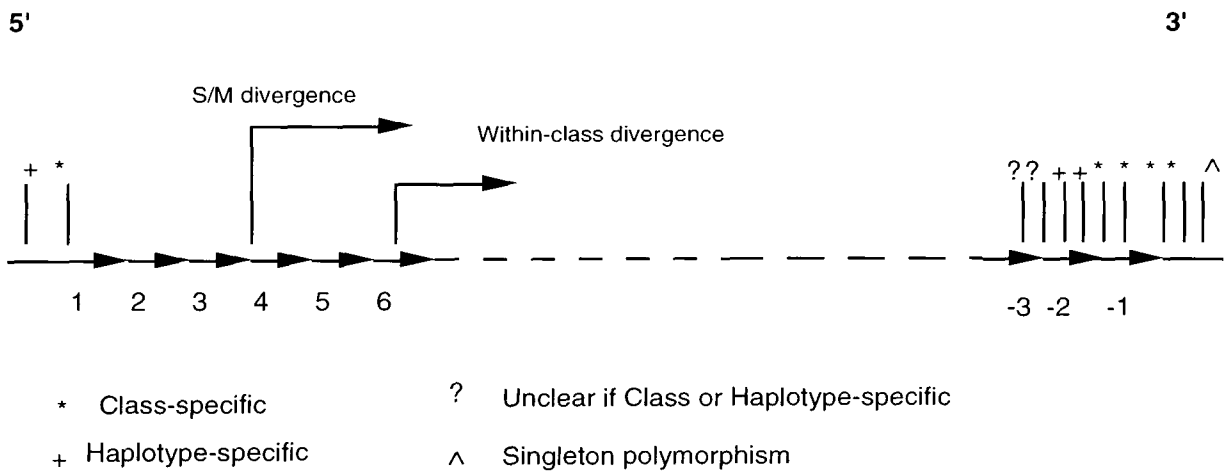
A total of eleven polymorphic sites were found within the sequenced IZHVR alleles (refer to Appendix 4 for IZHVR sequence); of these, five sites appear to be class-specific (S/M) sites of variation: 179, 1339, 1345, 1367, and 1450 (Table 4.7). Within the M class, a haplotype-specific mutation at site 127 is present, distinguishing M alleles on the IIIa haplotypes from those on other haplotypes (Table 4.7). One sample, homozygous for the rare S allele associated with the *Bgl I* (+) allele, does not have the S-associated mutation at site 179 in the 5' flanking region, although it does possess S-associated sequences at sites 1345 and 1367 in the 3' flanking region. It therefore seems possible that the *Bgl I* (+) site has been transferred to an S from an (+) M allele (perhaps on a IIIa haplotype, which is common in the Pacific) by gene conversion. This conversion event has not produced a concurrent size change in the recipient repeat array, and occurred at the 5' end of the IZHVR array, suggesting that the *Bgl I* site is situated at the 5' end of the array.

The presence of these size class-specific mutations suggests that conversion between M and S class alleles, of the sort that could have formed the IIf haplotype, are rare; interactions are more frequent within the same allele classes, resulting in within-class sequence homogenisation. Within the 5' end of the IZHVR array, differences in sequence between S and M alleles were not found until after the first four repeats, after which sequences diverge (Figure 4.9). M alleles have the same sequence extending up to six repeats 5' into the array; beyond this, M *Bgl I* (+) and M *Bgl I* (-)

alleles diverge in sequence (Figure 4.9). Two sequence polymorphisms were found in the 5' flanking region and up to repeat six in the array (Table 4.7); this compares with nine sequence polymorphisms in the 3' flanking region and three repeats into the array (Table 4.7; Figure 4.9). This corresponds to a level of sequence polymorphism at the 5' and 3' ends of the IZHVR (L, Figure 4.9; Equation 3.3, Chapter 3) which is high compared to that observed in the non-repetitive regions (Chapter 3; Table 3.3), although the polymorphism at the 3' end of the HVR is an order of magnitude higher. This indicates a considerable polarity in the accumulation of sequence variation at the IZHVR locus (Figure 4.9), consistent with many other VNTRs, as will be discussed later in this chapter.

---

Figure 4.9: Polarity of Sequence Divergence and IZHVR Allele Class Divergence




Polymorphic sites flanking and contained within the IZHVR. More polymorphisms are present at the 3' end than at the 5' end: L=0.0122 versus 0.00446, respectively.

---

## Characterisation of Variation at the IVS1.HVR Locus

### Size Variation at the IVS1.HVR locus in the ψζ1 and ζ1/ζ2 Genes

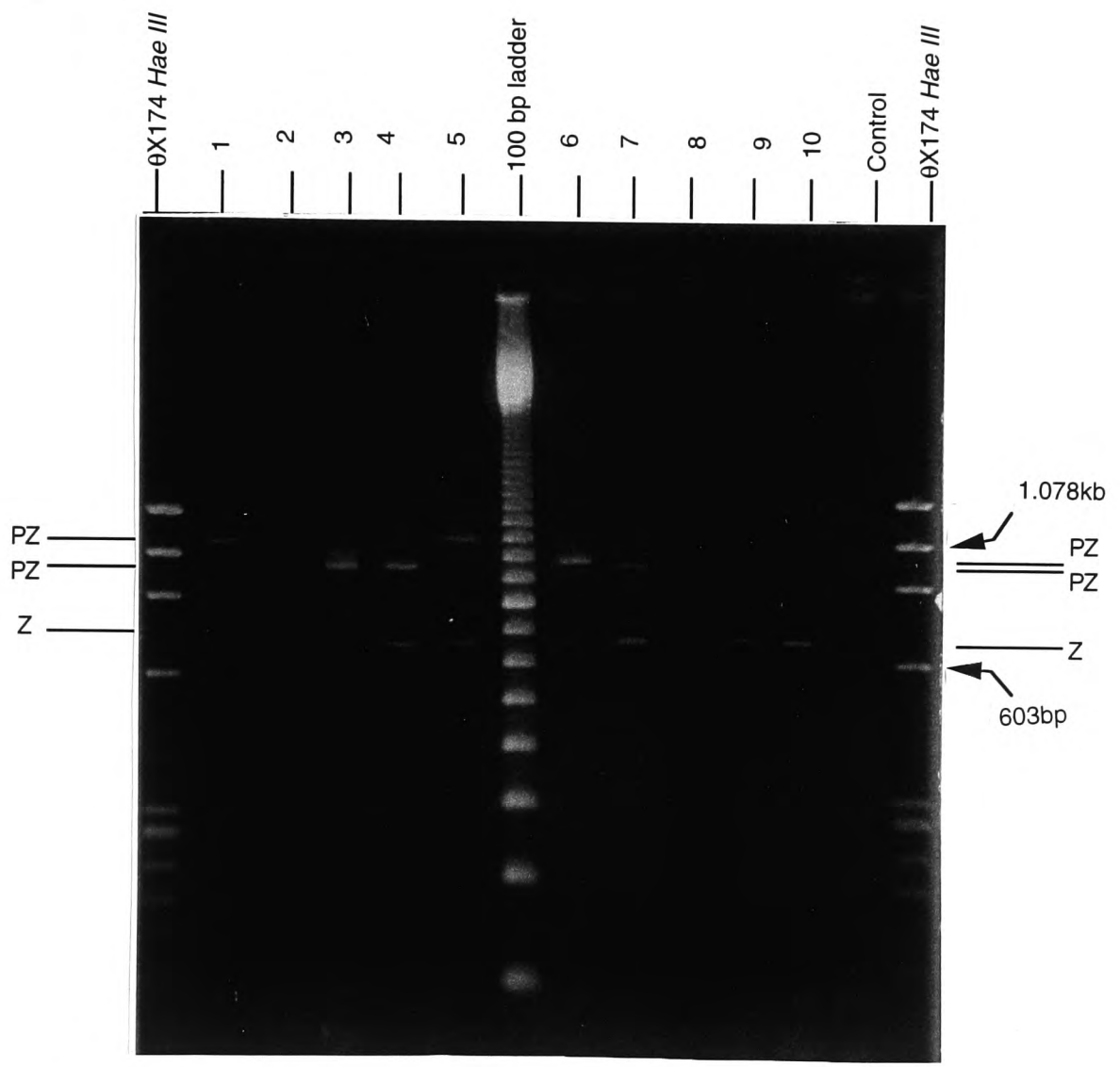
The potential exists, as at the IZHVR, for characterisation of further variation at the IVSI.HVR locus (Figures 4.2 & 4.4) by virtue of the PCR haplotyping protocol. The

amplified region of the ζ genes includes 5' flanking sequence, exon 1 and part of intron 1 including the HVR (refer to Figure 4.4 and Appendix 4). Size variation at the ψζ1/ζ1 genes was detected by the use of the downstream locus-specific primer, PZ.F1 (Chapter 2) and either reverse primer PZ.R7 (used for α-haplotyping, Chapter 2) or PZ.R2 (Appendix 5) (Figure 4.10) which is positioned closer to the 3' end of the HVR. Amplification of the IVS1.HVR within the ζ2 gene was achieved by the use of an upstream-specific primer, ζ2.F3 (Appendix 5) and the primer PZ.R7.

26 human chromosomes of known haplotype, and 2 chimpanzee chromosomes, were sized at the ψζ1 IVS1.HVR (Table 4.8)(Figure 4.10). One of the more common alleles, estimated to have a 39-repeat array, corresponds to that previously reported in both human [Proudfoot *et al.*, 1982] and chimpanzee [Willard *et al.*, 1986]. Nine differently sized alleles were found in humans, more than have previously been reported [Chapman *et al.*, 1986]. Much of the allelic variation was present as relatively small deviations from the "standard" $PZ_{39}$ allele, but alleles which were very different in size (e.g. $PZ_{53}$ and $PZ_{33}$; Table 4.8) were found in African samples, suggesting higher African diversity (as noted at the IZHVR, and from α-haplotypes and sequence variation; Chapters 2 and 3).

16 chromosomes were characterised at the IVS1.HVR in the ζ1 gene from both group III and VII haplotypes (Table 4.8), and, in contrast to the ψζ1 gene, alleles consisted of an invariantly sized 16-repeat array. This is consistent with the argument that the ζ1 conversion represents a single, recent event [Hill *et al.*, 1985]. A previous study, using Southern blotting, has reported that the ζ2 IVSI.HVR has two sizes of allele, consisting of 16 or 12 repeats [Hill *et al.*, 1985]. This dimorphism was confirmed by PCR in a survey of 492 chromosomes of known haplotype, and including previously underrepresented African populations (Figure 4.11). Most populations have approximately equal numbers of the 16 and 12 repeat alleles, although the

Figure 4.10: IVSI.HVR Allelic Variation




ψζ1 (PZ) and ζ1 (Z) IVS1.HVR alleles on 2% agarose; three differently sized ψζ1 HVR alleles are present, but there is no detectable variation within the ζ1 alleles. The control is water.

Table 4.8: IVS1.HVR Size Variation

| Gene | Sample | IVS1.HVR Allele | Haplotype |
|---|---|---|---|
| ψζ1 | Senegal | 41/49 | Ia/IIc |
| | W.Samoa | 39/39 | Ia/IVa |
| | Britain | 39/39 | Ia/IVc |
| | Britain | 37/41 | Ia/VIa |
| | Iceland | 41/41 | Ia/IIa |
| | Brunei | 43 | IIa |
| | Austr Aborig | 41 | IIc |
| | Senegal | 41/45 | IIc/VIb |
| | Philippines | 37/43 | IIc/IIe |
| | Senegal | 49 | IIe |
| | Ponape | 37/43 | VInovel/Vc |
| | San | 53/53 | Vc/IIg |
| | New Guinea H | 39/39 | Vd/IVb |
| | Bantu | 49 | Vc |
| | Senegal | 43 | IIc |
| | Bantu | 41 | IVa/VIIi |
| | Chimp | 39/39 | "II/II" |
| ζ1 | Senegal | 16/16 | IIIa/IIIf |
| | Iceland | 16/16 | IIIa/IIIb |
| | Austr Aborig | 16 | IIIa |
| | Brunei | 16 | IIIa |
| | Senegal | 16 | IIId |
| | Senegal | 16 | VIIc |
| | Senegal | 16/16 | VIIa/VIIa |
| | Kenya | 16/16 | VIIf/VIIf |
| | Senegal | 16/16 | VII/VII |
| | Senegal | 16/16 | VIIa/VIIa |

Alleles are named according to the estimated number of 14 bp repeats. ψζ1 gene alleles are sized according to a system which resolves only two repeats confidently (see text) so sizes are estimates and are open to revision at greater resolution. Refer to Table 3.1 for RFLP status of haplotypes.

Figure 4.11: Variation at the ζ2 IVSI.HVR

A)




B)

| Population | ζ2 Allele | | Total | Source |
|---|---|---|---|---|
| | 16 | 12 | | |
| Senegal | 38 | 46 | 84 | This study |
| Sri Lanka | 33 | 53 | 86 | " |
| Iceland | 33 | 54 | 87 | " |
| Austr Aborig | 13 | 89 | 102 | " |
| Kenya | 48 | 50 | 98 | " |
| Bantu | 1 | 3 | 4 | " |
| San | 2 | 0 | 2 | " |
| Polynesia | 7 | 7 | 14 | " |
| Bangladesh | 8 | 12 | 20 | " |
| North European | 0 | 5 | 5 | Hill et al, 1985 |
| Mediterranean | 1 | 4 | 5 | " |
| Asian Indians | 2 | 4 | 6 | " |
| SE Asia | 0 | 6 | 6 | " |
| Isl. Melanesia | 0 | 15 | 15 | " |
| New Guinea H | 0 | 14 | 14 | " |
| Saudi Arabia | 1 | 5 | 6 | " |
| Jamaica | 0 | 8 | 8 | " |
| Total | 187 | 375 | 562 | |

A) The 12 and 16 repeat alleles at the ζ2 IVS1.HVR are easily distinguished on agarose (3% NuSieve, GTG). B) Populations typed in this and a previous study are shown; H= highlands, SE= South-east, Isl.= island.

Figure 4.11: Variation at the ζ2 IVSI.HVR

A)



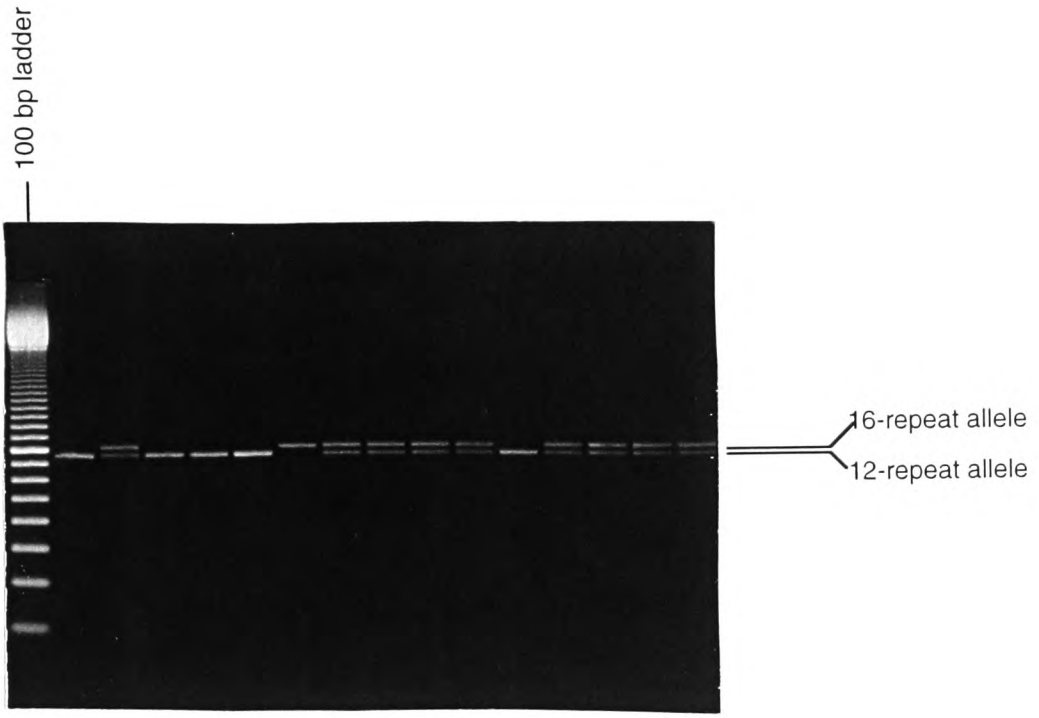

B)

| Population | ζ2 Allele | | Total | Source |
|---|---|---|---|---|
| | 16 | 12 | | |
| Senegal | 38 | 46 | 84 | This study |
| Sri Lanka | 33 | 53 | 86 | " |
| Iceland | 33 | 54 | 87 | " |
| Austr Aborig | 13 | 89 | 102 | " |
| Kenya | 48 | 50 | 98 | " |
| Bantu | 1 | 3 | 4 | " |
| San | 2 | 0 | 2 | " |
| Polynesia | 7 | 7 | 14 | " |
| Bangladesh | 8 | 12 | 20 | " |
| North European | 0 | 5 | 5 | Hill et al, 1985 |
| Mediterranean | 1 | 4 | 5 | " |
| Asian Indians | 2 | 4 | 6 | " |
| SE Asia | 0 | 6 | 6 | " |
| Isl. Melanesia | 0 | 15 | 15 | " |
| New Guinea H | 0 | 14 | 14 | " |
| Saudi Arabia | 1 | 5 | 6 | " |
| Jamaica | 0 | 8 | 8 | " |
| Total | 187 | 375 | 562 | |

A) The 12 and 16 repeat alleles at the ζ2 IVS1.HVR are easily distinguished on agarose (3% NuSieve, GTG). B) Populations typed in this and a previous study are shown; H= highlands, SE= South-east, Isl.= island.

Figure 4.11: Variation at the ζ2 IVSI.HVR

A)




B)

| Population | ζ2 Allele | | Total | Source |
|---|---|---|---|---|
| | 16 | 12 | | |
| Senegal | 38 | 46 | 84 | This study |
| Sri Lanka | 33 | 53 | 86 | " |
| Iceland | 33 | 54 | 87 | " |
| Austr Aborig | 13 | 89 | 102 | " |
| Kenya | 48 | 50 | 98 | " |
| Bantu | 1 | 3 | 4 | " |
| San | 2 | 0 | 2 | " |
| Polynesia | 7 | 7 | 14 | " |
| Bangladesh | 8 | 12 | 20 | " |
| North European | 0 | 5 | 5 | Hill et al, 1985 |
| Mediterranean | 1 | 4 | 5 | " |
| Asian Indians | 2 | 4 | 6 | " |
| SE Asia | 0 | 6 | 6 | " |
| Isl. Melanesia | 0 | 15 | 15 | " |
| New Guinea H | 0 | 14 | 14 | " |
| Saudi Arabia | 1 | 5 | 6 | " |
| Jamaica | 0 | 8 | 8 | " |
| Total | 187 | 375 | 562 | |

A) The 12 and 16 repeat alleles at the ζ2 IVS1.HVR are easily distinguished on agarose (3% NuSieve, GTG). B) Populations typed in this and a previous study are shown; H= highlands, SE= South-east, Isl.= island.

Aboriginal Australians showed a predominance of the 12-repeat allele (Figure 4.11). This allele was previously reported to be predominant in world populations [Hill *et al.*, 1985]; however, more Melanesian and New Guinea highland population samples were studied than from any other population and since these populations have affinities with the Aboriginal Australian population [Roberts-Thomson *et al.*, in press], predominance of the 12-repeat allele is expected.

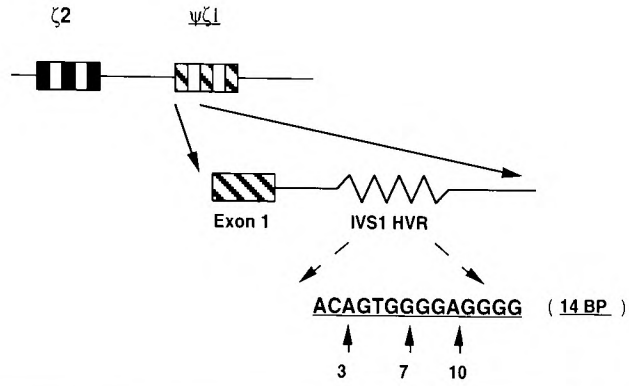
**Sequence Variation within the IVS1.HVR**

Extensive Variation within the IVS1.HVR in the ψζ1 Gene

Incomplete sequences of the ψζ1 IVSI.HVR were obtained from 22 human and 2 chimpanzee chromosomes (Table 4.9), using the sequencing protocols outlined in Chapter 3. Diploid sequences were obtained for samples homozygous for the ψζ1 gene. Single allelic sequences were obtained for four chromosomes from individuals who were heterozygous for ψζ1 and ζ1 genes (haplotype status, PZ/Z) (Table 4.9); the two differently sized alleles were separated on low-melting point agarose and extracted using β-Agarase (Appendix 1; A1.5.3). The primer ψζ.SF2 (Appendix 3)(Figure 4.12) was used to sequence the 5'-3' strand of the IVSI.HVR and provided information for repeats 1-18 in the array of the diploid samples. It was not possible to sequence the 3'-5' strand due to sequence failure close to the HVR, despite the use of a number of sequence primers at various positions relative to the HVR. Therefore, sequence information for repeats 19-39 is not available, although, for the purposes of this study (the determination of the presence of any underlying sequence differences between alleles of the same size) the incomplete sequence data were sufficient.

The repeat sequences obtained were aligned from the 5' end, allowing a human consensus sequence of repeats 1-18 to be determined (Table 4.9 and Figure 4.13). In the samples studied, the 14 bp repeat unit was regular in size, but differences in

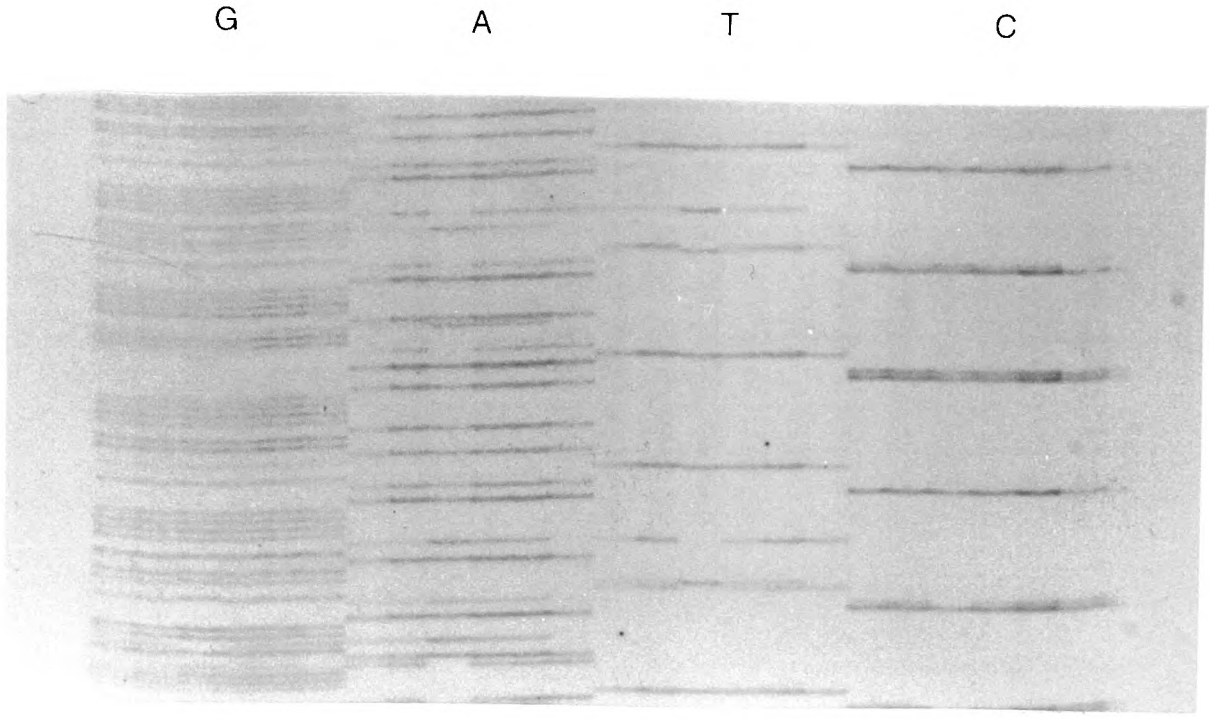
Table 4.9: ψζ1 IVS1.HVR Sequences




| | | Repeat Number and Position of Polymorphism in 14 bp Repeat | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | Haplotype | 1 | 2 | 3 | 4 @ 9 | 4 @ 13 | 5 @ 9 | 5 @ 12 | 6 @ 3 | 6 @ 10 | 7 | 8 @ 6 | 8 @ 9 | 9 @ 3 | 9 @ 8 | 9 @ 10 | 10 | 11 @ 4 | 11 @ 6 | 12 @ 7 | 13 @ 3 | 13 @ 7 | 13 @13 | 14 @ 3 | 14 @ 7 | 14 @ 13 | 15 @ 3 | 15 @7 | 16 @3 | 16 @7 | 17 @3 | 17 @7 | 18 @13 |
| | | | | | G | G | G | G | A | A | | G | G | A | G | A | | G | G | G | A | G | G | A | G | G | A | G | A | G | A | G | G |
| WS | Ia/IVa | con | con | con | -/- | -/- | A/A | -/A | -/C | -/T | con | A/A | -/- | -/C | -/A | -/T | con | -/A | -/A | -/A | -/C | -/A | -/- | -/C | -/A | -/- | -/C | -/A | -/C | -/A | -/C | -/A | -/A |
| GB | Ia/IVc | " | " | " | -/T | -/- | A/A | -/A | -/C | -/T | " | A/A | -/- | -/C | -/A | -/T | " | -/A | -/A | -/A | -/C | -/A | -/- | -/- | -/A | -/- | -/C | -/A | -/- | A/A | -/C | -/A | -/A |
| GB | Ia/VIa | " | " | " | T/T | -/A | -/- | A/A | -/- | -/- | " | -/- | A/A | C/C | A/A | T/T | " | A/A | A/A | G/G | -/- | A/A | -/- | C/C | G/G | -/A | -/- | A/A | C/C | G/G | -/- | A/A | -/- |
| PP | VInovel/Vc | " | " | " | -/T | -/- | A/A | -/A | -/C | -/T | " | A/A | -/A | -/C | -/A | -/T | " | -/A | -/A | -/A | -/C | -/A | -/- | -/C | -/A | -/- | -/C | -/A | -/- | -/A | -/C | -/A | -/A |
| Fil | IIc/IIe | " | " | " | -/T | -/- | A/A | -/A | -/C | -/T | " | A/A | AG | -/C | -/A | -/T | " | -/A | -/A | -/A | -/C | -/A | -/- | -/C | -/A | -/- | -/C | -/A | -/- | -/A | -/C | -/A | -/A |
| PNG | Vd/IVb | " | " | " | -/- | -/- | A/A | -/- | C/C | T/T | " | A/A | -/- | -/- | -/- | -/- | " | -/- | -/- | -/A | C/C | -/A | -/- | -/- | -/A | -/- | C/C | -/- | -/- | -/A | C/C | -/A | A/A |
| Ice | Ia/IIa | " | " | " | T/T | -/A | -/- | A/A | -/- | -/- | " | -/A | A/A | C/C | A/A | T/T | " | A/A | -/- | A/A | -/- | A/A | -/- | C/C | -/A | -/- | -/- | A/A | C/C | G/G | -/- | A/A | -/- |
| Sen | IIc/VIb | " | " | " | -/- | -/- | A/A | -/A | C/C | -/T | " | A/A | -/A | -/C | -/A | -/T | " | -/A | -/A | A/A | -/C | -/A | A/A | -/C | -/A | -/- | -/C | -/A | -/- | -/A | C/C | -/A | -/- |
| San | Vc/IIg | " | " | " | -/- | -/- | A/A | -/- | C/C | T/T | " | A/A | -/- | -/- | -/- | -/- | " | -/- | -/- | A/A | -/C | -/A | A/A | -/- | -/A | -/- | -/C | -/A | -/- | -/A | C/C | -/A | -/- |
| Sen | IIe | " | " | " | T | G | G | A | A | A | " | G | A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Brunei | IIa | " | " | " | T | G | G | A | A | A | " | G | A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| YJ | IIc | " | " | " | T | A | G | A | A | A | " | G | A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Sen | II | " | " | " | G | G | A | G | C | T | " | A | G | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Chimp | "II" | " | " | " | GT | GG | GG | GA | AA | AA | " | GG | GG | AA | AA | AT | " | AA | GG | GG | AA | AA | GG | CC | GG | GG | AA | AA | C | GG | - | - | - |

Sequence variation within 14 bp repeat array. Invariant sequence repeats are shaded. Consensus sequence is boxed (refer to Figure 4.13).

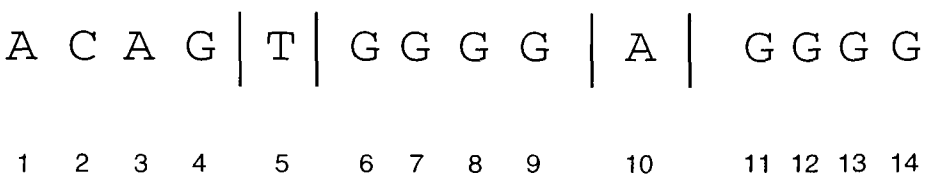
Figure 4.12: Extensive Sequence Polymorphism at the ψζ1 IVSI.HVR Locus




Six samples sequenced with primer ψζSF2 (Appendix 3) and analysed on 5% denaturing acrylamide gel (Appendix 1: A.1.9); termination mixes are run together.

Figure 4.13: ψζ1 IVS1.HVR Sequence

Structure of the 14 bp IVSI.HVR Repeat




"Consensus" Sequence of Repeats 1-18

1. ACAGTGGGGAGCGG
2. GCAGTGGGGAGGGG
3. ACCGTGGGGAGGGG
4. ACAGTGAGGAGGGG
5. ACCATGGGAAGGGG
6. ACAGTGGAGAGGGG
7. ACAGTGAGGAGGGG
8. ACCATAGGGAGGGG
9. ACAGTGGGGAGGGG
10. ACAGTGAGGAGGGG
11. ACCATAGGGAGGGG
12. ACAGTGGGGAGGGG
13. ACCGTGGGGAGGGG
14. ACCGTGGGGAGGGG
15. ACCGTGGGGAGGGG
16. ACCGTGGGGAGGGG
17. ACCGTGGGGAGGGG
18. ACATTGGGGAGGGG

Structure of the 14 bp repeat (positions in 14 bp indicated below sequence). Positions 5 and 10 within the repeat are invariant in sequence in repeats 1-18, and separate three blocks of four bp which demonstrate sequence variability.

sequence between repeats within an array were found, as reported previously [Proudfoot *et al.*, 1982]. Comparison of the sequences of different alleles reveals sequence polymorphism at the variable four-bp domains of the repeat, throughout the array (Table 4.9 and Figure 4.13). For example, in repeat number four, a polymorphic site (G/T) is present at position 9 within the 14 bp repeat sequence; within the same repeat, another polymorphic site (G/A) is present at position 13. These polymorphic positions associate between repeats within an array; that is, they are polymorphic at the same position in numerous repeats. Position 3 (C/A), for instance, is polymorphic in repeat numbers six, nine, thirteen, fourteen, fifteen, sixteen and seventeen (Tables 4.9 & 4.10). Other such widespread polymorphic sites include positions 6, 7, 9, and 10 (Table 4.10). The most plausible explanation for this pattern of variation within the HVR array is that sequence information is being transferred from one repeat to another by recombination: within and between alleles, and in both aligned and misaligned arrangements. It seems likely that gene conversion, rather than crossover, is the predominant recombinatory mechanism, since novel haplotypes are not found at high frequency in most populations (which would be the case if frequent crossover occurred to produce the extensive variation observed).

---

Table 4.10: Spread of Some Polymorphic Positions within the IVS1.HVR Array

| Position | Repeats in which position is polymorphic |
|---|---|
| 3 | 6, 13, 14, 15, 16, 17 |
| 4 | 11 |
| 6 | 8, 11 |
| 7 | 12, 13, 14, 15, 16, 17 |
| 9 | 4, 5, 8 |
| 10 | 6, 9 |
| 12 | 5 |
| 13 | 4, 13, 14, 18 |

---

There are a number of repeats that do not contain any sequence polymorphisms, and whose sequence is constant in all the human and chimpanzee alleles compared. These invariant sequence repeats are numbers one to three, seven and ten in the IVS1.HVR array (Table 4.9); their presence suggests that alignment of the HVR sequences from the 5' end is valid, and also that sequence differences between alleles are accumulating in a polar manner, as observed at the IZHVR. Repeats one to three form an invariant 42 bp block, have different sequences to each other and the sequences of repeats one and two are not found elsewhere in the repeat array (Figure 4.13) [Proudfoot *et al.*, 1982]. The invariant repeats, seven and ten, are identical in sequence to each other and also to other repeats in the array (Figure 4.13). The chimpanzee allele is the same size as one of the common human alleles ($PZ_{39}$), but its underlying sequence is different; in Table 4.9, only the sequence of the chimp at positions which are polymorphic in humans is noted (although some positions are polymorphic in both species). However, the chimpanzee sequence is different at a large number of other positions compared to the human sequence [Willard *et al.*, 1986], which were not shown here for ease of comparison. Nonetheless, the invariant repeats were also found in the chimpanzee sequenced by Willard *et al.*, [1986], and therefore, their presence is not a peculiar feature of the chimpanzee sample sequenced in this study.

Sequence Variation at the IVS1.HVR within the ζ2/ζ1 Genes

The IVSI.HVR in the converted ζ1 gene of 16 chromosomes of group III and VII haplotypes and from different populations was sequenced. Since these alleles consist of sixteen repeats, they could be sequenced in their entirety with primer ψζ.SF2. However, just as there was no size variation of the ζ1 IVS1.HVR alleles, these HVR alleles show no variation in sequence; this is further support for a recent origin for the converted gene [Hill *et al.*, 1985]. The sequence of the HVR was in good

agreement with the previously published sequence [Hill *et al.*, 1985] (Figure 4.14), with only one change found at position ten in repeat four; since all sixteen chromosomes showed the same sequence at this position, it is possible either that this is the position of a rare polymorphism or that the original sequence is erroneous.

---

Figure 4.14: ζ1 IVS1.HVR Consensus Array Sequence

| Repeat No. | Sequence | |
|---|---|---|
| 1. | ACAGTGGGGAGGGG | |
| 2. | GCGGTGGGGAGGGG | |
| 3. | ACAGTGGGGAGGGG | |
| 4. | ACCATGGAGAGGAG | |
| 5. | ACAGTGGGGAGGGG | |
| 6. | ACAGTGGGGAGAGG | |
| 7. | ACAGTGAGGAGGGG | ----- |
| 8. | ACCATGGGGAGGAC | |
| 9. | ACAGTGGGGAGGGC | |
| 10. | ACTGTGGGGAGGGG | ----- |
| 11. | ACAGTGAGGAGGGG | |
| 12. | ACCTTGGGGAGGGG | |
| 13. | AGAGTGAGGAGGGA | |
| 14. | ACCGTGGAGAGGGG | |
| 15. | ACAGTGAGGAAGGG | |
| 16. | ACAGTGAGGACAGA | |

The consensus (invariant) sequence of the ζ1 16-repeat HVR alleles. Lines indicate the region deleted to form the 12-repeat allele (between repeats 7 and 10).

---

The sequence of the ζ1 (and thus the progenitor ζ2) IVS1.HVR is related to that of the ψζ1 HVR, and features of invariant repeat structure described in the ψζ1 HVR are present [Hill *et al.*, 1985] (Figure 4.14). Invariant repeats one and two are related to those of the ψζ1 HVR, and repeat seven remains identical in sequence in both the ζ2/ζ1 and ψζ1 HVRs. The invariant repeat number ten in the ψζ1 HVR is identical in sequence to repeat number eleven (and seven) in the ζ HVRs. The smaller 12-repeat allele of the ζ2 HVR was produced by the deletion of four repeats from the 16-repeat

allele [Hill *et al.*, 1985]; this deletion event occurred between the invariant repeat seven and its homologue, repeat eleven, thereby removing repeats seven to ten, and it most likely represents an intramolecular recombination event.

## Mutational Mechanisms at α-Globin HVRs

### Neutrality

The variation at the IVS1.HVR in the ψζ1 gene is unusual in that the size and sequence of some repeats and positions of two polymorphic sites of the HVR, were found to be consistent in both chimpanzees and humans (Table 4.9). Neutral polymorphisms are not expected to persist for the 5 myrs that separate the human and chimpanzee lineages (see Chapter 6), and the existence of specific sizes of both the IVS1.HVR and IZHVR in humans and chimpanzees has led to the suggestion that alleles may be maintained by constraint on the mutational process [Chapman *et al.*, 1986]. Alternatively, although generally considered to be neutrally evolving loci, HVR alleles could persist under the influence of selection. The latter is an interesting possibility given evidence that the underlying sequence of HVRs does not seem to be random [Flam, 1994] and the evident association of some microsatellite and VNTR variation with disease.

In the VNTR flanking the insulin gene, some size alleles are associated with susceptibility to diabetes [Bell *et al.*, 1982; Bennett *et al.*, 1995; Lucassen *et al.*, 1993; Lucassen *et al.*, 1995], and some alleles at the H-Ras HVR associate with an increased risk of cancer [Green and Krontiris, 1993; Krontiris, 1995]. The insulin HVR locus is related in sequence to HVRs in the α-globin cluster (Figure 4.1), and has a similar trimodal size distribution to that of the IZHVR [Bell *et al.*, 1982]; therefore the α-globin HVR variation could also have functional significance. However, this would

seem more likely to be the case for IVS1.HVR which is situated within the introns of the ζ genes (including the functional human ζ2 gene), rather than the IZHVR which is positioned some distance from likely regulatory sequences of any of the α-globin genes.

To investigate the possibility of a selective influence, sufficient data on IZHVR distributions in the African populations (Mandenka and Luo) are available to compare with the allelic distributions predicted from the neutral mutation/infinite alleles model (IAM) [Kimura, 1983]. It seems more likely that HVRs will conform to the IAM than RFLP haplotypes (see Chapter 2), since the putative mechanisms of minisatellite evolution (USCE and replication slippage; Chapter 1) could theoretically produce an infinite number of alleles (although see Harding *et al.*, [1992] and Harding *et al.*, [1993]). The observed homozygosity can be used to estimate the expected number of alleles within a population by Equation 4.3:

$$E(k) = 1 + \theta/(\theta+1) + \theta/(\theta+2)..... + \theta/(\theta+ m-1) \qquad (4.3)$$

where m = the number of genes examined, θ = H/(1-H) and H = heterozygosity [Ewens, 1972]. The expected distribution of allele frequencies (the theoretical allele count configuration) can be predicted, from sampling theory, using the sample size and observed number of alleles, and the fit between the expected and actual configurations can be tested statistically [Ewens, 1972; Watterson, 1978]. 1000 allele configurations and expected distributions of homozygosity, commonest allele frequency, number of singletons and expected number of alleles were generated [Fuerst *et al.*, 1977; Clark, 1987](Table 4.11). On the basis of these criteria, the IAM could not be rejected. The results of this test alone are, however, insufficient to determine conclusively whether the IZHVR is evolving neutrally. The test has been criticised for being weak and inappropriate for HVRs [Deka *et al.*, 1991; Harding *et*

Table 4.11: Testing African IZHVR Distributions for Consistency with Neutral Theory

| Population | No of Chromosomes | No. of alleles | | No. of singletons | | | Freq of commonest allele | | | Homozygosity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Observed | Expected | Observed | Expected | p value | Observed | Expected | p value | Observed | Expected | p value |
| Kenya | 158 | 11 | 13.5 | 2 | 2.43 | 0.456 | 0.342 | 0.422 | 0.67 | 0.229 | 0.272 | 0.594 |
| Senegal | 230 | 12 | 15.8 | 3 | 2.45 | 0.677 | 0.37 | 0.42 | 0.583 | 0.214 | 0.272 | 0.677 |

*al.*, 1992; Martinson *et al.*, 1993]. Nevertheless, given its location in the α-globin gene cluster, it does seem improbable that the IZHVR is constrained by selection.

However, the size of the ζ2 IVS1.HVR may be limited by selection to the 12 and 16 repeat alleles (assuming that alleles of this size are still recombinogenic); an intron which is too large in size might not be spliced efficiently, or might interfere with some regulatory role played by the intronic sequence or even the HVR itself. For example, intronic sequences have a role to play in the regulation of the human β-globin [Collis *et al.*, 1990] and adenosine deaminase genes [Aronow *et al.*, 1989], and also in some *Drosophila* genes [Laurie and Stam, 1994]. Perhaps in the ψζ1 gene, the HVR has been able to expand in size following release from selection pressure when the gene became inactivated. Counter to this argument, the chimpanzee "PZ" gene is apparently functional [Willard *et al.*, 1986], and yet possesses a large HVR size (39 repeats); this may indicate that there is no selective effect on the size of the IVS1.HVR, or that the "PZ" gene in the chimpanzee is actually a pseudogene despite having an apparently functional sequence (as is the case for the converted ζ1 gene [Hill *et al.*, 1985]).

The presence of invariant repeats within the IVS1.HVR in different species is intriguing. Structural constraint has been noted at the apolipoprotein B 3' HVR locus [Ellsworth *et al.*, 1995], and the positioning of invariant repeats at the 5' end of the HVR may cause the formation of hairpin structures in this region of the array. The conservation in different species of secondary structure would imply that the HVR may have some functional role. In the absence of conclusive evidence, however, it is perhaps more likely that the unusual features of internal structure are caused by the mechanisms of HVR evolution, and particularly by mutational constraints.

**Mutational Constraint and Polarity**

The allelic distributions of the IZHVR and IVS1.HVR suggest the operation of mutational constraint; IZHVR distributions are bi- or trimodal in most human populations, and intermediate-sized alleles are generally not found (although an allele intermediate in size between M and L classes has been found in an Aboriginal Australian population [Roberts-Thomson *et al.*, in press]). Variation within the M size class is continuously distributed but S allele variation (and possibly also L allele class) is not. Size variation at the IVS1.HVR is uni-modal and probably continuous (at the appropriate level of size resolution); the absence of class sizes akin to those of the IZHVR, however, suggests that it may be constrained to fairly small size changes.

Constrained HVR evolution makes the IAM an inappropriate model; in this case, the stepwise mutation model (SMM)[Kimura and Ohta, 1978] which envisages size changes consisting of only one or a few repeats may be more appropriate [Deka *et al.*, 1991], and is consistent with evidence that mutant VNTR alleles are similar in size to their progenitors [Jeffreys *et al.*, 1988]. The model is also generally appropriate for microsatellite mutation [Sclotterer and Tautz, 1992; Valdes *et al.*, 1992] because replication slippage seems to explain, perhaps simplistically, their evolution. It is not clear to what extent VNTRs evolve by slippage, but unequal sister-chromatid exchange (USCE) with strong mutational constraints is certainly feasible [Harding *et al.*, 1993], and would mimic slippage. In terms of the IZHVR trimodal distribution, constrained USCE could explain the majority of alleles being closely related in size to one another, with size classes being created by much rarer, larger events.

The relative predominance of the M class IZHVR alleles in human populations [Higgs *et al.*, 1986] could be explained if array size is linked to mutation rate.

Computer simulations and interspecific comparisons have shown that large alleles (such as the L class of IZHVR) with correspondingly high mutation rates, can become extinct very easily and hence are more transient in evolutionary time [Gray and Jeffreys, 1991; Harding *et al.*, 1992; Harding *et al.*, 1993]. Shorter alleles, such as those in the S class of IZHVR alleles, may be more stable, with a lower mutation rate, and thus persist and show less variation. Therefore, although M alleles show the greatest size variation and are most common, they may not be the oldest IZHVR class; it has not proven possible to determine the ancestral state since the chimpanzee locus could not be amplified with human primers.

The theory of constrained USCE as a predominant mechanism of variation would create the expectation of a structural or sequence feature of the IZHVR array that would permit only small misalignments between arrays on sister chromatids. The mutational polarity evidenced by the positioning of invariant repeats at the 5' end of the IVS1.HVR array and the preferential accumulation of sequence differences between alleles at the 3' end of the IZHVR array, suggest another possibility: gene conversion. From the sequences of these HVRs, conversion clearly has an important role to play in creating patterns of variation. Conversion is capable of causing both size and sequence changes [Klein, 1995], and is consistent with the types of complex mutational events seen at the IVS1.HVR and other loci [Buard and Vergnaud, 1994; Fletcher, 1994; Jeffreys *et al.*, 1994]. The polarity of size and sequence variation observed in this and other (although not all, see Armour *et al.*, [1996]) studies of HVRs [Armour *et al.*, 1993; Desmarais *et al.*, 1993; Jeffreys *et al.*, 1994; Monckton *et al.*, 1994], suggests the presence of an initiator region at one end of the repeat array for gene conversion events (refer to Chapter 1). An initiator region may be present at the 3' end of both the IZHVR and IVS1.HVR arrays, facilitating conversion between homologous chromosomes as well as between sister chromatids (also see Jeffreys *et al.*, [1994]). It is interesting that polarity is a feature of the mutational events at these

α-globin HVRs because other studies have utilised HVRs with very high mutation rates. The IZHVR and IVS1.HVR, with mutation rates likely to be lower than that of the 3' HVR (<0.4%; [Jarman *et al.*, 1986]), indicate that polarity may be a property of many HVRs, and not just those which are exceptionally variable.

## HVRs and α-Haplotypes

Many questions about the specific mechanisms governing the formation of the variation at the IVS1.HVR and IZHVR remain unanswered, and are unanswerable with the population data presented here. Indeed, the processes operating to determine variation at repeated sequence loci may operate in concert, making these questions difficult to answer [Dover, 1993]. However, further elucidation requires more direct analyses of the mutation processes, such as is possible by the study of mutational events in sperm [Jeffreys *et al.*, 1994]. In practical terms, however, the variation present at the HVR loci described in this chapter has a number of implications for comparisons involving α-haplotypes. Alleles which are considered to be equivalent in the α-haplotyping strategy, such as M-class IZHVR alleles, or "PZ" (ψζ1 HVR) alleles, are, in fact, not equivalent in terms of size or sequence between individuals, haplotypes or haplotype groups (although the invariant ζ1 alleles are an apparent exception to this).

Allelic variation at the IVS1.HVR is greater than that observed at the IZHVR, perhaps due to its smaller repeat unit. Clearer association between IZHVR allele size and haplotype is notable; for example, most Ia haplotypes associate with an $M_{40}$ allele, whereas the same haplotype is associated with at least two different IVS1.HVR alleles (Tables 4.3 & 4.8). The IZHVR locus is thus potentially informative for haplotype relationships. However, it is clear that allele size is not the only important consideration when investigating HVR variation. In agreement with

many other studies of the internal structure of repeated sequences (for example, [Buard and Vergnaud, 1994; Decorte *et al.*, 1994; Desmarais *et al.*, 1993; Jeffreys *et al.*, 1990; Jeffreys *et al.*, 1994]) extensive internal sequence variation is also present. Thus, it is important to determine whether alleles of the same size are actually the same at the sequence level. Without such information it is impossible to compare alleles of the same size confidently and to derive conclusions about haplotype relationships, or to consider the likely mechanisms of HVR evolution.

The size and sequence diversity at the ψζ1 IVS1.HVR is not clearly linked to haplotype (Table 4.9). The partial sequence information from single alleles indicates that haplotypes of the same group do not carry the same IVSI.HVR sequence (Table 4.9). This is because gene conversion is occurring probably relatively frequently between arrays on different haplotype backgrounds. This is in contrast to the suggestion from 3' HVR data that inter-allelic exchange between heterologous chromosomes is a rare occurrence [Martinson *et al.*, 1994]. However, if inter-allelic exchanges are primarily resolved as gene conversion, as is suggested from this and other studies [Jeffreys *et al.*, 1994], it would not be possible to identify these events from studying size variation. In this respect, the IVS1.HVR is different to the IZHVR in which interactions are generally limited to occur within size classes, and to some extent, within haplotypes. Again, gene conversion is implied as a major mechanism influencing patterns of variation, but the large differences in size between the different allele classes, particularly between S and M/L, may reduce its frequency. Thus, S class alleles ($S_{12}$ and $S_{11}$) are similar in size and sequence, with only the relatively rare allele, $S_9$, having a considerably different internal structure and size. The presence of class-specific mutations is supportive of sequence homogenisation, and therefore sequence exchange, within classes, but not between classes. Evidence for haplotype group-specific sequences also exists (site 127; Table 4.9), indicating that the $M_{40}$ and $M_{41}$ alleles on the Ia and IIIa, may not be equivalent at the sequence

level, and possibly that inter-allelic exchange between heterologous haplotypes is less frequent than at the IVS1.HVR.

Thus, the assumption based upon sequence data, that haplotypes and haplotype groups are valid ways to summarise sequence variation at the α-globin gene locus, does hold despite the variability at the HVR loci contained in the haplotype. However, it is clear that, with increasing levels of resolution, the haplotype nomenclature increasingly becomes a simplification. Nevertheless, due to the properties of the IZHVR - similarly sized alleles are likely to be related to each other and more so, if on the same haplotype - it will be informative to consider the haplotype context of its variation in more detail. This approach, coupled with the sequence information gained in the other regions of the gene cluster, will help address not only further questions of haplotype mutation, but also of α-haplotype relationships.

# Chapter 5

# α-Globin Haplotype Relationships

## Sequence Data from Non-Repetitive Regions and Haplotype Context

It was shown in Chapter 3 that the Sac, Rsa and α Pst regions of the α-haplotype contained sites, other than the RFLPs, which were polymorphic in a large number of the samples tested. These regions are likely to be informative for investigating haplotype relationships. The six commonly polymorphic sites, Sac: 240, Sac: 421, Rsa: 347, Rsa: 331, Rsa: 330, α Pst: 347, were characterised in 122 chromosomes (as described in Chapter 3) in which either full haplotype or haplotype group was known (from previous RFLP analysis; D R Higgs, personal communication). The number of chromosomes from each haplotype group is primarily related to their frequencies in world populations. The patterns of variation of the six dimorphic sites across the α-globin cluster, from 5' end to 3' end of the haplotype, will be referred to as sequence alleles. 117 of the 122 chromosomes were unambiguously assigned a sequence allele (Table 5.1); this high level of success was aided considerably by the fact that sites Rsa: 331-330 behave as a dinucleotide polymorphism (Chapter 3), and also that some haplotype groups (as will become clear) have characteristic sequence alleles allowing heterozygous samples to be assigned allelic sequences.

Some sites are more polymorphic than others, indicated by differences in site heterozygosity (calculated according to Equation 4.1; Chapter 4) (Table 5.2). Rsa: 347 has the highest heterozygosity, and most haplotype groups are polymorphic at this site (Table 5.1); its homologue, α Pst: 347 is less polymorphic, with the (A) allele being less common. As discussed in Chapter 3, it is likely that the (A) allele arose initially on Zα2 (and presumably some time ago since this allele is commonly

# Chapter 5

# α-Globin Haplotype Relationships

## Sequence Data from Non-Repetitive Regions and Haplotype Context

It was shown in Chapter 3 that the Sac, Rsa and α Pst regions of the α-haplotype contained sites, other than the RFLPs, which were polymorphic in a large number of the samples tested. These regions are likely to be informative for investigating haplotype relationships. The six commonly polymorphic sites, Sac: 240, Sac: 421, Rsa: 347, Rsa: 331, Rsa: 330, α Pst: 347, were characterised in 122 chromosomes (as described in Chapter 3) in which either full haplotype or haplotype group was known (from previous RFLP analysis; D R Higgs, personal communication). The number of chromosomes from each haplotype group is primarily related to their frequencies in world populations. The patterns of variation of the six dimorphic sites across the α-globin cluster, from 5' end to 3' end of the haplotype, will be referred to as sequence alleles. 117 of the 122 chromosomes were unambiguously assigned a sequence allele (Table 5.1); this high level of success was aided considerably by the fact that sites Rsa: 331-330 behave as a dinucleotide polymorphism (Chapter 3), and also that some haplotype groups (as will become clear) have characteristic sequence alleles allowing heterozygous samples to be assigned allelic sequences.

Some sites are more polymorphic than others, indicated by differences in site heterozygosity (calculated according to Equation 4.1; Chapter 4) (Table 5.2). Rsa: 347 has the highest heterozygosity, and most haplotype groups are polymorphic at this site (Table 5.1); its homologue, α Pst: 347 is less polymorphic, with the (A) allele being less common. As discussed in Chapter 3, it is likely that the (A) allele arose initially on Zα2 (and presumably some time ago since this allele is commonly

## Table 5.1: Sac, Rsa and α Pst Sequences and α-Haplotypes

| Haplotype Group | Subtype | Sac 240 | Sac 421 | Rsa 347 | Rsa 331 | Rsa 330 | aPst 347 | Region | Number | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | | C | 15 | G | C | G | G | | | |
| I (PZ + + - -) | Ia (+ + - M) | T | | | | | | Africa/ M East | 5 | |
| | | T | | | | | | Pacific | 2 | |
| | | T | | | | | | Europe | 5 | |
| | | T | | | | | | Asia/ SE Asia | 3 | |
| | Ib (+ - - M) | T | | | | | | Africa | 1 | |
| | Ic (- + - M) | | | | | | | Africa | 1 | |
| | Id (- + - L) | | 14 | | | | | SE Asia | 2 | |
| | | | 14 | A | G | A | A | SE Asia | 1 | 20 |
| II (PZ + - - -) | IIa (+ - - L) | | | A | G | A | | Europe | 1 | |
| | | | 14 | A | G | A | | M East | 1 | |
| | | | 14 | | | | | SE Asia | 1 | |
| | | | 14 | A | | | | SE Asia | 2 | |
| | | | 14 | A | | | | Asia | 1 | |
| | | | | A | | | | Asia | 1 | |
| | IIc (+ + - M) | T | | A | G | A | A | Africa | 1 | |
| | | T | 14 | A | G | A | | Africa | 1 | |
| | | | | A | | | | Austr Aborig | 1 | |
| | | | | A | | | | Africa | 2 | |
| | | | | A | | | | N Guinea C | 1 | |
| | | | 14 | A | | | | Africa | 1 | |
| | | T | 14 | A | | | | SE Asia | 1 | |
| | | T | | A | | | | SE Asia | 3 | |
| | IIe (+ - - S) | | | A | | | | Austr Aborig | 1 | |
| | | | | A | | | | SE Asia | 1 | |
| | | | | A | | | A | Africa | 1 | |
| | IIg (+ - - M) | | 14 | | | | | Africa | 1 | |
| | | | | A | | | | Africa | 1 | |
| | | | | | G | A | | Africa | 1 | |
| | | | | | G | A | | SE Asia | 1 | |
| | IIj (- + - M) | | | | | | | Africa | 2 | 27 |
| III (Z - - - -) | IIIa (- - + M) | | | A | | | | Europe | 1 | |
| | | | | A | | | | Africa | 1 | |
| | | | | A | | | | Austr Aborig | 7 | |
| | | | | A | | | | SE Asia | 1 | |
| | | | | A | | | | New GuineaH | 2 | |
| | | | | A | | | | New Guinea C | 1 | |
| | | | | A | | | | Pacific | 4 | |
| | IIIb (+ - + M) | | | A | | | | New Guinea H | 2 | |
| | | | | A | | | | Europe | 1 | |
| | IIId (+ - - M) | | | A | | | | Africa | 3 | |
| | IIIf (- - - M) | | | A | | | | Africa | 2 | |
| | III (undefined 5' end) | | | A | | | | Europe | 1 | |
| | | | | A | | | | New Guinea H | 1 | |
| | | | | A | | | | Africa | 1 | 29 |
| IV (PZ + - + +) | IVa (+ - - S) | | 14 | A | | | A | Melanesia | 1 | |
| | | | | | G | A | A | Melanesia | 1 | |
| | | | | | G | A | A | Austra Aborig | 1 | |
| | | | | | G | A | A | Africa | 1 | |
| | | | | | | | A | Polynesia | 1 | |
| | | | | | | | A | Europe | 1 | |
| | | | | | | | A | Africa | 1 | |
| | IVb (- - - S) | | | | | | A | New Guinea H | 1 | |
| | IVd (+ + - M) | T | | | | | A | Africa | 1 | |
| | IV (undefined 5' ends) | | | | G | A | A | Europe | 1 | |
| | | | | | G | A | A | New Guinea H | 1 | |
| | | | | | | | A | New Guinea H | 1 | 12 |
| V (PZ + - - +) | Vb (- - + M) | | | A | G | A | | Asia | 1 | |
| | | | | A | | | | Pacific | 1 | |
| | Vc (+ - - S) | | | | G | A | A | Africa | 1 | |
| | | | | | G | A | A | New Guinea C | 2 | |
| | | | | | | | A | Africa | 1 | |
| | | | | A | | | | Pacific | 2 | |
| | Vd (- - - S) | | | | | | A | New Guinea H | 1 | |
| | V (undefined 5' end) | | | | | | | New Guinea H | 1 | |
| | | | | | | | | | | 10 |
| VI (PZ - - - -) | VIa (+ + - L) | | | | | | | Europe | 1 | |
| | VIb (+ + - L) | | | A | | | | Africa | 1 | |
| | VI (undefined 5' ends) | T | | | | | | Pacific | 1 | |
| | | | | A | | | | Africa | 2 | |
| | | | | A | | | | Austr Aborig | 1 | |
| | | | | | G | A | | Africa | 1 | 7 |
| VII (Z + - - -) | VIInovel (- - + M) | | | | G | A | | Austr Aborig | 1 | |
| | VIIa (+ - - M) | T | | | | | | Africa | 2 | |
| | | | | | | | | Africa | 1 | |
| | VIIg (+ + - S) | | | | | | | Africa | 1 | |
| | | | | A | | | | Africa | 2 | |
| | VIIi (- + - S) | | 14 | A | | | | Africa | 2 | |
| | | | | A | | | | Africa | 1 | |
| | VII (undefined 5' end) | | | | | | A | Africa | 1 | |
| | | | | | | | | | | 11 |
| VIII (PZ + - + -) | VIIIa (+ - - S) | | | | | | A | M East | 1 | 1 |
| | | | | | | | | | | 117 |

The consensus sequence is boxed (D R Higgs, personal communication). Polymorphic sites are named after the region amplified and position in PCR fragment (Sac) or relative to the α-gene Cap site (Rsa and α Pst). Pacific= Micronesia and Polynesia.

associated with nearly all haplotype groups; Table 5.1) and then transferred to Zα1 by a gene conversion event (Figure 3.9; Chapter 3). The resulting Zα1 allele is likely to have arisen relatively recently because of its fairly limited haplotype distribution: characterising group IV haplotypes, being found on ~ 50% of group V haplotypes and on isolated examples of group I, II and VII (Table 5.1). Likewise, the dinucleotide polymorphism, Rsa: 331-330 is relatively widespread on haplotype groups, but because it has not been transferred onto the homologous α2 Z box, it is also likely to be relatively recent in origin; that it has spread between haplotype groups is indicative of frequent interallelic conversion between groups, as well as within groups.

---

Table 5.2: Polymorphic Site Heterozygosity

| Site | Heterozygosity |
|---|---|
| Sac: 240 | 0.336 |
| Sac: 421 | 0.224 |
| Rsa: 347 | 0.494 |
| Rsa: 331-0 | 0.260 |
| α Pst: 347 | 0.305 |

---

At Sac: 240 and Sac: 421, the (T) and $A_{14}$ alleles, respectively, are relatively uncommon (Tables 5.1 & 5.1), although it is likely that these are ancestral (from the chimpanzee sequence; Chapter 3). Their haplotype association is fairly limited: Sac: 240 (T) is characteristic of Ia haplotypes, and is present at low frequency on group II haplotypes; the Sac: 421 $A_{14}$ allele is effectively restricted to group II haplotypes (Table 5.1). The low frequency and limited haplotype distribution of these old alleles may be a consequence of population genetic factors in humans, such as genetic drift. The Sac polymorphisms are situated at the 5' end of the α-haplotype and are not obviously associated with the polymorphisms in the Rsa and α Pst regions at the 3' end of the haplotype. This is to be expected if the 5' and 3' ends of the haplotype

recombine (Chapter 2), and if some of these sites are being "swapped" between chromosomes by conversion (Chapter 3).

The 3' end polymorphisms are also not clearly associated with each other, indicative again of their individual recombinatory activity. As a result, 18 of the possible 32 different sequence alleles are present in the sample (counting Rsa: 331-330 as one polymorphism gives five dimorphic sites, and the number of possible sequence alleles as $2^5$) (Table 5.3). Many of these sequence alleles are associated with more than one haplotype background; for example, sequence allele 10 has the highest frequency and is found associated with five different RFLP haplotype groups (Table 5.3 & 5.4). Many rarer sequence alleles, however, are associated with a number of haplotypes; sequence allele 7 is associated with five haplotype groups, but is present at a frequency of only ~8 %. Such alleles are likely to be either relatively old but have undergone drift, or represent sequence motifs generated by multiple recombination events. The ancestral sequence allele (based on the chimpanzee sequence and the reasoning that the two α-gene sequences must originally have been identical) is likely to have been T, $A_{14}$, G, CG, G; this allele is not found in human populations (Table 5.3). This, as discussed for the Sac sites, is probably a result of the scrambling effect of recombination, as well as population genetic forces such as drift.

The different haplotype groups are associated with different numbers of sequence alleles (Table 5.4). Group II haplotypes are associated with the most sequence alleles (and many with ancestral sequences, such as Sac: 240 (T)); this is not a sample size effect, since an approximately equal number of group III as group II chromosomes were studied but the former are associated with only one allele (Table 5.1). Thus, group II haplotypes have the greatest sequence heterozygosity, with group V, VI and VII haplotypes also being relatively diverse. The low levels of sequence diversity (regardless of geographical origin) associated with groups I and III is

Table 5.3: Frequency of Sequence Alleles Formed from Sac, Rsa and α Pst Regional Polymorphisms

| | Sac: 240 | Sac: 421 | Rsa: 347 | Rsa: 331 | Rsa: 330 | aPst: 347 | Frequency | No.Groups |
|---|---|---|---|---|---|---|---|---|
| | C | 15 | G | C | G | G | (%) | |
| 1 | T | . | . | . | . | . | 16.24 | 3 |
| 2 | T | . | . | . | . | A | 0.85 | 1 |
| 3 | T | . | A | . | . | . | 2.56 | 1 |
| 4 | T | . | A | G | A | A | 0.85 | 1 |
| 5 | T | 14 | A | . | . | . | 0.85 | 1 |
| 6 | T | 14 | A | G | A | . | 0.85 | 1 |
| 7 | . | . | . | . | . | . | 8.55 | 5 |
| 8 | . | 14 | . | . | . | . | 4.27 | 2 |
| 9 | . | . | . | . | . | A | 8.55 | 2 |
| 10 | . | . | A | . | . | . | 35.04 | 5 |
| 11 | . | 14 | A | . | . | . | 5.13 | 2 |
| 12 | . | 14 | A | . | . | A | 0.85 | 1 |
| 13 | . | 14 | A | G | A | . | 0.85 | 1 |
| 14 | . | 14 | A | G | A | A | 0.85 | 1 |
| 15 | . | . | . | G | A | . | 3.42 | 3 |
| 16 | . | . | A | G | A | . | 1.71 | 2 |
| 17 | . | . | A | . | . | A | 1.71 | 2 |
| 18 | . | . | . | G | A | A | 6.84 | 2 |

117 chromosomes were analysed (refer to Table 5.1). The number of haplotype groups each allele is associated with is also shown.

Table 5.4: Sequence Allele Heterozygosity of Haplotype Groups

| Haplotype Group | Sequence Allele | Number (Chrms) | H |
|---|---|---|---|
| I | 1 | 16 | |
| (PZ + + - - ) | 7 | 1 | |
| | 8 | 2 | |
| | 14 | 1 | |
| | | | 0.354 |
| II | 3 | 3 | |
| (PZ + - - - ) | 5 | 1 | |
| | 4 | 1 | |
| | 6 | 1 | |
| | 7 | 2 | |
| | 8 | 2 | |
| | 10 | 8 | |
| | 11 | 4 | |
| | 13 | 1 | |
| | 15 | 2 | |
| | 16 | 1 | |
| | 17 | 1 | |
| | | | 0.854 |
| III | 10 | 29 | |
| ( Z - - - - ) | | | 0 |
| IV | 9 | 6 | |
| (PZ + - + + ) | 12 | 1 | |
| | 18 | 5 | |
| | | | 0.569 |
| V | 7 | 1 | |
| ( PZ + - - + ) | 9 | 2 | |
| | 10 | 3 | |
| | 19 | 1 | |
| | 18 | 3 | |
| | | | 0.8 |
| VI | 1 | 1 | |
| ( PZ - - - - ) | 7 | 1 | |
| | 10 | 4 | |
| | 15 | 1 | |
| | | | 0.66 |
| VII | 1 | 1 | |
| ( Z + - - - ) | 7 | 3 | |
| | 9 | 1 | |
| | 10 | 3 | |
| | 11 | 2 | |
| | 15 | 1 | |
| | | | 0.831 |

H= heterozygosity, which was calculated with correction for sample size according to Equation 4.1 (Chapter 4). Sequence alleles are numbered according to Table 5.3.

striking. As a consequence for population affinity studies, the Ia haplotypes and all group III haplotypes, are strictly comparable with each other. However, any other haplotype may not have the same underlying sequence between, or even in some case, within populations.

There are numerous possible reasons why some haplotypes should be associated with more sequence variation than others. In the human HLA complex, a recombination hotspot is found at the TAP2 gene; some haplotypes which encompass this region, however, show linkage disequilibrium across the hotspot [Cullen *et al.*, 1995]. This can be interpreted to mean either that these haplotypes are of recent origin, or that directional selection is operating to maintain its particular sequence [Cullen *et al.*, 1995]. Also, in the case of both α-globin and HLA loci, many of the observed patterns of variation are determined by recombination, and it is possible that haplotypes differ in their recombinogenic capacity; for example, group III haplotypes could, due to an unidentified mutation, be less recombinogenic than the more highly variable group II. It is difficult to argue the case against selection with the available sequence data, but it is clear that the high degree of variability of group II haplotypes is due to their age, rather than diversifying selection or greater mutability, since the chimpanzee has group II haplotypes (Chapter 3). This suggests that the observed relative sequence variability of the different haplotype groups is a reliable indicator of their relative ages. In this case, it seems likely that group I and III haplotypes are recent in origin.

## HVR Variation and Haplotypes

### Size Variation at the IZHVR

Of the 726 chromosomes characterised for IZHVR size variation in Chapter 4, a total of 320 chromosomes were of known haplotype, sampled from a number of world

populations, including the Senegalese Mandenka for which much haplotype information is available [Martinson *et al.*, 1995]. The associated haplotype and IZHVR size alleles are shown in Table 5.5. The most common alleles in this sample are $M_{42}$, $M_{41}$, $M_{40}$ and $S_{12}$, and these are associated with a large number of different haplotypes (Table 5.5). Most of the rarer alleles also associate with a number of haplotypes, with only a few alleles characteristic of one haplotype or haplotype group, such as the $S_9$ allele which is limited to two group VII haplotypes and $M_{46}$, which was found only on the IIg.

Some haplotype groups are associated with more IZHVR alleles than others; the group II haplotypes show the greatest number of alleles, and highest heterozygosity levels (Table 5.6). Group III haplotypes show the least variability despite being represented by the largest sample size (Table 5.6). It would be misleading to include group IV and V haplotypes in this comparison because they are associated with S alleles which demonstrate little sub-allelic size variation (Chapter 4).

---

Table 5.6: Number of IZHVR on Different Haplotype Groups

| Haplotype Group | No. IZHVR Alleles | No. Chromosomes | Heterozygosity |
|---|---|---|---|
| I | 7 | 84 | 0.594 |
| II | 10 | 93 | 0.762 |
| III | 5 | 101 | 0.578 |
| VII | 7 | 21 | 0.790 |

Heterozygosity calculated by Equation 4.1; Chapter 4.

---

The striking association of Ia and IIIa haplotypes with particular IZHVR alleles, noted in Chapter 4, is evident in this larger data set, being responsible for lower levels of heterozygosity of group I and III haplotypes (Table 5.6): ~ 80 % of Ia haplotypes are associated with $M_{40}$, and ~ 90 % of IIIa haplotypes with $M_{41}$ alleles. α-Haplotype association has also been demonstrated in studies of the haplotype

Table 5.5: IZHVR Size Variation and Its Association with Haplotype

| Haplotype | | IZHVR Allele | | | | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | L59 | M47 | M46 | M45 | M44 | M43 | M42 | M41 | M40 | M39 | M38 | S12 | S11 | S9 | |
| Ia | + + - M PZ + + - - | | | | | | | 6 | 6 | 62 | 2 | 1 | | | | 77 |
| Ib | - + - M PZ + + - - | | 2 | | | | | 1 | | 2 | | | | | | 5 |
| Ic | + - - M PZ + + - - | | | | | 1 | | | | | | 1 | | | | 2 |
| IIa | - + - L PZ + - - - | 1 | | | | | | | | | | | | | | 1 |
| IIa/b | +/- + - L PZ + - - - | 3 | | | | | | | | | | | | | | 3 |
| IIc | + + - M PZ + - - - | | | | | | | 11 | 6 | 13 | | 1 | | | | 31 |
| IId | - - - S PZ + - - - | | | | | | | | | | | | 1 | | | 1 |
| IId/e | +/- - S PZ + - - - | | | | | | | | | | | | 16 | | | 16 |
| IIe | + - - S PZ + - - - | | | | | | | | | | | | 19 | | | 19 |
| IIf | + - + S PZ + - - - | | | | | | | | | | | | 1 | | | 1 |
| IIg | + - - M PZ + - - - | | | 1 | | 1 | 3 | 2 | | 2 | | | | | | 9 |
| IIj | - + - M PZ + - - - | | | | | | 2 | 3 | | 3 | | | | | | 8 |
| IIk | - - - M PZ + - - - | | | | | | 1 | | 1 | | | | | | | 2 |
| IIm | + + - S PZ + - - - | | | | | | | | | | | | | 2 | | 2 |
| IIIa | - - + M Z - - - - | | | | | | | 4 | 48 | 2 | | | | | | 54 |
| IIIb | + - + M Z - - - - | | | | | | 1 | 2 | 8 | 2 | | | | | | 13 |
| IIIc | - + - L Z - - - - | 1 | | | | | | | | | | | | | | 1 |
| IIId | + - - M Z - - - - | | | | | | | 9 | | | | | | | | 9 |
| IIIe | - + - M Z - - - - | | | | | | 1 | | | 1 | | | | | | 2 |
| IIIf | - - - M Z - - - - | | | | | | 4 | 15 | 2 | 1 | | | | | | 22 |
| IVa | + - - S PZ + - + + | | | | | | | | | | | | 6 | 2 | | 8 |
| IVb | - - - S PZ + - + + | | | | | | | | | | | | 1 | 1 | | 2 |
| Vc | + - - S PZ + - - + | | | | | | | | | | | | | 1 | | 1 |
| Vd | - - - S PZ + - - + | | | | | | | | | | | | | 1 | | 1 |
| VIb | + + - L PZ - - - - | 2 | | | | | | | | | | | | | | 2 |
| VIe | + - - M PZ - - - - | | 2 | | 1 | | | | | | | | | | | 3 |
| VIIa | - + - M Z + - - - | | | | | | | 2 | | | | 1 | | | | 3 |
| VIIc | + - - M Z + - - - | | | | 6 | | 1 | 2 | | | | | | | | 9 |
| VIIe | - - - M Z + - - - | | | | | | | 1 | | | | | | | | 1 |
| VIIf | + + - M Z + - - - | | | | | | | 2 | 1 | | | | | | | 3 |
| VIIg | + + - S Z + - - - | | | | | | | | | | | | 1 | | 1 | 2 |
| VIIh | + - - S Z + - - - | | | | | | | | | | | | | | 1 | 1 |
| VIIi | - + - S Z + - - - | | | | | | | | | | | | 2 | | | 2 |
| VIIIc | + + - M PZ + - + - | | | | | | | | 1 | | | | | | | 1 |
| VIIId | + - - M PZ + - + - | | | | | | 2 | | | 1 | | | | | | 3 |
| | Total | 7 | 4 | 1 | 7 | 2 | 15 | 60 | 73 | 89 | 2 | 4 | 47 | 7 | 2 | 320 |

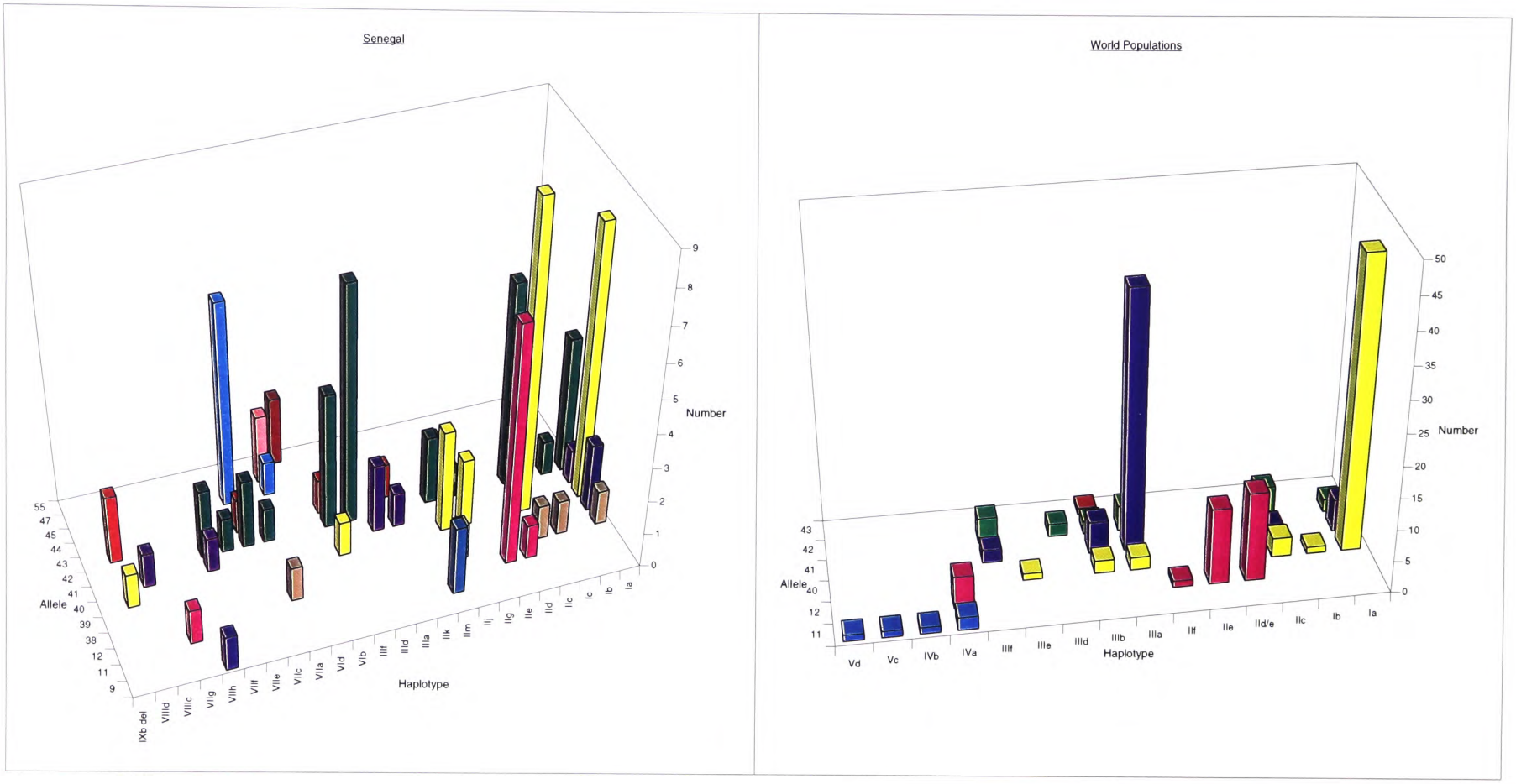
Variation in samples from world populations (refer to Chapter 4 for details), Senegalese Mandenka, Icelandic and Kenyan Luo populations.

context of 5' and 3' HVR variation [Martinson *et al.*, 1994]. A practical implication of association, for some haplotypes, is that ambiguous chromosomes in population surveys can be assigned haplotypes with greater confidence. For example, the presence of an $M_{41}$ allele in a heterozygous sample would add weight to a suggested resolution which included one IIIa haplotype.

The strength of allele and haplotype associations is affected by geographical origin, however, as demonstrated by plotting separately allelic variation and associated haplotype for the Senegalese Mandenka and for an artificial sample set consisting of chromosomes from a number of world populations (Figure 5.1). Outside of Africa, Ia and IIIa haplotypes are predominant and are clearly associated with $M_{40}$ and $M_{41}$ alleles, respectively; within Africa, many more haplotypes and IZHVR alleles are present (Figure 5.1). The African Ia haplotypes, for example, are associated with a number of different alleles, not just $M_{40}$; nevertheless, the Ia is still most commonly associated with $M_{40}$. Also, although the IIIa is rare in Africa, and thus is not represented by many chromosomes, it is associated with an $M_{41}$ allele in Senegal as outside of Africa (Figure 5.1) (Table 5.5).

On younger haplotypes, which the Ia and IIIa are likely to be (from the sequence data discussed earlier), there is clear association with IZHVR size. On older haplotypes, however, a lack of association is evident. This is in part due to the swapping of the 5' ends of the haplotype, which, over sufficient time, will serve to obscure the original associations of HVR size and haplotype. Since association is destroyed, this suggests that when 5' motifs are swapped between haplotypes, the IZHVR variation is included. This would further localise the point of preferential crossover as between IZHVR and the $\psi\zeta1/\zeta1$ genes. Some evidence for this comes from the associated IZHVR size with novel haplotypes. In the Australian Aborigine population, many novel haplotypes have the 5' end of the IIIa, which is extremely

Figure 5.1: IZHVR Size Variation and Its Association with Haplotype



common in the population [Roberts-Thomson *et al.*, in press]. A number of haplotypes with the distinctive ( - - +M) motif found on the IIIa, also are associated with an $M_{41}$ allele (Table 5.7).

The allelic associations of Ia and IIIa haplotypes indicate how linkage disequilibrium becomes weakened across the IZHVR following haplotype formation. The alleles present on the Ia, for example, other than the most common, $M_{40}$, are within a range $M_{42}$-$M_{38}$, but in increments of one repeat. This suggests that mutation upon a haplotype is constrained to small changes in size, in a slippage-type manner (see Chapter 4). Further size changes, and possibly larger, non-incremental changes, are perhaps then caused by the 5' and 3' end recombination outlined above. The ultimate result will be the type of allelic distribution present on group II haplotypes: multi-modal and non-continuous (Table 5.5).

---

Table 5.7: IZHVR Alleles Associated with Novel Haplotypes in the Australian Aborigine Population

| Haplotype | IZHVR Allele |
|---|---|
| IIIa/- - + M Z + + - - | $M_{41}/M_{41}$ |
| VIg/- - + M PZ + - - - | $M_{40}/M_{41}$ |
| IIIa/- - + M PZ + - - - | $M_{41}/M_{41}$ |
| IIIa/- - + M Z + - - - | $M_{41}/M_{41}$ |

---

Despite the lack of association between IZHVR size and most haplotypes, it seems likely that closely related haplotypes and haplotype groups would have similar allelic distributions. The IZHVR M-allele distributions on haplotype groups and a number of full haplotypes within the Senegalese population were compared (non-parametric Mann-Witney U test; Table 5.8). This indicates that many haplotypes are significantly different from one another (Table 5.8). Two pairs of groups (I and II, &

III and VII) are, however, not significantly different to each other (Table 5.8), suggesting that these groups are related. This seems likely given that III and VIIs are the only haplotypes to carry the ζ1 converted gene, which, along with its absence in chimpanzees [Willard *et al.*, 1986], suggests this gene has been produced recently [Higgs *et al.*, 1986; Hill *et al.*, 1985]. The RFLP structure of groups I (PZ + + - -) and II (PZ + - - -) is also consistent with the notion that these groups are related. This relationship is supported by the lack of significant difference between Ia and IIc/IIg haplotypes (Table 5.8), although there are significant differences between the Ia and haplotypes from groups III and VII (Table 5.8).

---

Table 5.8: Comparison of IZHVR Distributions on Different Haplotypes

| | Comparison | | | | | |
|---|---|---|---|---|---|---|
| | I/II | I/III | I/VII | II/III | III/VII | II/VII |
| p value | 0.7896 | 0.0031* | 0.0398* | 0.0023* | 0.7491 | 0.0132* |
| | Ia/Ib | Ia/Ic | Ia/IIc | Ia/IIg | Ia/IIIf | Ia/VIIn1 |
| p value | 0.398 | 0.15447 | 0.2265 | 0.0662 | 0.0071* | 0.0112* |

Significant values, ($p < 0.05$), are indicated with an asterisk.

---

**Size and Sequence Variation at the ζ Genes**

Association of ζ2 IVS1.HVR Size Variation and ψζ1/ζ1 Genes

It was shown in Chapter 4 that there is extensive size variation in the ψζ1 IVS1.HVR, which was not associated with specific haplotypes. In this section, the dimorphic ζ2 IVS1.HVR is considered in relation to the downstream locus (ψζ1 and ζ1 genes) on different haplotypes. The ζ2 IVS1.HVR size of 67 chromosomes associated with a ψζ1 gene of known haplotype or haplotype group, and from a number of

populations, was determined (Table 5.9). The chimpanzee 'PZ' gene is linked to a 16-repeat allele, indicating that this is the ancestral arrangement, which is consistent with the scenario for the production of the 12-repeat allele outlined in Chapter 4. Approximately equal numbers of the 16 and 12-repeat alleles are associated with most haplotypes. Within group II haplotypes, slightly more 16-repeat alleles were found, although this is not significant ($\chi^2$: $p>0.1$) and is attributable to the IIg and IIj haplotypes, which, like group VIII haplotypes, are almost exclusively associated with $\zeta 2$ 16-repeat alleles (Table 5.9). Conversely, the Ia and Ib, and group VI haplotypes, although represented by small sample sizes, are associated exclusively with the 12-repeat $\zeta 2$ allele.

Since the $\zeta 2$-$\psi\zeta 1$ region extends across the region where crossover events between 5' and 3' ends of the haplotype occur, lack of association between the two genes is expected. In the cases where association is found, such as at the Ia haplotypes, the explanation is likely to be recent origin of that haplotype. If this is so, then the haplotypes carrying the recently $\zeta 1$ converted gene, groups III and VII, are expected to show association with $\zeta 2$ IVS1.HVR size. A total of 176 chromosomes carrying the $\zeta 1$ gene were therefore characterised at the upstream $\zeta 2$ HVR (Table 5.9): 139 group III and 37 group VII chromosomes, of varying subtypes and geographical origin.

As expected, the group III and VII haplotypes are strongly associated with only one $\zeta 2$ HVR allele. Contrary to expectation, however, the two groups are linked to different alleles: the VIIs most frequently with the 16-repeat allele, and the IIIs almost exclusively with the 12-repeat allele. This constitutes a significantly different association between the two haplotype groups ($\chi^2$, $p<0.001$). Association of those $\zeta 1$ genes present in non-African populations (and therefore from group III haplotypes) with the 12-repeat $\zeta 2$ HVR allele was found in a previous study, and interpreted as meaning that the $\zeta 1$ conversion was interchromosomal [Hill *et al.*, 1985]. However,

Table 5.9: Linkage of ψζ1/ζ1 Genes to ζ2 IVS1.HVR Alleles on Different Haplotype Backgrounds

| ζ1 Status | Haplotype Group | Subtype | ζ2 IVSI.HVR | | Total |
|---|---|---|---|---|---|
| | | | 16 | 12 | |
| Pseudozeta (PZ) | I (PZ + + - -) | Ia ( + + - M) | - | 4 | 4 |
| | | Ib ( - + - M) | - | 1 | 1 |
| | | Ic ( + - - M) | 1 | - | 1 |
| | II (PZ + - - -) | IIa (- + - L) | 1 | 2 | 3 |
| | | IIc (+ + - M) | 3 | 5 | 8 |
| | | IIe ( + - - S) | 3 | 2 | 5 |
| | | IIg ( + - - M) | 7 | 1 | 8 |
| | | IIj ( - + - M) | 3 | - | 3 |
| | | - | 5 | 6 | 11 |
| | IV (PZ + - + +) | - | 2 | 3 | 5 |
| | V (PZ + - - +) | - | 4 | 2 | 6 |
| | VI (PZ - - - -) | - | - | 5 | 5 |
| | VIII (PZ + - + -) | - | 3 | - | 3 |
| | | Total | 36 | 31 | 67 |
| Zeta (Z) | III (Z - - - -) | - | 3 | 131 | 139 |
| | VII (Z + - - - ) | - | 34 | 8 | 37 |
| | | Total | 37 | 139 | 176 |

ζ2 HVR alleles were amplified and sized as described in Chapter 4. "-" indicates that subtype information is not available.

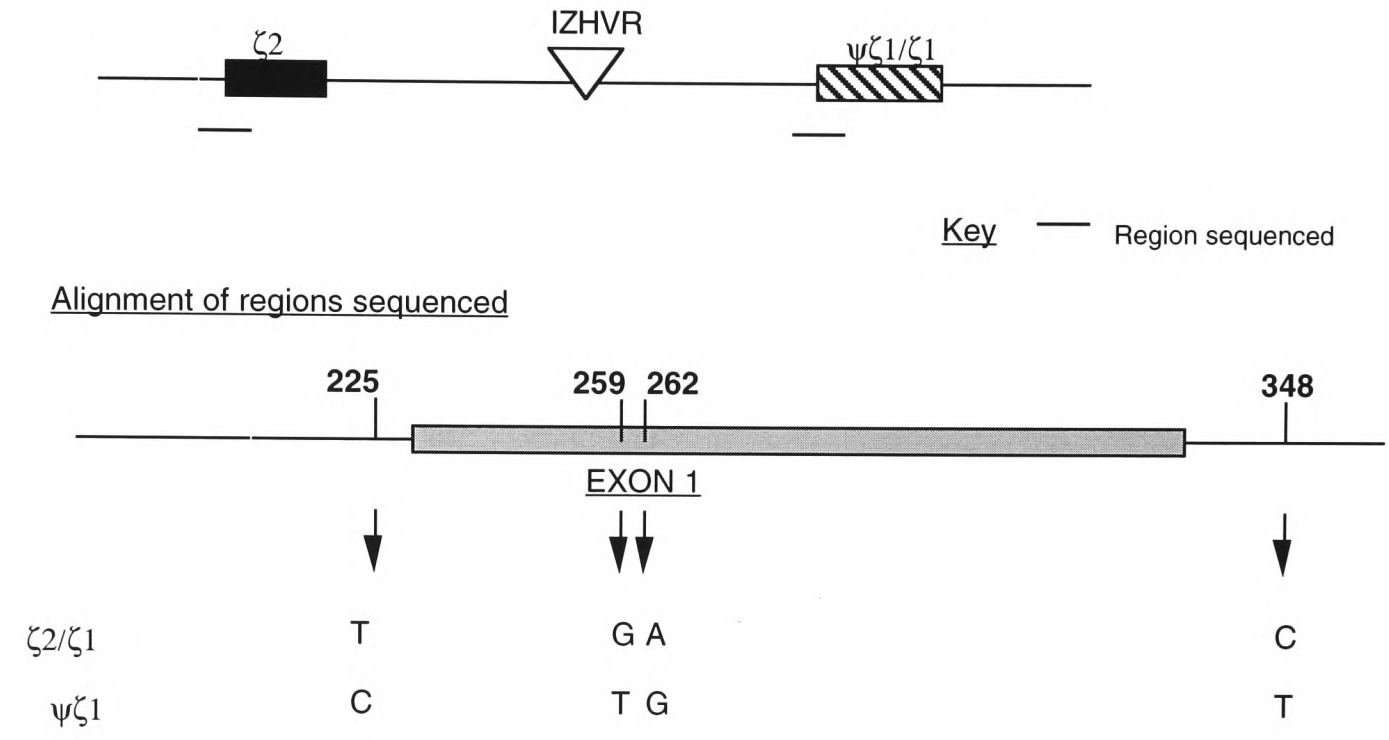
the linkage of group VII ζ1 genes with a different ζ2 HVR is consistent with an intrachromsomal gene conversion event, and thus a separate conversion to that on the group III; this implies that group III and VII haplotypes are not related. It seemed possible to resolve this issue by looking at some of the point mutational differences between the upstream and downstream loci [Proudfoot *et al.*, 1982], in addition to the size differences at the IVS1.HVR.

ζ2/ζ1 Sequence Polymorphism Flanking the IVS1.HVR

There are four sequence differences between the ψζ1 gene and ζ2 genes in the region sequenced in this study (Figure 5.2). Site 348 (in the first intron, 47 bp 5' to the start of IVS1.HVR; Figure 5.2; refer to Appendix 4) can be detected with sequencing primer, ψζ1.SF2, which was used to sequence the HVR of both ψζ1 and ζ2/ζ1 genes (Chapter 4). Two of the sixteen ζ1 genes sequenced in the previous chapter were found to have the ψζ1 sequence at site 348, rather than that of the ζ2 gene as expected (Figure 5.2). These two chromosomes were both from group VII haplotypes. Therefore, 68 further chromosomes of various haplotypes and from a variety of populations were sequenced at site 348 in both upstream and downstream ζ genes. 25 chromosomes were haplotypes carrying the ψζ1 gene, and 43 carrying the ζ1 gene (28 group III, 15 group VII haplotypes) (Table 5.10). The sequence of position 348 at both genes, and size of the ζ2 HVR, are used to classify ζ-sequence alleles (Table 5.10).

Haplotypes carrying the ψζ1 gene are all associated with the predicted ζ-sequence allele (C---T: Figure 5.2; Table 5.10) and when associated with both 16- and 12-repeat ζ2 HVR alleles. The chimpanzee 'PZ' gene is associated with the sequence of the ζ2 gene at position 348 (C--C), which is expected given its apparently functional sequence [Willard *et al.*, 1986]. The ζ1 genes of group III haplotypes are also all

Figure 5.2: Point Mutational Differences between ψζ1 and ζ2/ζ1 Genes




The positions of four sequence polymorphisms (225, 259, 262 and 348) are given relative to the forward PCR primer (ψζ.F1; Chapter 2). All four can be typed with sequencing primer, ψζ1.SF(T) (Appendix 3); 348 and the HVR can be sequenced with ψζ1.SF2 (Chapter 4; Appendix 3). Site 225 can also be detected by the use of a *Pvu II* digestion: the ζ2 sequence is (+), the ψζ1 sequence is (-). Site 262 is the position of the codon 6 inactivating mutation in the ψζ1 gene [Proudfoot *et al.*, 1982].

Table 5.10: ζ2/ζ1/ψζ1 Sequence Haplotypes on Different Haplotype Backgrounds

| Downstream Locus | Haplotype | ζ Sequence Allele | | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| | | 16 C---T | 12 C---T | 16 T---T | 16 T---C | 16 C---C | 12 C---C | |
| Pseudozeta | Ia | | 4 | | | | | 4 |
| | IIa | 1 | 1 | | | | | 2 |
| | IIc | 2 | 2 | | | | | 4 |
| | IIe | 1 | | | | | | 1 |
| | IIj | 2 | | | | | | 2 |
| | IIg | 2 | | | | | | 2 |
| | II | 1 | 2 | | | | | 3 |
| | Vc | 2 | | | | | | 2 |
| | VI | 1 | | | | | | 1 |
| | VId | | 1 | | | | | 1 |
| | VIII | 1 | | | | | | 1 |
| | Chimp | | | | | 2 | | 2 |
| Zeta | IIIa | | | | | 3 | 13 | 16 |
| | IIIb | | | | | | 1 | 1 |
| | IIId | | | | | | 4 | 4 |
| | IIIe | | | | | | 1 | 1 |
| | IIIf | | | | | | 6 | 6 |
| | VIIa | | | | 1 | | | 1 |
| | VIIc | 1 | | | | 4 | | 5 |
| | VIIf | | | | | | 1 | 1 |
| | VIIe | | | | 1 | | | 1 |
| | VII | 2 | | 3 | | | | 5 |
| | --+M Z++-- | | | | | 1 | 1 | 2 |
| | Total | 15 | 10 | 3 | 2 | 11 | 27 | 68 |

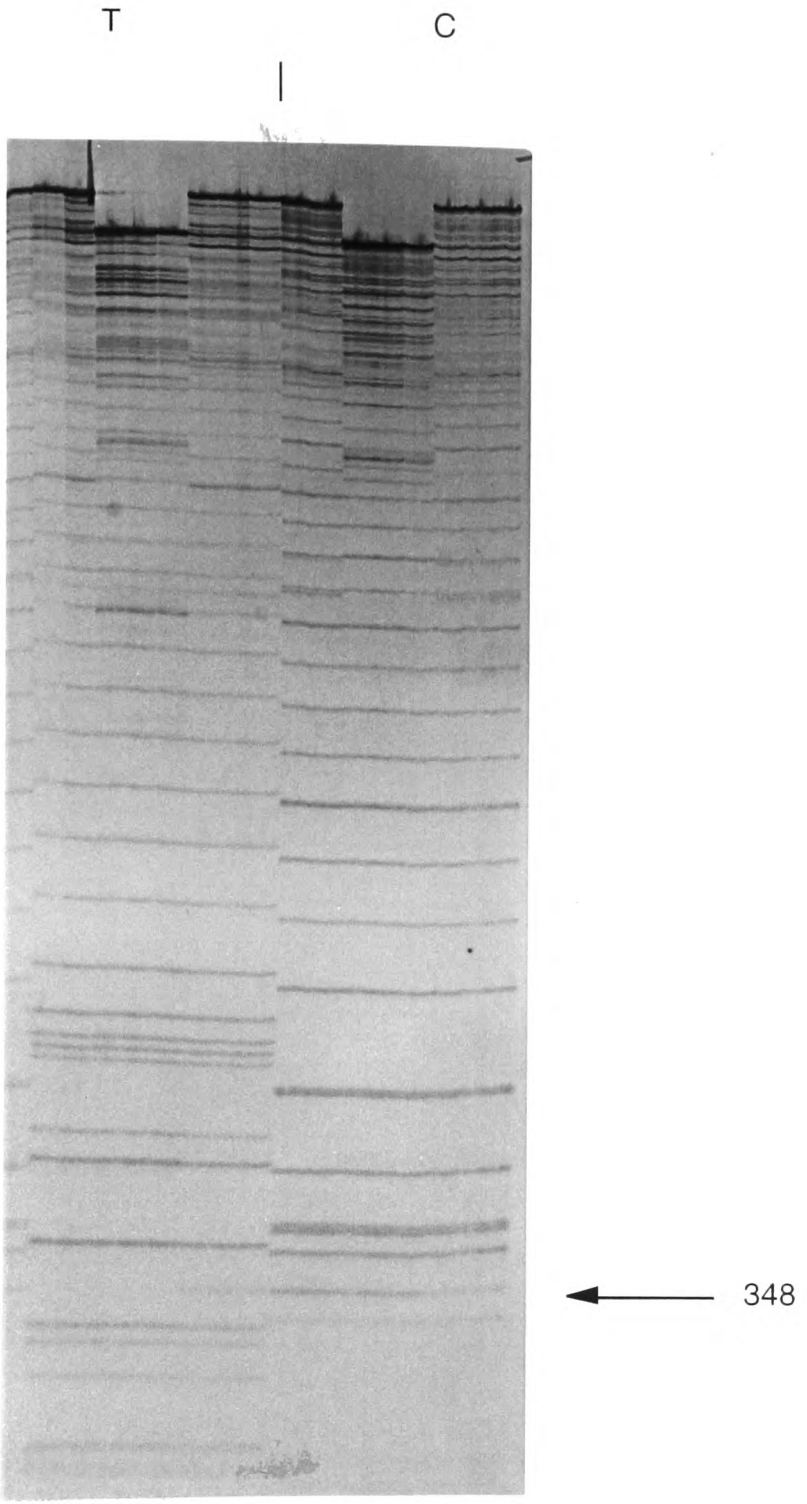
Sixty-eight chromosomes were sequenced with ψζ.SF2 on both ψζ1/ζ1 and ζ2 genes. The size of the ζ2 IVS1.HVR is shown (16 or 12 repeats). Position 348 is shown on ζ2 and ζ1, in 5' to 3' order (i.e. C--T indicates the sequence,T, at 348 on downstream ψζ1 or ζ1 gene). Refer to Tables 5.1, 5.5 and 5.9 for RFLP information of haplotypes shown here.

associated with the expected sequence allele (C--C), which is that of the ζ2 gene. However, the group VII haplotypes are strikingly different to group III, and all other haplotypes, by being associated with a number of different ζ-sequence alleles (Table 5.10).

The most common ζ-sequence allele associated with group VII haplotypes is ζ2-like (C--C), as expected, but three other alleles are also observed (T--T, T--C and C--T) confirming the earlier observation that position 348 is polymorphic on group VII haplotypes, through occasional association with the ψζ1 "specific" sequence (Figure 5.3). Given the rarity of the ψζ1 (T) sequence on the ζ2 gene, and its presence only on ζ2 and ζ1 genes of group VII haplotypes (Table 5.10), this sequence is likely to have initially been acquired by the ζ1 gene in an interallelic gene conversion event with a ψζ1 gene. This event, forming the C--T sequence allele, must have occurred in the 5' flanking region of the IVS1.HVR, but it is clear from the invariant sequence of all ζ1 HVRs (Chapter 4) that this did not involve the closely adjacent HVR itself, perhaps because length heterogeneity of this type inhibits conversion (as suggested for S and M IZHVR alleles, Chapter 4). The other sequence alleles observed, T--T and T--C, indicate that since the transfer of information from ψζ1 to ζ1, both intrachromosomal and interchromosomal conversions have occurred between the ζ2 and ζ1 genes (Figure 5.4).

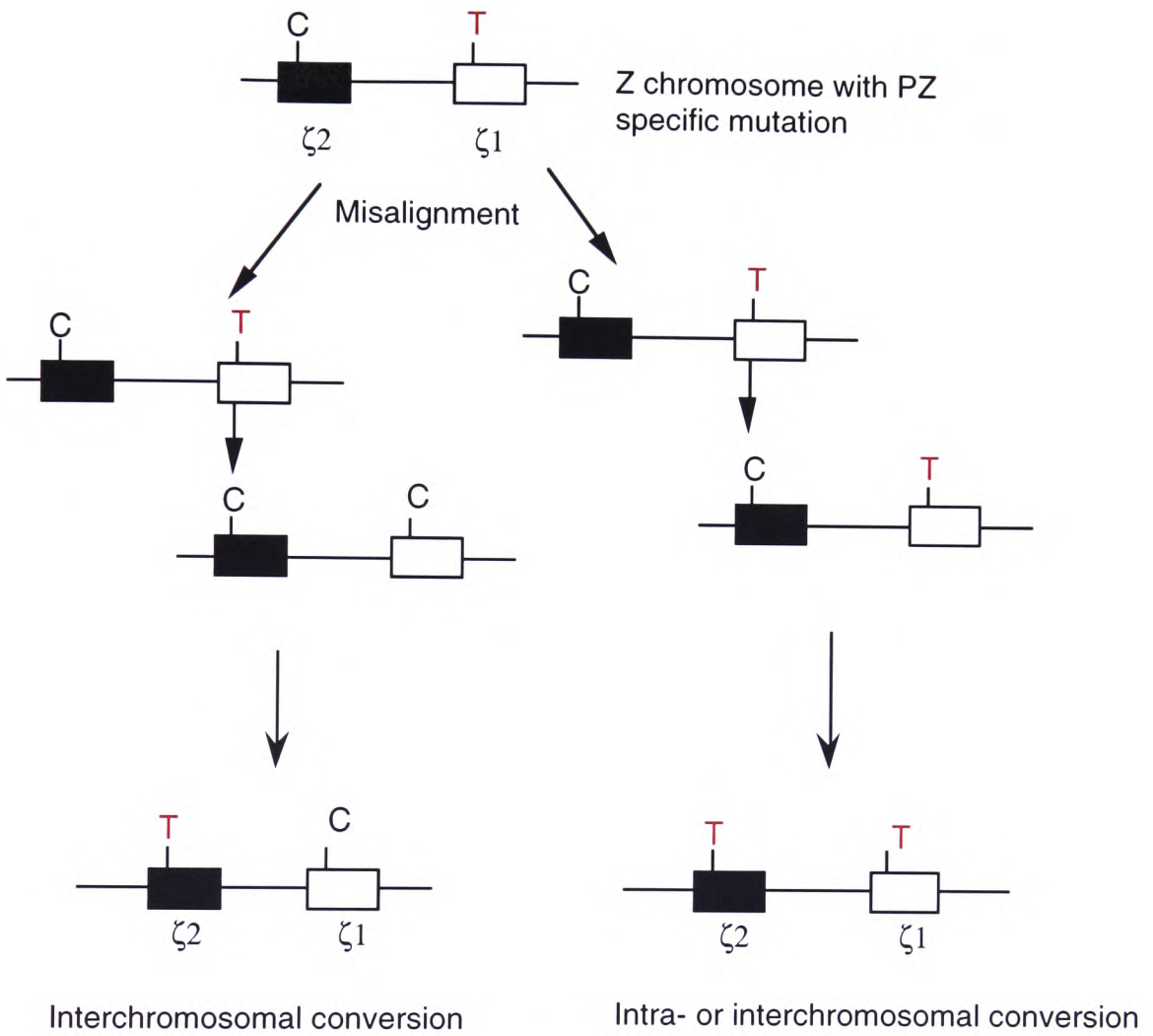
Since position 348 is converted between ζ genes, it seemed likely that some of the other sites which differ between the ψζ1 and ζ2/ζ1 genes (Figure 5.2), might also be transferred in these conversion events. Therefore, 10 chromosomes of group VII haplotypes with the sequence 'T' at position 348 (on either the ζ2 or ζ1 genes) were sequenced using primer, ψζ.SFT1 (Appendix 3) (Table 5.11). This primer types all four sequence differences between the ζ genes 5' to the IVSI.HVR (Figure 5.2). All ζ1 and ζ2 genes, with a T at 348, show no corresponding alterations in sequence at the

Figure 5.3: Sequence Polymorphisms at Position 348 on Group VII Haplotypes




Eight group VII chromosomes sequenced using only ddTTP and ddCTP termination mixes, with primer ψζ1.SF2 (Appendix 3); samples 6 to 8 are T/C heterozygotes at position 348.

Figure 5.4: Conversion Events between ζ Genes Transferring a ψζ1-Specific Mutation




The formation of the T--T sequence allele could occur by intrachromosomal gene conversion (between sister chromatids) or between homologous chromosomes in a C--T/C--T homozygote. Since the C--T allele is rare (Table 5.9), intrachromosomal exchange seems more likely.

other three positions. This indicates that conversions involving position 348 do not extend upstream. In two chromosomes from the same individual (haplotype VIIf/IIg; Table 5.11), the ζ2 genes have the ψζ1-specific sequence at position 225, although they do not have the ψζ1 sequence at 348. This provides evidence for conversion of the ζ2 gene by a ψζ1 gene.

---

Table 5.11: Sequences at Four Positions that Differ between the ζ2 and ψζ1 Genes

| Haplotype | 225 | | 259 | | 262 | | 348 | |
|---|---|---|---|---|---|---|---|---|
| | ζ2 | ζ1 | ζ2 | ζ1 | ζ2 | ζ1 | ζ2 | ζ1 |
| | T | T | G | G | A | A | C | C |
| VIIf/IIg | C/C | -/- | -/- | -/- | -/- | -/- | -/- | T/T |
| II/VII | -/- | -/- | -/- | -/- | -/- | -/- | -/T | T/T |
| VI/New | -/- | -/- | -/- | -/- | -/- | -/- | -/- | -/T |
| VIIf/VIIf | -/- | -/- | -/- | -/- | -/- | -/- | -/T | -/T |
| VIIe/IIIf | -/- | -/- | -/- | -/- | -/- | -/- | -/T | -/- |

---

It is possible that many conversions of ζ2 by ψζ1 would be deleterious (for example, transfer of the codon 6 mutation from ψζ1 would inactivate the ζ2 gene) and therefore, are not observed. Rarity of transfer is confirmed by the fact that no observed haplotypes carrying the ψζ1 gene are associated with a ζ2 gene with a T at position 348 (i.e. no intrachromsomal conversion on these chromosomes) (Table 5.10). The greater homology between the ζ2 and ζ1 genes may enhance conversion and have less functionally significant results. Since conversion of this type clearly does occur on group VII haplotypes, it would be expected also on group III haplotypes. The reason why it has not, is probably that the group III haplotypes arose more recently than group VIIs. This is consistent with all other sequence and HVR data presented so far.

The question of the exact relationship between the group III and VII haplotypes, however, remains unclear. The likelihood that they are related depends, to some extent, on whether a gene conversion event spanning the entire ψζ1 gene to form the ζ1 gene is likely to have occurred more than once in humans. The fact that it has not yet been observed in chimpanzees suggests that the event is unlikely, although there may be some selective pressure against conversion of the "PZ" gene if it is indeed functional in chimpanzees. If haplotype groups III and VII do have the same ζ1 conversion, it is necessary to appeal to population genetic forces, such as drift, to account for their association with differently sized ζ2 IVS1.HVRs.

## Determination of an α-haplotype network of relationships

### Principles of the Approach

The α-haplotype RFLP structure can be used as a framework for a network of haplotype relationships, based on minimising the necessary number of mutational steps; that is, use of the principle of parsimony[1] which states that evolutionary change is improbable. For example, a group I haplotype (PZ + + - -) is more likely to be related to a group II (PZ + - - -), through one mutation at the *Rsa I* site, than to a group VII haplotype (Z + - - -), which would require two mutations. This approach of using RFLP structure is validated by the finding of little in the way of extra polymorphism, except within the Sac, Rsa and α Pst regions (Chapter 3); this extra polymorphism has allowed sub-classification of RFLP-defined variation but has not changed haplotype nomenclature. A similar relationship between sequences and restriction site haplotypes has been reported in the mitochondrial genome [Graven *et al.*, 1995].

[1]Parsimony is being used here as a general principle of phylogenetics rather than the specific statistical technique for drawing phylogenetic trees known as "parsimony".

The process of reconstructing an intraspecific phylogeny or gene tree is not, however, straightforward. Methods designed to estimate between-species phylogenies are inappropriate because intra-species diversity may be low, thus limiting the resolving power, and populations are not reproductively isolated units, as species are. Recently, population genetics theory has been used in an attempt to refine the estimation of intraspecific phylogeny and overcome some of the limitations imposed by inter-specific methods of reconstruction [Crandall and Templeton, 1993]. Assuming a constant population size and random mating, older haplotypes are predicted to have higher frequencies within populations [Donnelly and Tavare, 1986] and possess a larger number of mutational connections to other haplotypes [Castelloe and Templeton, 1994; Crandall and Templeton, 1993; Excoffier and Langaney, 1989]. In addition, the geographic location of rare (low frequency) haplotypes can indicate their origins from other, older haplotypes [Crandall and Templeton, 1993]; in a geographically sub-divided population with limited gene flow, a singleton haplotype is more likely to be connected to a common allele in the same population than to an allele in a different population.

Since so much α-haplotype frequency information is available, this approach is potentially very useful, in combination with the extensive data on sequence and HVR variation described in this and previous chapters. However, recombination can present problems because it destroys historical associations, and it is clear that all of the extra sequence polymorphisms identified (and possibly also the RFLP sites) are undergoing recombination. Segmental exchange of the type responsible for patterns of variation at the sequence level of α-haplotypes, and the HVRs they contain, means that a single diverging sequence tree cannot realistically be determined. Thus, such recombinogenic sites are usually excluded from the determination of gene trees. Nevertheless, earlier in this chapter, recombination has been used as an implicit

indicator of haplotype age; greater recombination indicates older haplotypes, in much the same way as high frequency suggests that haplotypes are older. In this sense, recombinatory activity is useful and it is valid to use it in a phylogenetic context provided that the underlying assumption, that haplotypes recombine at equal frequencies, is not violated.

The obvious exception to this assumption are group VII haplotypes which recombine more frequently than other haplotypes at the ζ loci (Table 5.10). This is not problematic, however, because the ζ loci allelic sequence information is complemented by other sequence and HVR data which also serve as measures of relative amounts of recombination and haplotype age. However, because of recombination (and probably also because of demographic violations of constant population size), it seems likely that conflicts will occur between the relationships suggested from sequencing and HVR data, and from α-haplotype frequencies. Indeed, haplotype frequency may not always the most reliable indicator of age [Castelloe and Templeton, 1994] since, for example, the frequency of a recently arisen haplotype may be elevated by selection.

**The Relationships of α-Haplotypes**

The ancestral haplotype group is the group II. This is based on the chimpanzee sequence (Chapter 3), association with the largest number, and highest heterozygosity, of sequence alleles (Sac, Rsa, α Pst regions; Table 5.1), the largest number and high heterozygosity of IZHVR alleles (Table 5.6), and lack of linkage between ψζ1 and ζ1 (Table 5.9). Additionally, if assigned α-haplotypes in previous world-wide surveys are considered, totalling 2214 chromosomes when deletion chromosomes are excluded [Higgs *et al.*, 1986; Martinson, 1991; Ramsay and Jenkins, 1988; Martinson *et al.*, 1995; O'Shaughnessy *et al.*, 1990], group II haplotypes are the

most common, having the highest frequency and being present in many populations (Table 5.12). Further, the group also has the largest number of subtypes, presumably resulting from the most 5' and 3' end recombination (Table 5.12).

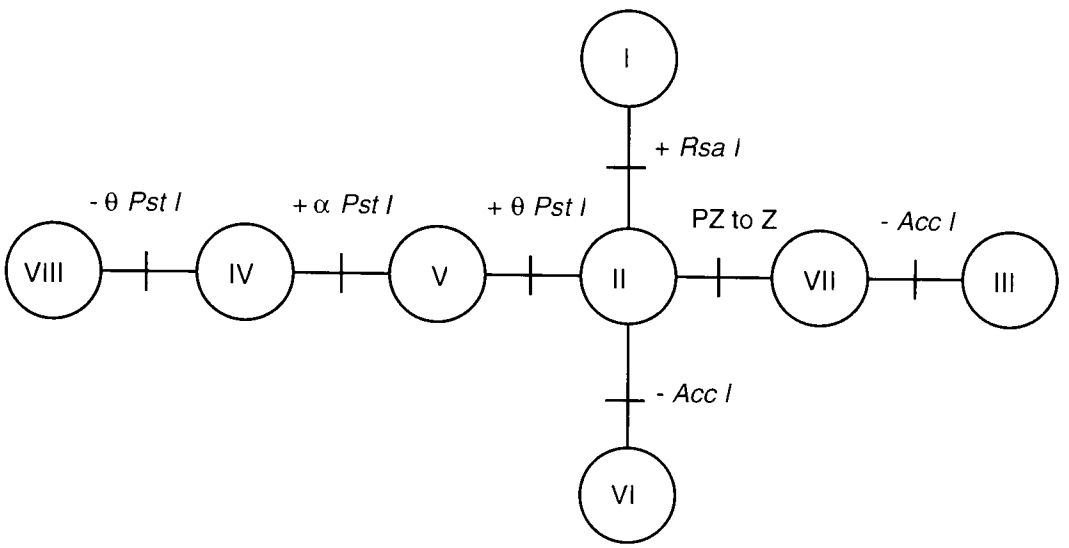
Table 5.12: Frequencies of Haplotype Groups in World Populations

| Group | No. Chromosomes | No. Populations | No. Subtypes |
|---|---|---|---|
| I | 582 | 28 | 7 |
| II | 697 | 28 | 14 |
| III | 498 | 32 | 7 |
| IV | 275 | 14 | 5 |
| V | 55 | 13 | 6 |
| VI | 15 | 5 | 6 |
| VII | 81 | 4 | 8 |
| VIII | 11 | 7 | 4 |

Given that group II haplotypes are oldest, it is expected that this group will have undergone the greatest amount of mutation to give rise to other haplotypes, and hence be connected to more haplotypes in the network than other groups. Indeed, this group can be connected via one mutational step at the RFLP sites to groups I, V, VI, and VII (Figure 5.5).

Group IV and V haplotypes share distinctive sequence alleles at the Sac, Rsa and α Pst regions, particularly in being associated with the α Pst: 347 (A) sequence (Table 5.1) which is only otherwise found rarely on group II and VII haplotypes and not on other groups. Group IV and V haplotypes are also evidently related in terms of their RFLP structure, differing only at the α *Pst I* site (Figure 5.5). Since group V haplotypes are associated with more sequence haplotypes than group IV (Table 5.1) and show greater sequence heterozygosity (Table 5.4), and group IV haplotypes always have the rare (A) sequence at α Pst: 347, group IV can be considered to be derivatives of group V haplotypes. This seems feasible because group V can be directly derived from group II haplotypes (Figure 5.5), but is in conflict with the

Figure 5.5: The α-Haplotype Network Depicting Proposed Relationships Between Major Haplotype Groups




| Group | RFLP Structure |
|---|---|
| I | PZ + + - - |
| II | PZ + - - - |
| III | Z - - - - |
| IV | PZ + - + + |
| V | PZ + - - + |
| VI | PZ - - - - |
| VII | Z + - - - |
| VIII | PZ + - + - |

This outline is based upon RFLP structure, with groups related by one, presumably point mutational, step. The relationships are refined by use of sequencing and HVR information, and consideration of α-haplotype frequency and amount of recombination. The group III haplotypes have an ambiguous origin; they are shown here to be related to group VII haplotypes, the pathway most favoured by the evidence presented in the text.

world-wide frequency data which indicate that group V haplotypes are rarer than group IV haplotypes, although found in a comparable number of human populations (Table 5.12). The weight of evidence suggests, though, that the frequency data are not indicative of age in this case.

Group VIII haplotypes are represented by only one chromosome in this study of sequence variation (Table 5.1) because this group is relatively uncommon (Table 5.12). The sequence data, however, indicate probable relationship with group IV/V haplotypes, since α Pst: 347 has the sequence (A) (Table 5.1), and RFLP structure suggests a specific relationship with group IV (Figure 5.5; Table 5.1). Its low frequency world-wide is consistent with positioning this group at the tip of the haplotype network, and its mutational relationship indicates that although group IV haplotypes are younger than group V haplotypes, they are still old enough to have given rise to another group. Likewise, although the frequency of group V haplotypes is low, their position within the network suggests that they are a relatively old group.

The origins and interrelationships of the remaining haplotype groups (I, III, VI and VII) are less straightforward to unravel. Group I haplotypes are extremely common in world populations being comparable to group II haplotypes in this respect (Table 5.12). This would suggest that they are old haplotypes. However, all other evidence suggests that they are in fact young haplotypes and that the frequency data are misleading; they are primarily associated with one sequence allele at the Sac, Rsa and α Pst, and ζ2-ψζ1 regions (Tables 5.1 & 5.10), and with relatively low IZHVR heterozygosity (Table 5.6). Their associated IZHVR allele distribution (Table 5.5) and RFLP structure suggests that their recent origin is from group II haplotypes, and so form one mutational branch from the root, without further mutational connections (Figure 5.5).

It would seem that the frequency data of group III haplotypes is also misleading in indicating that the group is high frequency, found in more populations than any other group and therefore perhaps one of the oldest groups (Table 5.12). The association of the group with just one sequence allele at the Sac, Rsa and α Pst regions and one ζ2 IVS1.HVR allele (Tables 5.1& 5.9) and low variability at the IZHVR (Table 5.5) suggest that this is probably the most recently arisen group. Accepting that the group must be positioned at the tip of the network, however, does not resolve the question of its origins. Consideration of the group's RFLP structure indicates that it could be mutationally related, by one step, to either group VII or group VI haplotypes (Figure 5.5). Sequence data are consistent with either relationship (Table 5.1); associated IZHVR allele distributions suggest a relationship with group VII haplotypes, but the ζ1-ζ2 linkage data suggests that group III and VII haplotypes are different (Table 5.9) although, as previously discussed, this depends on how likely two independent ζ1 conversions are to have occurred in the human lineage.

Unlike group VII haplotypes, which are only found in Africa, group VI haplotypes are not restricted to one continent [Higgs *et al.*, 1986] suggesting that they may be older and therefore more likely to have been progenitors to the group III haplotypes. However, they are less frequent in these populations than group VII haplotypes are in Africa (Table 5.12). From RFLP structure and sequence alleles (Table 5.1), the progenitors of the group VI are likely to be group II haplotypes. If the frequencies of group II, III and VI haplotype groups in different populations are compared, it might be expected that the frequencies of either groups II or III would be high when the group VI is present: if group II haplotypes are the progenitor group, then they would be expected to be frequent. This comparison shows some correlation between

group II and VI frequency, but little between group III and VI frequencies within geographical regions (Table 5.13).

Table 5.13: Frequencies of Haplotype Groups II, III and VI in World Populations

| | Frequency (%) | | |
|---|---|---|---|
| Region | II | III | VI |
| Taiwan | 55.4 | 5.5 | 0.8 |
| Britain | 11.4 | 6.8 | 2.3 |
| Asian Indian | 31.6 | 2.7 | 1.4 |
| SE Asian | 21.4 | 2.9 | 1.4 |
| Senegal | 33.9 | 19.1 | 2.5 |
| San | 70.3 | 5.4 | 2.7 |
| Austr Aborig | 10.6 | 73.1 | 3.4 |
| New Guinea (H) | 0 | 28.1 | 0 |
| Iceland | 2.6 | 23.7 | 0 |

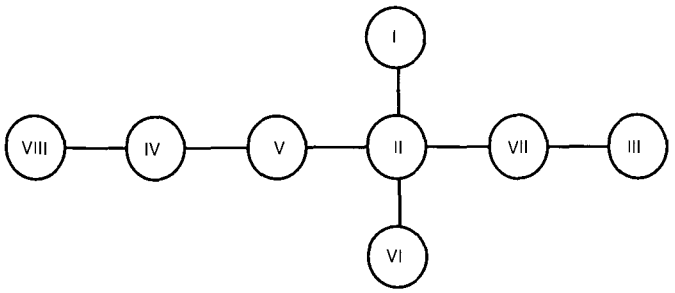
Populations with high frequencies of group II haplotypes have group VI haplotypes, whereas populations such as New Guinea highlands, which have no group II, also have no group VI haplotypes (Table 5.13). Thus, a link between these two groups is possible. One of the problems of this argument, however, is that haplotype groups may not be locally produced; that group VI and III haplotypes are found in a number of different populations, from most continents, indicates that they probably predate modern human dispersal. An exception to this is seen in the Australian Aborigine population, where group IIIa haplotypes are predominant (73%; Table 5.13), with only a few group VI and II haplotypes also being present [Roberts-Thomson *et al.*, in press]. The IIIa (- - + M Z - - - -), is not easily related to the only group VI haplotype present, the VId (+ - - S PZ - - - -). However, the most common group II haplotype present in this population is the IIe (+ - - S PZ + - - -), which could be related to the VId by one mutational event at the *Acc I* site. In this particular case, the group VI haplotypes may have been produced locally, and the high frequency of group III haplotypes is not related to local formation from a group VI haplotype.

Thus, both frequency and sequencing/HVR variation data do not allow the origin of the group III haplotypes to be unambiguously determined. However, it is possible to adjust the frequency data to allow more accurate estimation of haplotype age. Population genetic simulation studies have shown that the best indicator of haplotype age is not frequency but the number of mutational connections in the proposed network; thus a haplotype may still be ancestral even if it is considerably less frequent than a younger haplotype at the tip of the network [Castelloe and Templeton, 1994]. Accordingly, the observed haplotype frequencies can be modified in relation to the position of a haplotype within the network to obtain an "outgroup weight value" whose magnitude correlates with age (Figure 5.6)[Castelloe and Templeton, 1994]. Outgroup weight values of haplotype groups I to VIII were calculated for two networks: one in which group III haplotypes are related to group VII and the other in which they are related to group VI haplotypes (Figure 5.6 A & B).

In both scenarios, the adjusted frequencies of all haplotype groups fit the proposed scheme based on sequence and HVR variation; for example, the outgroup weight of the progenitor group V haplotypes is now showing that these are older than group IV haplotypes, although the original frequency data suggested the reverse. The first scenario thus also suggests that group VI are older than group III haplotypes so that the latter must have been produced from group VII haplotypes (Figure 5.6 A). The second scenario envisages group VII haplotypes at the tip of the network, with an age younger than that of the group III haplotypes (Figure 5.6 B). However, data on sequence allele heterozygosity (Table 5.4), IZHVR heterozygosity (Table 5.6) and, particularly, $\zeta 2$-$\zeta 1$ sequence variability (Table 5.10) all strongly suggest that group VII haplotypes must be older than group III haplotypes, regardless of their relationship to them. On this basis, the second scenario is rejected, so that group III

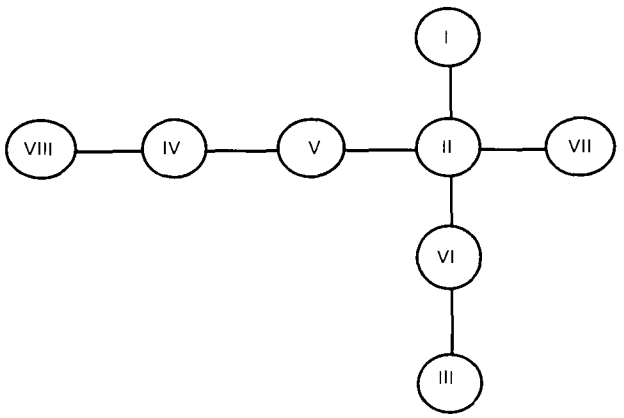
Figure 5.6: Calculation of Outgroup Weights of α-Haplotype Groups

A)




| Tip Haplotype Groups | | Interior Haplotype Groups | | Sum of Weights |
|---|---|---|---|---|
| I | 0.132 | II | 0.649 | 2.1 |
| III | 0.113 | V | 0.466 | |
| VI | 0.0035 | IV | 0.155 | |
| VIII | 0.0025 | VII | 0.579 | |
| | Outgroup Weight | | | |
| I | 0.063 | V | 0.222 | |
| II | 0.309 | VI | 0.002 | |
| III | 0.054 | VII | 0.276 | |
| IV | 0.074 | VIII | 0.002 | |

B)




| Tip Haplotype Groups | | Interior Haplotype Groups | | Sum of Weights |
|---|---|---|---|---|
| VII | 0.0185 | VI | 0.549 | 2.08 |
| I | as above | II | as above | |
| III | " | V | " | |
| VIII | " | IV | " | |
| Outgroup Weight | | | | |
| VI | 0.263 | | | |
| VII | 0.009 | | | |

Rest as above.

Calculation of weighted frequencies according to heuristic approach of Castelloe and Templeton, [1994]: tip haplotypes are assigned a weight half of their frequency in the sample, interior haplotypes a weight that equals their own frequency plus the frequencies of all haplotypes that are one mutational step from them. The weights are summed over all haplotypes, and the weight of a specific haplotype is divided by this sum to generate the outgroup weight.

haplotypes must be considered to be mutationally related to group VII haplotypes, and representing a subset of the progenitor's diversity.

This leaves the question of whether the group VI haplotypes are indeed as young as they seem from the original and corrected frequency data (Table 5.12; Figure 5.6), since they are associated with considerable diversity at the sequence and HVR level (Tables 5.4, 5.6 & 5.9) and therefore would not be expected to appear at the tip of the haplotype network. The group may have, by chance, drifted to low frequency in present human populations, or perhaps has multiple origins resulting in artificially raised levels of variation (if progenitor haplotypes have different sequences). Since group VI haplotypes have probably arisen locally in the Australian Aboriginal population, the latter hypothesis may be true and, as will be seen, other data also support this.

The outgroup weights of the other haplotype groups, however, are likely to be indicative of their relative ages; groups II, V and VII are considerably older than other groups, and groups I and III, although widespread in extant populations, are more recent in origin. The explanation for these young haplotypes being so widespread may be demographic effects in recent human history, such as population expansion (as discussed in Chapter 6).

**The Origins of Full Haplotypes**

The $\alpha$-haplotype network (Figure 5.5) represents the primary mutational relationships between haplotype groups and does not attempt to include the origins of individual haplotypes (members of a haplotype group). From the perspective of specific haplotypes, the network is simplistic since their relationships are likely to be considerably more complicated, given the scrambling effect of 5' and 3'

recombination which is responsible for much haplotypic diversity (Chapter 2). Nevertheless, information about the origins of a few haplotypes is available, and the network itself highlights further aspects of the mechanisms producing haplotypes.

Group V haplotypes are identified as having originated directly from the group II haplotypes (Figure 5.5). From a consideration of RFLP structure, one of the most common group V haplotypes, the Vc (+ - - S PZ + - - +), could be related by one mutational step at the θ *Pst I* site to IIe haplotypes (+ - - S PZ + - - -). This relationship is suggested by the sharing of similar sequence alleles (Table 5.1), similar geographical distribution [Higgs *et al.*, 1986; Martinson, 1991], and from a further sequence polymorphism identified in the Acc region, site Acc: 594 (Table 3.1; Chapter 3). This site usually has the sequence (G), but it was found to be (A) on Southeast Asian IIe, Vc, IIc and VIg haplotypes (Tables 3.1 & 5.14). The same haplotypes from other geographic regions are associated with the more common sequence (G) (Table 3.1; Chapter 3).

Ten additional chromosomes were therefore sequenced at Acc: 594 (Table 5.14). The Southeast Asian (Philippines and Taiwan) and Australian Aborigine IIe haplotypes are associated with (A) at Acc: 594, but the same haplotypes in Africa and the coastal regions of New Guinea are not (Table 5.14). Vc haplotypes in Southeast Asia and Micronesia also carry the (A) allele, but not in Africa, or the highlands of New Guinea (Table 5.14). The same sequence allele is associated with the Southeast Asian and Australian IIe and Vc (sequence allele 10, Table 5.14), but this is different to those associated with the same haplotypes in New Guinea and Africa (Table 5.1).

Table 5.14: Sequence at Site Acc: 594 in IIe and Vc chromosomes

| Region | Haplotype | Acc: 593 |
|---|---|---|
| Philippines | IIc/IIe | A/A |
| Ponape | Vc/VIg | A/A |
| Taiwan | IIe/IIe | A/A |
| Austr Aborig | IIe/IIe | A/A |
| Philippines | IIc/IIe | G/A |
| New Guinea (C) | IIe/IIe | G/G |
| New Guinea (H) | IVa/Vc | G/G |

Note: (C) = coast, (H) = highlands. Ponape is a Micronesian island.

These data suggest that Vc haplotypes in Southeast Asia have originated locally from a IIe haplotype, and that it is the Southeast Asian IIe present in the Micronesian and Australian Aborigine populations, despite evidence from α-haplotype data to suggest that there is little recent Southeast Asian influence on the latter [Roberts-Thomson *et al.*, in press]. Thus, Southeast Asian-type haplotypes (Vc and IIe) differ from those with identical RFLP structure in Melanesia and Africa. The group VI (and IIc) haplotypes are polymorphic at Acc: 594, also suggesting multiple origins for the group in world populations, as considered feasible from the amount of associated variation with the group, despite its position at the tip of the network (Figure 5.5).

The proposed network demands that multiple mutation events at the same nucleotide have occurred. This suggests that some RFLP sites are subject to gene conversion events, since it has already been shown that the α-globin gene complex is not hyper-mutable and that all (-) alleles at any one RFLP site have the same sequence (Chapter 3). For example, the ancestral θ *Pst I* (-) allele mutated to become (+) on the group V and IV haplotypes, but then was lost to form the group VIII. The *Acc I* RFLP site, being ancestrally (+), has been lost on more than one occasion, to

form both the group III and group VI haplotypes. If group VI haplotypes have multiple origins, evidence of multiple conversions involving the *Acc I* RFLP is expected; multiple conversions on group II backgrounds, in certain geographical regions, could produce local-specific group VI haplotypes (as seen in the Australian Aborigine population discussed earlier).

Additional evidence for gene conversion at the *Acc I* site is provided by preliminary sequencing of novel haplotypes. Such haplotypes are found at low frequency in many populations, and are usually new recombinants between the 5' and 3' end of the haplotype (Chapter 2). Occasionally, new 3' end patterns, i.e. haplotype groups, are observed. One such group, PZ - + - - , is found in a number of geographical regions, including Asia and the Pacific [O'Shaughnessy *et al.*, 1990]. Two chromosomes were sequenced at the Sac, Rsa and α Pst regions, in order to determine the associated sequence allele and likely origin of this novel group (Table 5.15).

---

Table 5.15: Sequence Alleles of Novel Haplotype Groups

| Region | Sequence Allele | Haplotype |
|---|---|---|
| Sri Lanka | T, 15, G, C, G, G | + + - M PZ - + - - |
| Guam | T, 15, G, C, G, G | + + - M PZ - + - - |
| Australia | C, 15, G, C, G, G | - - + M Z + + - - |

---

According to RFLP structure, the group PZ - + - -, could be derived from a Ia haplotype by the loss of the *Acc I* (+) site; the associated sequence allele is suggestive of this relationship (Tables 5.1 & 5.15). Another novel haplotype group, Z + + - -, found only in the Australian Aborigine population [Roberts-Thomson *et al.*, in press] has the 3' end of the sequence allele associated with group I haplotypes, combined with a 5' Sac region sequence characteristic of a IIIa (Table 5.15). Given this information and the RFLP structure of this novel haplotype, the group may be the

result of an exchange event between a group I and group III haplotype at the Acc region of the aligned haplotypes. Thus, in this case, a recombination event in the Acc region has resolved as a crossover, rather than conversion. Therefore, although homologous exchange between the 5' and 3' ends of the haplotype may be frequent, events are also likely occur at other regions of the haplotype and may be responsible for some of observed haplotype diversity.

Multiple origins for some group I, as well as group VI, haplotypes also seem likely; all Ia and Ib haplotypes are associated with the same sequence allele, regardless of geographical origin (Table 5.1). The Id, however, is associated with two sequence alleles which are very different to that seen on the Ia/Ib (Table 5.1). Id haplotypes have a 5' end (- + - L) that is found on the IIa; the IIa is associated with sequence alleles similar to those found on the Id, particularly in having the otherwise quite rare (14) allele at Sac: 421 (Table 5.1). The Id therefore may have originated from a IIa which underwent a conversion event, transferring the *Rsa I* (+) site. If so, it is not clear whether the resulting Id haplotype should be classified as a true group I haplotype, or as a variant IIa.

Similar problems are likely to be encountered with other haplotypes, particularly rare ones, and highlight the difficulty of assigning origins to haplotypes when their sequence alleles or HVR sizes turn out to be non-distinguishing. Only young haplotypes retain evidence of their origins because there has not been enough time for recombination to obscure them. For other full haplotypes, such as those in group II, it is impossible on the basis of the data presented in this thesis to determine the precise origins of each haplotype. It is also clear that geographical context is important for the origins of full haplotypes. The consequence of this is that haplotypes such as the Vc, IIe and group VI, are not strictly comparable in different regions of the world, and thus are less suitable for population affinity studies using

α-haplotype data. Much information on the origins of full haplotypes, and the relationships between different populations, may be gained from future studies of geographical sequence variation at this locus. This is also evident from the associated sequence alleles of a number of other haplotypes from different regions (Table 5.1); for example, Southeast Asian IIc haplotypes appear to be different to those in Africa, and the Polynesian and European IVa differs to that in Melanesia and Australia (Table 5.1).

Regardless of these complications, it is important that so much about haplotype relationships can be gleaned from recombining sequences. Clearly there are different problems associated with segmental exchange of sequence polymorphisms and extremely variable HVRs, which can be used as an indicator of haplotype age, and crossover events such as those which swap 5' and 3' ends of the haplotype. The latter serve to obscure the origins of particular haplotypes, except when looking at the youngest haplotypes. However, given the likely future development of statistical methods for dealing with recombinogenic sequences in a phylogenetic context, the success of using such sequences in this study promises that more sophisticated analyses will be possible, thus overcoming one of the problems of using nuclear DNA loci for evolutionary studies. Further, the derived network of α-haplotype relationships reveals the importance of demography, in addition to recombination, for world-wide patterns of variation. Particularly, the high frequency of recently arisen haplotypes, at the tip of the network, has important implications for theories of human evolution.

# Chapter 6

## α-Globin Gene Sequence Variation and Human Evolution

The development of a method to determine α-haplotypes by PCR has enabled the characterisation of considerable sequence and VNTR length variation within the α-globin gene cluster. This has answered many of the questions raised in Chapter 1. For example, since (-) RFLPs always have the same sequence and little extra common polymorphism was found in the cluster, the RFLP haplotype represents well the sequence variation at this locus and, although some geographical structure to this variation was observed, it is valid to use α-haplotypes to investigate human population affinities. It is also apparent that the patterns of variation observed in human populations at this is locus are affected by recombination. Within populations, RFLP haplotype diversity is generated by 5' to 3' end recombination (Chapter 2), generating population differences primarily in sub-haplotype composition, rather than in new 3' ends (haplotype groups). The 3' end of the α-haplotype (PZ/Z, *Acc I*, *Rsa I*, α *Pst I*, and θ *Pst I*) represents a fairly stable linkage group, with relationships between haplotype groups based primarily on point mutational processes (Chapter 5). However, extensive recombination, not evident from the RFLP structure, occurs at localised conversion "hotspots", both at the α genes at the 3' end of the haplotype, and at the ζ genes and within the HVRs at the 5' end of the haplotype.

This localisation of recombination is important because it indicates that the compartmentalisation of recombination rates in the genome, based on cytogenetic observations and chromosome banding (Chapter 1), is a simplification. Thus, the β-globin gene cluster has a recombination hotspot but is contained in a characteristically "cold" isochore; the α-gene cluster does not have a crossover

hotspot despite being contained in a "hot" isochore. The localisation of recombination will likely be of importance at other gene loci, with the effect of allowing the persistence of polymorphism in the face of drift or selection, and also in sequence homogenisation and reduction in diversity, depending on factors such as the balance of conversion to crossover and lengths of conversion tracts (refer to Chapter 3). Explaining patterns of sequence variation, however, is not simple and it is clear from the work described in this thesis that both molecular and population genetic forces operate in a complex and interdependent way to determine locus variation. Demography has been an implicit part of the explanation for α-globin sequence variation; for example, specific demographic information is contained at the Acc: 594 site, indicating that the IIe/Vc haplotypes have multiple and geographically specific origins (Chapter 5). In addition, size variation at the IZHVR contains useful population information (Chapter 4). There are also demographic implications for recent human evolution from the α-haplotype network derived in the previous chapter.

## Human Evolution and Genetics

### Hypotheses of Human Evolution and an African Origin for Modern Humans

Fossil data indicate that the earliest hominids arose in Africa. From there, a migration of *Homo erectus* to Europe and Asia occurred more than 1 mya (and perhaps as early as 1.8 mya [Swisher *et al.*, 1994]). However, the more recent history of human ancestry has been, and continues to be, a source of considerable debate. There are two contrasting extreme hypotheses of human evolution; the first implies that present population differences are very old. One form of this hypothesis, the multiregional hypothesis, maintains that the transition of *H.erectus* to *sapiens* was made several times in different regions, with species identity maintained by gene

flow [Wolpoff *et al.*, 1984]. This is supported by recent dating of archaic *H.sapiens* fossil finds in China [Tiemei *et al.*, 1994].

The second argument suggests that after the *H.erectus* migration, *sapiens* emerged from *erectus* in Africa in one speciation event[1] and underwent a second exodus from there, replacing archaic populations elsewhere. In this rapid replacement or Out-of-Africa scenario, racial differences are not very old [Stringer and Andrews, 1988]. The accumulating data on human genetic variation lend support for an ultimately African origin for hominids since African populations are usually more diverse than other populations and often contain ancestral alleles (see Chapter 1 for references). These data, along with the finding of comparatively low diversity in humans [Li and Sadler, 1991], have been interpreted as being supportive of the Out-of-Africa model. Perhaps the most influential study in this context was of mtDNA variation, which suggested that the African origin of modern humans was recent (~250,000 ya) and occurred in a bottleneck event [Cann *et al.*, 1987]. Despite much criticism, this rapid replacement model has become the orthodoxy of human evolutionary scenarios.

However, HLA and other nuclear data suggest that a significant bottleneck did not occur during human evolution and that the long-term human population size has been $10^4$ to $10^5$ [Ayala *et al.*, 1994; Takahata, 1993; Takahata, 1994]; also the inference of a bottleneck is based upon confusion between a gene genealogy and a population phylogeny (because mtDNA is non-recombinogenic, it is inherited as a single gene) [Jones and Rouhani, 1986]. In addition, high levels of African diversity are not unexpected if African populations have been larger for longer periods of time; genetic drift is a more significant force in smaller populations, as those in Europe and Asia may have been during the Pleistocene (perhaps due to the effects of

[1]Or racial divergence event if *H.sapiens* and *erectus* are considered to represent sub-species rather than different species - this question is not easy to resolve from the fossil data.

glaciations). This could therefore have reduced diversity in non-African populations. Certainly, there are also problems for the alternative, multiregional hypothesis, in that it is not clear whether plausible rates of gene flow could maintain species diversity in isolated populations while still allow regional variation. Also, the model is at variance with the current speciation paradigm [Mayr, 1963].

Genetic variation from a larger number of loci can potentially address some of the issues in this debate. There are two aspects of the α-globin data which are relevant in this context and which have been used in other studies to distinguish between the human evolutionary hypotheses. Firstly, there is the ability to determine the age of variation, and secondly there are the patterns of variation in different populations; in this way, genetic data can be used to consider the specific issues of "when" and "how" human evolution occurred, and through the determination of ancestral states, some idea of "where" (see Excoffier, [1990]).

### α-Globin Molecular Divergence Times

There are two methods for estimating molecular divergence times; the first, the relative branch method, involves the calibration of mutation rate via the fossil record, assuming a constant molecular clock, and time is calculated according to Equation 6.1:

$$T = \theta / 2\mu \qquad \textbf{(6.1)}$$

The second method determines coalescence time (CT); that is, the age of the common genetic ancestor if lineages are traced back through time [Hudson, 1990]. It can be approximated according to Equation 6.2 [Tajima, 1983; Templeton, 1993]:

$$CT = \theta (1+k)/2\mu (1 + n\theta) \quad (6.2)$$

where $\theta$ is the average nucleotide diversity, k is the total number of sequence differences between the two most divergent sequence haplotypes, $\mu$ is the mutation rate and n is the total number of nucleotides studied. The variance is calculated by Equation 6.3:

$$\sigma^2 = \theta^2(1+k)/4\mu^2(1 + n\theta)^2 \quad (6.3)$$

CTs are usually larger than molecular clock derived estimates (hereafter referred to as molecular divergence times) unless the number of nucleotides studied and segregating sites in the sequence data are large [Ruvolo, 1996]. Both estimates based on the sequence data presented in Chapter 3, with the results of comparable estimates made from β-globin sequence variation, are shown in Table 6.1.

Table 6.1: Molecular Divergence Times of α- and β-Globin Gene Cluster Sequence Variation

| Locus | Regions Considered | μ (x10-9) | θ (= L) | T (kyr) | n (bp) | k (max.) | CT (kyr) |
|---|---|---|---|---|---|---|---|
| α | All regions | 3.15 | 0.00103 | 160 | 4986 | 10 | 290 +/- 90 |
| | - θ Pst/ Xba | 2.1 | 0.00129 | 300 | 3149 | 7 | 600 +/- 170 |
| β | + R | 1.4 | 0.0018 | 480 | 2658 | 13 | 1580 +/-420 |
| | - R | 1.4 | 0.0013 | 460 | 2158 | 10 | 1350 +/- 400 |

Refer to Chapter 3 for details of α-globin sequence variation. β-globin times are from Fullerton, [1995] and Fullerton *et al.*, [1994]. +R sequence information includes 500 bp of sites which are known to undergo recombination due to close vicinity to a recombination hotspot; -R calculation excludes the polymorphic sites contained in this 500 bp.

The ranges of dates estimated for the α- and β-globin sequence variation do not coincide with one another (Table 6.1); the dates for α-globin are more similar to estimates based on apolipoprotein alleles (~500,000 yrs, [Xiong *et al.*, 1991]) and the ZFY locus on the Y chromosome (270,000 yrs, but note 95% confidence limits of 0-800,000 yrs [Dorit *et al.*, 1995]), although larger than that of mtDNA (~200,000 years [Ruvolo, 1996; Ruvolo *et al.*, 1993]). Other nuclear loci have yielded older dates than both α- and β-globin: > 30 myrs for HLA genes [Ayala *et al.*, 1994; Ayala and Escalante, 1996; Takahata, 1993] and between 1-3 myrs for the γ and ψη genes and a region 5' to the δ genes (all contained in the β-globin gene cluster) [Bailey *et al.*, 1992; Maeda *et al.*, 1983].

Most of these times are within neutral expectations: a neutral nuclear polymorphism takes ~$4N_e$ generations to become fixed and since coalescence theory envisages the reverse of this process, nuclear autosomal alleles are expected to coalesce ~$4N_e$ generations ago (actually it is $4N_e$ [1-(1/i)] where i = number of sampled alleles). Thus, assuming a generation time of 15-20 years, and an effective population size of $10^4$, the expected coalescence time of a human neutral nuclear polymorphism is ~600,000-800,000 years. This is not inconsistent with the α-globin or apolipoprotein dates or the molecular divergence date of β-globin (Table 6.1) [Fullerton *et al.*, 1994]; it is also approximately four times the age of mtDNA, which is not unexpected given that mtDNA is maternally inherited and haploid (so that its expected neutral coalescence time is ~$2N_f$, where $N_f$ is the effective number of females).

In the case of human HLA alleles, balancing selection causes the maintenance of variation over times greatly exceeding the 600,000-800,000 year range expected for neutral loci [Ayala *et al.*, 1994]; the same explanation may also be applicable to the region 5' to the δ genes, which has a molecular divergence time of 3 myrs [Maeda *et al.*, 1983]. The question remains of what causes the variation in dates obtained with

other nuclear loci; β-globin variation appears to be older than α-globin and other nuclear loci, although still within the boundaries of neutral expectation (Table 6.1). Since all dates are associated with considerable variances, it is impossible to determine whether the differences between loci are significant. However, if the difference between α and β globin dates is real, then a possible explanation is provided by comparing the date obtained including all regions of the α-gene cluster with that obtained excluding the Xba and θ Pst regions. The latter contain *Alu* repeats which undergo gene conversion in humans and chimpanzees, with the result that divergence is elevated whilst human polymorphism is lowered (Chapter 3).

Thus, the calculated CT including the converting Xba/θ Pst regions is considerably lower to that obtained excluding these regions (Table 6.1). However, it was clear in previous chapters (Chapter 3 & 5) that polymorphic sites, other than those contained in the Xba/θ Pst regions, also undergo gene conversion, although they do not have an associated elevation in species divergence. The estimate excluding of the Xba/θ Pst regions is less than half that of the β-globin gene variation (Table 5.1) and it is possible that gene conversion is operating in the α-globin gene cluster to reduce the overall level of human polymorphism relative to the expectation from a relatively high mutation rate (Chapter 3). This is not inconceivable considering that telomeric regions, in which the α-gene cluster is situated, may be prone to conversion rather than crossovers (see Chapter 1 for discussion). However, localised, and perhaps more frequent, conversion such as that which occurs around the α-genes, will operate to offset the influence of large-scale conversion and sequence homogenisation. Such localisation allows for the persistence of polymorphism, such as the Rsa/α Pst: 347 sites, in the face of selection at the α-genes.

Localised recombination may thus overcome a hitchhiking and selective-sweep-effect around the α-genes, by uncoupling neutral sites from selected ones. However,

in regions of the genome experiencing low levels of recombination, hitchhiking (due to advantageous or deleterious mutations) can result in lower within-species diversity [Begun and Aquadro, 1992; Charlesworth *et al.*, 1993; Kaplan *et al.*, 1989], as noted to be important in the *Drosophila* genome [Begun and Aquadro, 1995]. The effect of hitchhiking is to reduce the time to the most recent common ancestor (CT), with the magnitude of the reduction dependent on population size, the strength of selection operating on non-neutral substitutions, when the selected substitution occurred and the amount of crossover [Kaplan *et al.*, 1989].

An overall reduction in diversity across the α-gene cluster, due to selection at the α- or ζ-genes, could perhaps explain lower than expected human polymorphism levels relative to divergence, as compared with β-globin. Certainly, despite the presence of regional conversion hotspots, linkage disequilibrium does occur across the α-gene cluster, evidenced from the existence of haplotypes. The CTs of other loci, particularly of the non-recombining mtDNA and Y chromosome, are also likely to be reduced through hitchhiking, as well as through the $N_e$ size effect discussed earlier, and may explain a number of studies that argue that the mtDNA locus does not behave neutrally [Ballard and Kreitman, 1995; Excoffier, 1990; Jorde *et al.*, 1995; Takahata, 1994].

**Coalescence Times and Human Evolution**

In a human evolutionary context, the CT of a locus is not informative, and does not date the human speciation event. Unless the human species underwent a profound reduction in population size just as it was founded, so that almost all variation inherited from the ancestral population was lost by drift, there is no reason to assume that allelic divergence began when the species originated. It has already been noted that HLA and other data indicate that such a bottleneck event did not

occur, and the molecular divergence dates obtained for α-globin variation are consistent with a long-term population size of $10^4$ to $10^5$ ($N \sim \theta/4\mu$; refer to Table 6.1). However, the estimation of the time of human population divergence does potentially provide information about human evolution, and in particular for distinguishing between the Out-of-Africa and multiregional hypotheses. This estimation cannot be made with the α-globin data because an extensive survey of intra-population sequence variation was not performed. However, from a variety of other loci studied so far, this population divergence time (African vs. non-African populations) is reported to be consistently in the region of ~150,000 years [Nei and Takezaki, 1996].

When a population divides, each daughter population takes some sample of alleles from the original and if divisions were not too long ago, the CT of each population's alleles will be greater than or equal to the population divergence, and thus will represent an upper boundary of when populations split[2]. The CTs of a large number of loci will vary due to different mutational processes and selection as discussed above, but the resultant distribution should provide a consensus date (for discussion, see [Ruvolo, 1996]). Since the consensus is ~ 150,000 yrs, the Out-of-Africa model is favoured at present [Nei and Roychoudhury, 1993; Nei and Takezaki, 1996], a conclusion which is reinforced by the finding that sub-Saharan Africans are genetically more divergent from the rest of the world populations (as discussed earlier). In terms of α-haplotype diversity (Chapter 2), α-globin sequence polymorphism (Table 3.1, Chapter 3), IZHVR variation (Table 4.4, Chapter 4 & Figure 5.1, Chapter 5) and ψζ1 IVS1.HVR size variation (Table 4.8, Chapter 4), Africans are more diverse at the α-globin gene locus. In addition, the Sac region rare

[2]If, however, alleles are lost after divergence, such as might occur in a small population, the CT will underestimate population divergence.

ancestral allele, T-$A_{14}$ (Table 3.5, Chapter 3 & Table 5.1, Chapter 5) is present only in African samples.

These results are consistent with the Out-of-Africa theory because the most divergent and first-established population is likely to stay in place of origin and new populations would be formed when they move out of the original place. However, although the α-globin data do seem to support the hypotheses, it is not clear whether a simple model of rapid replacement is more likely or whether some gene flow with archaic populations outside of Africa has occurred during migration of modern humans from Africa.

**α-Haplotype-Related Sequence Variation and Recent Population Expansions**

It was striking that the α-haplotype groups, I and III were found to be almost monomorphic in sequence and associated with low levels of HVR variation (Chapters 3 and 4). This sequence monomorphism was observed in all populations, and on all subtypes of the group III and most of those of group I haplotypes. The implication of this is that these haplotype groups have arisen recently, and they are therefore depicted at the tips of the haplotype network (Figure 5.5; Chapter 5). However, they are present at high frequency in most world populations (Chapter 5); for tip haplotypes to have this distribution suggests recent population expansion.

The β-globin sequence haplotype tree (R Harding, personal communication; [Fullerton, 1995]) has a number of features in common with the α-haplotype network derived in Chapter 5. There are four β-globin lineages (A, B, C and D) which consist of sequence alleles originally defined by RFLPs. The B lineage contains the ancestral allele (as determined by interspecific comparisons) which, despite its age, is geographically confined to Africa; other regionally-specific haplotypes are

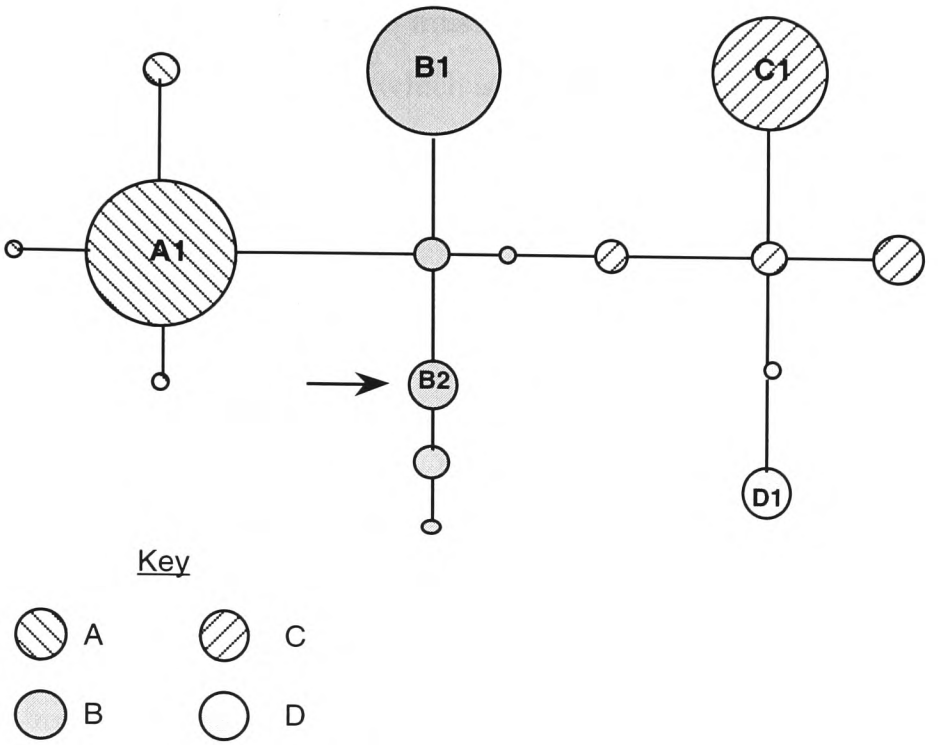
also internally situated on the tree. Likewise, group VII α-haplotypes are old, yet confined geographically to Africa (Chapter 5). The A1, B1, C1 and D1 alleles arise near the tips of the β-globin gene tree (Figure 6.1), but are common and widespread in modern human populations. In this sense, they are similar to the group I and III haplotypes. Thus, alleles A1, B1, C1 and D1 trace the recent global expansion indicated by the α-globin data.

Population size expansion has been proposed from the mismatch analysis of mtDNA variation in human populations. This analysis uses the shape of the frequency distribution of pairwise differences in a sample to determine past demographic influences [Rogers and Harpending, 1992]. Observations of uni-modal distributions, corresponding to a "star-like" mtDNA phylogeny, are suggestive of a recent (~50,000) population expansion [DiRienzo and Wilson, 1991; Harpending *et al.*, 1993]; following a population expansion, polymorphic variation is more likely to be retained due an increase in $N_e$ and a single mtDNA lineage may diverge rapidly. This interpretation has been criticised, however, since mutation rate heterogeneities and genetic hitchhiking can give rise to similar distributions [Aris-Brosou and Excoffier, 1996; Begun and Aquadro, 1995]. A prediction of the hitchhiking effect is that the frequency distribution of polymorphisms will show a skew towards rare polymorphisms, since most will have arisen since the last selective sweep and be recent (and also resulting in a "star-like" genealogy). Excess rare alleles are not evident in the α-globin data (or β-globin, S M Fullerton, personal communication), indicating that any recent *size* expansion is not evident in these nuclear loci, although the gene trees demonstrate the effect of a recent *geographical* expansion.

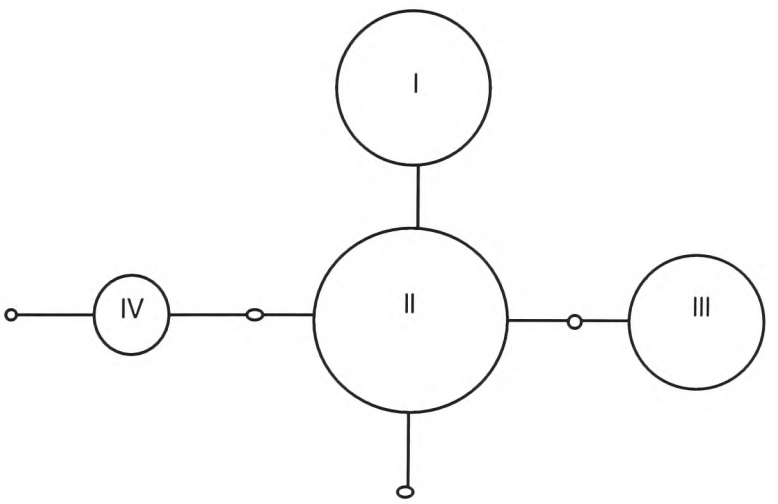
The source of this geographical expansion is not clear; C and D alleles are rare in Africa, although the former are common in Asia and the latter are present in Europe. Similarly, the Ia and IIIa haplotypes are common in Europe and Asia, respectively,

Figure 6.1: α and β-Globin Gene Trees

A)




B)




A) Gene tree showing mutational relationships of β-globin sequences (R M Harding, personal communication), which fall into four groups, or lineages: A to D. Size of circle indicates relative differences in world frequency (not to scale). Arrow indicates ancestral allele. B) Scaled, according to frequency, α-globin gene network (refer to Chapter 5). As in the β-globin tree, tip sequence haplotypes are very frequent in extant human populations, and some internal sequences are much less frequent. This pattern is consistent with recent population expansion.

but the IIIa is almost unknown in Africa and therefore could, perhaps like C and D alleles, have been introduced into that continent from elsewhere. The source population for a recent expansion must have contained the Ia and IIIa along with other haplotypes, such as the IIa, which is also not found in Africa, but is common in many other populations [Higgs *et al.*, 1986]. It appears that any recent expansion must have occurred from a population with an α-haplotype repertoire other than those found in present-day African populations, although it is always possible that new surveys will uncover previously undetected variation (including IIIa and IIa haplotypes) in African populations.

## Future Studies

Specific questions about human evolution and molecular mechanisms (such as the relative importance of conversion and crossover) cannot be determined from the data presented in this thesis, since the sequence survey was not designed to address such issues. However, the development of the PCR haplotype is the first step in the progression towards a more informative approach to the evolutionary and molecular genetics of this locus. Generally, studies of nuclear DNA have lagged behind those of mtDNA for technical reasons (due to diploidy) and because recombination makes data interpretation difficult. However, this study has illustrated the importance of looking at recombinogenic sequences, and also the possibility of doing so with success. Development of the issues raised in this thesis will require, nevertheless, some technical refinements.

To study recombination a more direct approach is desirable, such as the investigation of events in sperm [Li *et al.*, 1988], which is presently being undertaken in the β-globin gene cluster (Julie Schneider, this laboratory). Specific gene conversion events in the α-globin gene complex might be of particular interest for

further study, by direct or indirect analysis. For example, the conversion event of $\psi\zeta 1$ to $\zeta 1$, has been reported to end within the microsatellite repeat contained within the second intron of the $\psi\zeta 1$ gene [Hill *et al.*, 1985]. Primers have been designed to amplify this region (refer to Appendix 5), and conversion events could therefore be characterised in greater detail. This would also be useful for confirmation of the suggested relationship between the $\zeta 1$ converted gene present on the group III and VII haplotypes.

Further phylogenetic studies using this locus would benefit from the determination of allelic sequences. Allele-specific [Wu *et al.*, 1989] or the amplification refractory mutation system (ARMS) PCR [Newton *et al.*, 1989] use a mismatched 3' residue within the primer sequence (and often an additional internal mismatch) to allow amplification of only one allele. This approach has already been successfully applied to β-globin [Fullerton *et al.*, 1994]. The method would be applicable to highly variable regions of the α-gene cluster, where resolution of the phase of polymorphic sites is not possible. In this study, it is conceivable that the choice to sequence regions already known to contain RFLPs and using samples of known haplotype, has resulted in an underestimation of sequence diversity, with obvious important implications for the discussion above of sequence ages[3] (although the value of L determined from sequence data is not very different to θ determined from RFLP data; Chapter 3). The estimation of sequence diversity from a different region of the α-globin gene cluster would help to address this issue, although requiring an approach similar to that employed in the β-complex: sequencing of a larger contiguous region, and utilisation of an ARMS-approach [Fullerton *et al.*, 1994]. Such a strategy would allow within and between population surveys to be carried out,

[3]The samples were from many different populations and it is known that inter-population diversity is lower than intra-population diversity in humans [Lewontin, 1972]. This is countered, however, by the fact that the haplotypes sequenced represent a sample that is more variable than any extant population, possibly increasing observed sequence diversity.

thereby enabling population differences to be characterised and dated in order to contribute to the debate on alternative theories of human evolution.

Achieving more successful α-haplotype assignments within populations, particularly highly variable African populations, and therefore better resolution of population affinities, also relies on an allele-specific approach. In this case, amplification of single DNA molecules, followed by appropriate enzyme digestion [Ruano *et al.*, 1990], would enable haplotypes to be assigned unambiguously. However, to type sites long distances apart, PCR multiplexing might also be required. As with all of these technical applications, successful use with the α-globin gene cluster, which is particularly difficult to amplify by PCR (Chapter 2), relies to some extent on future improvements in PCR technology. Nevertheless, the recent ability to amplify very long pieces of genomic DNA [Cheng *et al.*, 1994] is encouraging, since many of the appropriate modifications and the improved understanding of the amplification reaction may be important for amplifying shorter, but GC-rich regions.

A useful and presently accessible future direction for the use of this locus would be to investigate regions of the complex which are not undergoing recombination. This would allow more standard approaches to phylogeny construction to be considered, and additional polymorphic and non-recombining sites are likely to help in refining the haplotype relationships suggested here (Chapter 5) and also in determining how novel haplotype groups arise. Possible regions to consider as containing such sites are the α-pseudogenes, such as ψα1 and ψα2, which are not undergoing sequence correction, and also the 5' and 3' regions outside of the presently defined α-gene cluster. Primers are already available to extend investigation around the θ gene at the 3' end of the cluster (Appendix 5). Such studies of regions flanking the cluster may indicate to what extent polymorphisms remain in linkage disequilibrium with

the α-haplotype as investigation is moved 5' and 3'; this will help address the question of whether the genome is characterised by regions of linkage disequilibrium (such as the α-globin gene complex), separated by recombination hotspots (such as that contained in the β-gene complex), and therefore provide an understanding about the localisation of recombination in the genome.

This study set out to characterise the sequence variation underlying α-globin haplotypes, and has demonstrated the extremely variable nature of the α-globin gene locus. This variability appears to be most promising for answering many questions of human evolution and molecular genetics, facilitated by the study of a number of different types of polymorphism including point mutations and length variation. The profound possibilities inherent in the comparative study of allelic diversity are also clear, and future investigations of both the α and β gene clusters will enhance our knowledge of the properties of loci contained in different functional compartments of the genome. What perhaps is becoming most clear from this and other studies, is that one human genome sequence, the goal of modern human molecular genetics, will represent only the starting point for a full appreciation of how highly complex and variable a "genome" truly is. It is this variable component with which future genetics, in all fields, will need to be concerned.

# Appendix 1: Standard Materials and Methods

# Appendix 1

## Materials and Methods

The procedures outlined in this Appendix represent standard protocols [Sambrook *et al.*, 1989], or protocols that have been modified only slightly. Methods or modifications of methods which were developed as part of this study are detailed in the text (Chapters 2, 3 and 4).

### A1. 1 DNA Samples Used in this Study

All samples were available as aqueous solutions, most of which had been stored at -20°C.

### A1.2 Restriction Enzyme and Southern Blotting Analysis

This section details the standard method to determine α-globin haplotypes by Southern blotting and use of a number of α-globin cluster specific hybridisation probes [Higgs *et al.*, 1986].

#### A1.2.1 Restriction Enzyme Digestion, Gels and Southern Blotting

~ 10μg DNA was incubated with restriction enzyme (typically 20U) in the appropriate buffer as supplied by the manufacturer (Boehringer Mannheim) at 37°C for 4-6 hours. Enzyme digests were separated on 0.8-1.5% agarose gels (25 x 20 cm) depending on the size of the fragments of interest, and run at ~2V $cm^{-1}$ in horizontal gel tanks (BRL, Ltd.) in 1x AGB, with 500ng of commercially available bacteriophage λ DNA digested with *Hind III* (Pharmacia) as a marker. Gels were stained with ethiduim bromide (0.5mg/ml), visualised on a short-wave UV transilluminator, and then denatured in 0.5M NaOH solution for 30 mins, followed

by 3-4 hours in gel neutralising buffer (GNB). Subsequently, transfer onto nylon blotting membrane (Hybond-N, Amersham) was achieved via a standard Southern blot stack with 6x SSC. Transfer was allowed to proceed overnight (~ 15 hours).

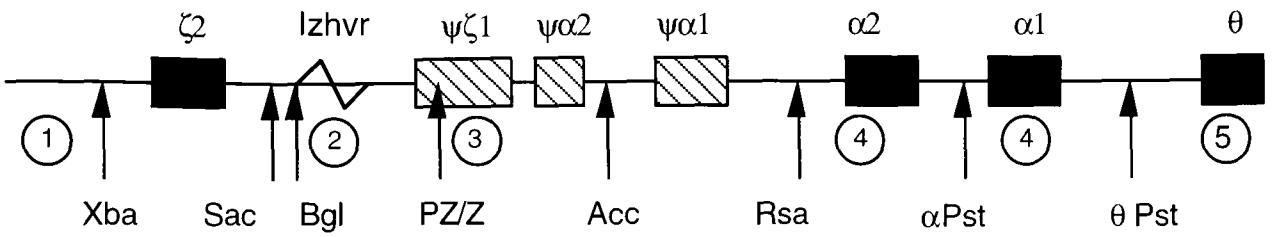
### A1.2.2 Preparation of Gels, Probe Labelling and Hybridisation

Following transfer, membranes were rinsed briefly in deionised water and the DNA immobilised on the membrane by ultraviolet irradiation (in a stratalinker; Stratagene, Ltd.). The membranes were stored dry or hybridised immediately. Prehybridisation and hybridisation were performed using a formamide buffer containing a blocking agent to suppress background binding of the probe. Membranes were placed in 3.5ml of formamide buffer in a sealed plastic hybridisation bag (up to 3 membranes per bag) and left at 42°C for at least 6 hours. Hybridisation was likewise performed at 42°C, using a fresh aliquot of formamide buffer containing a radioactively labelled probe (10ng $ml^{-1}$ formamide buffer).

Probes used in this study are shown in Figure A1.1. They were labelled according to the random-priming extension method of Feinberg and Vogelstein, [1983], which involves the synthesis, at a number of sites primed by random sequence hexanucleotides, of a complementary strand by the Klenow fragment of *E.coli* DNA polymerase I in the presence of labelled deoxynucleotides. The reaction was performed at 37°C with $\alpha^{32}$P-CTP, according to manufacturer's instructions (Multiprime Labelling System, Amersham). Unincorporated nucleotides were removed using Sephadex G50 (Sigma) columns which had been equilibrated with 3x SSC. Hybridisation was allowed to proceed for 12-16 hours.

Following hybridisation, washing procedures ensured removal of excess probe and weak unspecific binding of probe to sequences other than the target sequences. 2

Figure A1.1: Probes Used to Detect RFLPs in the α-Globin Gene Complex




Six probes (indicated by circled numbers) were used to determine the α-haplotype [Higgs *et al.*, 1986]:
1) a 600 bp *Bam*H1/*Eco*RI fragment from the plasmid pJW5, a subclone of the cosmid cSG21 containing ~ 9 kb 5' to the ζ2 gene [Goodbourn, 1983] (L1),
2) a 1.1 kb *Alu I* fragment from the plasmid pSG21, containing the IZHVR [Goodbourn *et al.*, 1983] (IZHVR),
3) a 1.8 kb *Sac I* fragment from the plasmid pBRζ, containing the ψζ1 gene [Proudfoot *et al.*, 1982] (ζ),
4) a 1.5 kb *Pst I* fragment from the plasmid pDH7, containing the α1 gene [Higgs *et al.*, 1981] (α),
5) a 800 bp *Bam*HI fragment from the plasmid pAM18θ, containing the θ gene [Higgs *et al.*, 1986] (θ).

The following digests (A1.2.1) are necessary to type all 9 RFLP sites:

| Digests | Probe | Sites Typed |
|---|---|---|
| *Bam*HI/*Xba I* | L1 | *Xba I* |
| *Bgl I* | IZHVR | *Bgl I* |
| *Sac I* | ζ, IZHVR | PZ/Z, IZHVR, *Sac I* |
| *Acc I* | α | *Acc I* |
| *Rsa I* | α | *Rsa I* |
| *Pst I* | α, θ | α *Pst I*, θ *Pst I* |

washes of 15 mins were performed at room temperature in ~500ml 2X SSC, followed by 2 washes of 30 mins each at 65°C in 0.1x SSC/0.1% SDS in a shaking water bath. Membranes were dried completely before exposure for 2 to 7 days at -70°C to high speed (Kodak XAR-50) X-Ray sensitive autoradiography film in the presence of intensifying screens (DuPont Cronex).

## A1.3 DNA Enzymatic Amplification

Refer to Chapter 2 for details of PCR α-haplotyping protocol.

### A1.3.1 Non-Radioactive and Radioactive 5' Labelling of Primers using T4 Polynucleotide Kinase

Bacteriophage T4 polynucleotide kinase catalyses the transfer of the γ-phosphate of ATP to the 5' hydroxyl terminus of DNA. Non-radioactive and radioactive labellings were performed depending on the required application (see A1.7.2 and Chapter 3 for PCR sequencing). 5000pmoles of forward (5') or reverse (3') primer were incubated with 10U of polynucleotide kinase in a 20μl reaction with kinase buffer and either 1mM ATP or 2μl of $\gamma^{32}$P ATP (75μCi $\mu l^{-1}$) at 37°C for 30 mins. The enzyme was heat-killed at 65°C for 2 mins, the reaction mixture phenol/chloroform extracted and the primers ethanol precipitated (see A.5.1) on dry ice with an equal amount of either 5' or 3' unlabelled opposite primer in the presence of 10μg of a carrier molecule, such as tRNA. Primers were then resuspended in sterile distilled water and were ready for use.

## A1.4 Visualisation of Amplification Products

### A1.4.1 Agarose Gels

To check amplification, 5-10% of a PCR reaction was added to 5μl of electrophoresis dye (A1.10) and run on standard agarose gels, usually at a 1.0-1.5% concentration,

at 10V $cm^{-1}$. 500ng of commercially available θX174 *Hae III* (Pharmacia) or λ *Hind III* were used as markers. Gels were stained in 0.5mg/ml of ethidium bromide and viewed on a short-wave ultra-violet transilluminator.

When attempting to achieve higher level of resolution of amplified products, in particular those containing HVRs, other agaroses and conditions were used (refer to Chapter 4). IZHVR alleles were analysed on 1.5-1.8% agarose gels run at ~1.5V $cm^{-1}$ against a variety of markers: 500ng of commercially available 100 bp marker (Pharmacia) and λ Hind III (Pharmacia). In addition, λ DNA digested with *Cla I* was used since this provides bands in the 1.7-1.9 kb range, which is useful for determining the sizes of M alleles. Typically 100μg of methyl-adenine-free λ DNA (*Cla I* is a methyl-sensitive enzyme) was digested with *Cla I* in buffer as supplied by the manufacturer (New England Biolabs) at 37°C for ~3-4 hours, ethanol precipitated and resuspended in sterile distilled water.

ψζ1, ζ1 and ζ2 alleles were analysed on either 2% Metaphor (FMC, BioProducts) or 2.5% NuSieve (FMC, BioProducts) horizontal gels run in 1x AGB at 5V $cm^{-1}$. Markers used included a 100 bp marker (Pharmacia). These gels provided optimal resolution for these alleles which were ≤ 1 kb in size.

### A1.4.2 Hybrid Agarose-Acrylamide Gels

As outlined in the text, in order to visualise differences between IZHVR amplified alleles at the highest level of resolution, it was necessary to run hybrid agarose-acrylamide gels [Arteiler and Porteous, 1992; Peacock and Dingman, 1968]. The appropriate amount of 30% Protogel (National Diagnostics) was added to 5ml 10x TBE, and 40μl TEMED (N,N,N',N'-Tetramethylethylenediamine, Sigma) and the mixture left in a 42°C water bath. A 1% agarose gel solution in water (to make up a

total volume of gel mixture of 50ml) was boiled, cooled to ~40°C and then added to the acrylamide mixture. 250µl fresh 10% ammonium persulphate (APS, Sigma) was added and the gel poured immediately between glass plates (20 x 20 cm, with the top plate treated with Surfacil siliconising fluid (Pierce, USA); refer to A1.10) separated by 1 mm spacers. The gel was allowed to set for ~ 1 hour and then was run at 250V. The gel was stained directly with ethidium bromide and visualised on a short-wave UV transilluminator.

## A1.5 Purification of Amplified Products

### A1.5.1 Ethanol Precipitation

1/10th of the volume of amplification reaction of sodium acetate and 2 volumes of 100% ethanol were added to the amplified product, mixed and left in a dry ice-ethanol bath for ~30 mins. The mixture was spun at high speed (13,000g) for 10-15 mins and supernatant removed. The pellet was rinsed with 70% ethanol, spun again, then dried and resuspended in an appropriate amount of sterile distilled water.

### A1.5.2 "Gene-cleaning"

This method utilises a silica matrix ("glassmilk") to bind DNA and to remove proteins and PCR primers in the amplification reaction mixture (GeneClean II, BIO 101, Inc.). The method was suitable for most amplification reactions used since all resulted in products of greater than 500 bp. This method of purification was used for preparing the samples for PCR sequencing (A1.8.1). 5µl of glassmilk was used per 5µg of PCR product and the procedure carried out according to manufacturer's instructions.

### A1.5.3 Extraction of Amplified Alleles from Agarose Gels: Gene-Cleaning, Qiaex, and β-Agarase

HVR alleles of different sizes in heterozygous individuals could be isolated on agarose and then extracted (Chapter 4). DNA bands were visualised on a long wave UV transilluminator to reduce damage to the DNA, and excised from the gel using sterile razor blades.

Gene-cleaning procedures could be used to extract the DNA from the agarose slice (standard agarose cast and run in 1x AGB) by 5-10 mins incubation at 55°C, prior to addition of the glassmilk, as directed by the manufacturer's instructions (GeneClean II, BIO 101, Inc.). The Qiaex Kit (Qiagen) utilises a silica matrix to bind DNA in a similar way to the gene-cleaning procedure. 10-15µl of glass beads were incubated with the agarose slice (also standard agarose set in 1x AGB) according to the manufacturer's instructions.

The technique producing the greatest yield involved the digestion of the agarose slice by the enzyme β-*Agarase I* (New England Biolabs). Since this method requires that the solution be liquid at the incubation temperature of 40°C, amplified products were separated on a 1.5% low melting point agarose gel (SeaPlaque GTG, FMC BioProducts). Gels were made and run in 1x AGB. Gel slices were incubated at 65°C for 10-15 mins with ~1/4 volume of TE buffer. 5U of β-*Agarase I* were added per 200µl of volume with the manufacturer's' supplied buffer, and the mixture left at 40°C for 4-6 hours. Following digestion, phenol/chloroform extraction and ethanol precipitation were carried out and the DNA pellet resuspended in an appropriate amount of sterile distilled water.

## A1.6 Restriction Enzyme Analysis of Amplified Products

### A.6.1 α-Haplotyping - Restriction Enzyme Digestion of Amplification Products

The procedure for performing restriction enzyme digests for the purposes of α-haplotyping is described in Chapter 2.

### A.6.1 Internal Mapping of IZHVR alleles using Restriction Enzyme Partial Digests [Smith and Birnstiel, 1976]

Amplification reactions were performed as usual with either forward (5') or reverse (3') primer labelled with $\gamma$-$^{32}$P[ATP] (A1.3). Products from 2-4 reactions were combined (100-200µl) and ethanol-precipitated to remove any substances inhibitory to enzyme digestion and resuspended in 40µl of sterile distilled water.

20µl of precipitated PCR product was incubated with 2U of either *Bsa J1* or *Bst N1* (New England Biolabs), 10x buffer as supplied by the manufacturer and 3 µg BSA (when using *Bst N1*) in a 30 µl reaction at 60°C in a dry heating block (Techne Dri-Block). 6µl aliquots of the digestion mixture were removed at time intervals of 0, 2.5, 5, 10 and 15 mins, and added to an equal volume of Sequenase stop dye (Sequenase Kit, USB). 1µl of each timed aliquot was combined (total 5µl) for each digest of one sample and loaded onto a 3% denaturing polyacrylamide gel (A1.9), often with a sequencing reaction run alongside as a marker. These gels were fixed and dried as normal sequencing gels and exposed to high speed autoradiography film (Kodak XAR-50) at room temperature for ~ 5 days.

## A1.7 Sample Preparation for Sequencing

Sequencing methods require some preparation of samples. Amplified products were gene-cleaned (A1.5.2) for PCR sequencing, but other purification methods were used and single-stranded DNA was produced in two ways.

### A1.7.1 CLB6 Columns to Remove Excess PCR Primers (refer to Chapter 3)

1ml Sepharose CLB6 (Sigma) columns were made and equilibrated with 100µl sterile distilled water. Phenol/chloroform extracted products were placed in the columns and spun at 1300g for 3 mins. The collected mixture was then ethanol precipitated and resuspended ready for further preparation or direct use in sequencing reactions.

### A1.7.2 λ-Exonuclease Digestion to Produce Single-Stranded DNA [Higuchi and Ochman, 1989]

Following an amplification using one kinased primer (T4 polynucleotide kinase and non-radioactive ATP; A1.3.1), the products were ethanol-precipitated or passed through CLB6 columns and then ethanol precipitated (A1.5.1). The resultant DNA was resuspended in ~40µl of sd$H_2O$ and incubated with 10U of λ-exonuclease and manufacturer's supplied exonuclease buffer at 37°C for 2 hours. A phenol/chloroform extraction and ethanol precipitation step followed and the product was resuspended as appropriate for the following sequencing reaction.

### A1.7.3 Use of Dynabeads to Produce Single-Stranded DNA [Thein and Hinton, 1991]

"Dynabeads" (Dynal) are streptavidin coated magnetic beads which bind biotin. Amplifications were performed with one 5' biotinylated PCR primer (Oswel) and

one unmodified primer to produce products which could then be bound to the magnetic beads. 50µl of Dynabeads per 100µl amplification reaction was washed in an equal volume of 1x Bind and Wash (B+W) solution and resuspended in twice the volume of 2x B+W (A1.10). All washing was carried out in a magnetic particle concentrator (MPC, Dynal). The mineral oil was removed from the amplification reactions and the suspended Dynabeads were added. Binding was allowed to continue for 15 mins at room temperature, with occasional gentle mixing.

The supernatant was removed in the MPC and the beads with attached DNA were washed with 1x B+W. Freshly made 0.1M NaOH solution was incubated with the beads to melt the DNA duplex for 10 mins. Following removal of the NaOH supernatant, the beads were washed three times in 0.1M NaOH, 1x B+W and TE buffer respectively, as recommended by manufacturer. Finally, the beads were resuspended in an appropriate amount of distilled water for the sequencing reaction. Note that the discarded NaOH supernatant contains the non-biotinylated strand; purification of the strand and use in sequencing reactions did not, however, result in high quality sequencing information.

## A1.8 Direct Sequencing of Amplified Products

Details of these chain-terminating methods [Sanger *et al.*, 1977] can be found in Chapter 3.

### A1.8.1 Cycle Sequencing using Vent (exo-) Polymerase

(CircumVent Thermal Cycle Dideoxy DNA Sequencing Kit) (New England Biolabs)

Amplification reactions were gene-cleaned prior to the use of this procedure, and generally double-stranded DNA was used. ~200ng of template was added to a 14µl

mixture containing 1.5µl of the supplied 10x buffer, 1µl 30x Triton-X-100 and 2.5pmol of sequencing primer. 2µl of [α-$^{35}$S]dATP (20 µCi) and 2U of Vent (exo-) DNA polymerase (refer to Chapter 2) were added to this solution, gently mixed and 3.2 µl was taken and added to each of four reaction tubes containing 3µl of supplied deoxy/dideoxynucleotide mixture. The reactions were overlain with mineral oil (Sigma) and placed in a thermal cycler with the conditions determined in the text (Chapter 3). 4µl of supplied stop dye was added beneath the mineral oil and the reactions stored at -20°C.

### A1.8.2 Sequencing using "Sequenase" Version 2.0 (USB)

Sequencing utilised single-stranded template that was produced by either the λ-exonuclease method or Dynabeads. Reactions were carried out with not less than 1µg of DNA (usually 1 amplification reaction). The manufacturer's recommended protocol was followed with the modifications as detailed in Chapter 3.

## A1.9 Visualisation of Sequencing Products: Denaturing Polyacrylamide Gels

Denaturing gels were made for visualising sequencing reactions (5%) and also for IZHVR internal maps (3%) (A1.4.2). 40ml gel mixtures were made using 4 ml of either 40% Bis-Acryl (Accugel, National Diagnostics) or high performance "Long Ranger"[1] (AT Biochem; supplied 50% concentration), 4.8ml 10X TBE, and 16.8g ultrapure Urea (GIBCO BRL, Life Technologies). Polymerisation was catalysed using 250µl of fresh 10% ammonium persulphate (Aldrich Chemical Co., Ltd.) and 25µl TEMED (Sigma). Gels were poured between two glass plates (20 x 40 cm; spacing 0.4 mm), one of which had been treated with Surfacil siliconising fluid

[1]Long Ranger gels provided greater sequence information (~30%) from one run compared to standard polyacrylamide.

(Pierce, USA) and one treated with a mixture of 15µl Bind-Silane (γ-methacryloxypropyltrimethoxysilane (Sigma)), 150µl 10% acetic acid and 5ml of 100% ethanol. Gels were allowed to set for ~1 hour before being pre-run for ~20 mins at 30 watts, maximum current and voltage. Samples were heated at ~80°C for 2 mins, and 1.5-3.0µl loaded to each well formed by the sharkstooth comb (usually 2.9 mm point-to-point spacing, 48 wells; BRL, Life Technologies Inc.). Gels were run for ~1-3 hours at < 50°C, and following fixing in 20% ethanol/10% acetic acid, were baked onto the glass plate in a 104°C oven. One cooled, gels were exposed to either high speed (Kodak XAR-50) or single emulsion maximum resolution (Biomax, Kodak) autoradiography film at room temperature for 12-48 hours.

## A1. 10 Buffers And Reagents

| | |
|---|---|
| Formamide Hybridisation Buffer (50 ml) | 50% Formamide<br>3x SSC<br>1x Denhardt's Solution<br>2% SDS<br>10µg Salmon Sperm DNA |
| Denhardt's Solution (100x) | 2% (w/v) BSA<br>2% (w/v) Ficoll (Pharmacia)<br>2% (w/v) Polyvinylpyrrolidone (PVP) |
| Gel Neutralising Buffer (GNB) | 1.5M NaCl<br>0.5M Tris-HCl pH 7.2<br>0.001M EDTA |
| Standard Saline Citrate, SSC (20x) | 3M NaCl<br>0.3M $Na_3$citrate |
| Bind and Wash (2x B+W) | 10mM Tris-Cl (pH 7.5)<br>1mM EDTA (pH 8.0)<br>2mM NaCl |
| TE Buffer | 10mM Tris-Cl<br>1mM EDTA (pH 8.0) |

| | |
|---|---|
| Agarose Gel buffer (AGB) | 0.04M Tris-HCl pH 8.0<br>0.02M Sodium acetate<br>0.2mM EDTA |
| TBE Electrophoresis Buffer | 89mM Tris base<br>89mM Boric acid<br>2.5mM EDTA (pH 8.0) |
| TBE loading dye (Non-denaturing gels) | 0.02% Bromophenol Blue<br>0.25% Xylene cyanole FF<br>40% Sucrose |
| PNK-labelling Buffer | 0.05M Tris-Cl pH 7.6<br>0.01M $MgCl_2$<br>5mM Dithiothreitol<br>0.1mM Spermidine<br>0.1mM EDTA |
| AGB loading dye | 15% Ficoll 400<br>0.05% Bromophenol blue<br>0.05% xylene cyanole |

## Appendix 2: Kenyan α-Haplotypes

Raw haplotype and α-genotype data for 72 Kenyan Luo, as described in Chapter 2. Only unambiguous haplotypes are shown. N/A = information not available due to PCR failure. * = presence of a rare (-) *Pst I* allele at a second polymorphic site contained in the α Pst amplified region (see Chapter 2).

| Sample | *Xba I* | *Sac I* | *Bgl I* | IZHVR | Subtype | PZ/Z | *Acc I* | *Rsa I* | *aPst I* | *tPst I* | Genotype | Haplotype |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1O | +/+ | +/- | -/- | MM | M39/M42 | PZ/PZ | +/+ | +/- | -/- | -/- | αα/αα | |
| 1Q | +/+ | +/+ | -/- | MM | M40/M40 | PZ/PZ | +/+ | +/- | -/- | -/- | αα/α−3.7 | IIc/Ia |
| 2F | -/- | +/+ | -/- | SS | S12/S12 | Z/Z | +/+ | -/- | -/- | -/- | N/A | VIIi/VIIi |
| 2G | -/- | -/- | -/- | MM | M43/46 | Z/Z | +/- | -/- | -/- | -/- | αα/αα | VIIe/IIIf |
| 2H | +/+ | -/- | -/- | MM | M43/M43 | PZ/PZ | +/+ | -/- | -/- | -/- | αα/α−3.7 | IIg/IIg |
| 2I | +/+ | +/- | -/- | MM | M41/M47 | PZ/PZ | +/+ | +/- | -/- | -/- | αα/αα | |
| 2J | +/+ | +/- | -/- | MM | M40/M41 | PZ/PZ | +/+ | +/- | +/- | +/- | αα/α−3.7 | |
| 2L | +/- | +/- | -/- | SS | S12/S12 | PZ/Z | +/+ | +/- | -/- | +/- | αα/αα | |
| 2M | -/- | +/- | -/- | MM | M44/M43 | PZ/PZ | +/- | -/- | -/- | -/- | αα/αα | |
| 2N | +/+ | +/- | -/- | SS | S12/S9 | PZ/Z | +/- | -/- | -/- | +/- | αα/α−3.7 | |
| 2O | +/- | +/- | -/- | SS | S12/S12 | Z/Z | +/- | -/- | -/- | +/- | αα/α−3.7 | |
| 2P | -/- | -/- | -/- | MM | M42/M42 | Z/Z | -/- | -/- | -/- | -/- | αα/α−3.7 | IIIf/IIIf |
| 2Q | +/- | -/- | -/- | MS | M42/S12 | PZ/Z | -/- | -/- | +/- | +/- | αα/αα | |
| 2R | +/- | +/- | -/- | SS | S12/S12 | PZ/PZ | +/- | -/- | -/- | +/- | N/A | |
| 2S | +/- | -/- | -/- | MS | M43/S12 | PZ/PZ | -/- | -/- | -/- | -/- | N/A | |
| 2T | +/- | -/- | -/- | MM | M42/M43 | Z/Z | +/+ | -/- | -/- | -/- | αα/αα | VIIe/VIIc |
| 2U | +/- | -/- | -/- | SS | S12/S12 | PZ/Z | +/- | -/- | +/- | +/- | αα/αα | |
| 2V | +/- | -/- | -/- | MS | M42/S12 | PZ/PZ | +/- | -/- | -/- | +/- | αα/α−3.7 | |
| 2W | +/- | +/- | -/- | MS | M42/S12 | PZ/PZ | +/+ | -/- | -/- | +/- | αα/α−3.7 | |
| 2X | +/- | +/+ | -/- | MM | M40/M40 | PZ/PZ | +/- | +/- | +/- | +/- | αα/αα | |
| 3A | +/+ | +/- | -/- | MS | M40/S12 | PZ/PZ | +/- | +/- | +/- | +/- | αα/αα | |
| 3B | +/+ | +/+ | -/- | MM | M40/M40 | PZ/PZ | +/+ | +/- | +/- | +/- | αα/αα | |
| 3C | -/- | -/- | -/- | MM | M43/M43 | Z/Z | -/- | -/- | -/- | -/- | αα/αα | IIIf/IIIf |
| 3D | -/- | +/- | -/- | MS | M43/S12 | Z/Z | +/- | -/- | -/- | -/- | αα/αα | |
| 3F | +/- | +/- | -/- | MS | M45/S12 | PZ/PZ | +/+ | -/- | -/- | -/- | αα/αα | |
| 3G | +/- | +/- | -/- | MM | M42/M42 | Z/Z | -/- | -/- | -/- | -/- | αα/αα | |
| 3H | +/- | +/- | +M/- | MS | M45/S12 | Z/Z | +/- | -/- | -/- | -/- | α−3.7/α−3.7 | |
| 3I | +/- | +/- | -/- | MM | M42/M40 | PZ/PZ | -/- | -/- | -/- | +/- | αα/αα | |
| 3K | -/- | +/+ | -/- | MM | M43/M43 | PZ/PZ | +/+ | -/- | -/- | -/- | αα/αα | IIj/IIj |
| 3L | +/- | +/- | -/- | MM | M42/M41 | PZ/PZ | +/- | -/- | -/- * | -/- | αα/αα | |
| 3M | -/- | +/- | -/- | MM | M43/M43 | Z/Z | -/- | -/- | -/- | -/- | αα/αα | IIIf/IIIe |
| 3N | +/+ | +/+ | -/- | SS | S9/S12 | Z/Z | +/+ | -/- | -/- | -/- | αα/αα | VIIg/VIIg |
| 3O | +/- | +/- | -/- | MS | M43/S12 | PZ/Z | +/+ | -/- | -/- | -/- | αα/αα | |
| 3P | -/- | +/- | -/- | MM | M44/M42 | PZ/Z | +/+ | -/- | -/- | -/- | αα/αα | |
| 3Q | +/- | +/+ | -/- | MS | M40/S12 | PZ/Z | +/+ | -/- | -/- | -/- | αα/αα | |
| 3R | +/- | +/- | -/- | MS | M42/S12 | PZ/Z | +/+ | -/- | -/- | -/- | N/A | |
| 3S | -/- | +/+ | -/- | MM | M42/M42 | PZ/Z | +/+ | -/- | -/- | -/- | αα/αα | IIj/VIIa |
| 3T | +/- | +/+ | -/- | MS | M40/M41 | PZ/Z | +/+ | +/- | -/- | -/- | αα/αα | |
| 3U | +/- | +/- | -/- | MM | M41/M41 | PZ/Z | +/- | +/- | -/- | -/- | αα/αα | |
| 3V | -/- | +/- | -/- | MS | M42/S12 | Z/Z | +/- | -/- | -/- | -/- | αα/αα | |
| 3X | +/+ | +/- | -/- | MS | M40/S12 | PZ/PZ | +/+ | -/- | +/+ | +/+ | αα/αα | |
| 4A | +/- | +/- | +/- | MS | M41/S12 | Z/Z | +/- | -/- | -/- | -/- | αα/αα | |
| 4B | +/+ | +/- | -/- | MS | M40/S12 | PZ/PZ | +/+ | -/- | -/- | +/- | αα/α−3.7 | |
| 4C | +/+ | +/- | -/- | MM | M42/M40 | PZ/PZ | +/+ | -/- | -/- | -/- | αα/αα | IIg/IIc |
| 4D | +/- | +/+ | -/- | MS | M40/S12 | PZ/Z | +/+ | +/- | -/- | -/- | αα/α−3.7 | |
| 4E | +/- | +/- | -/- | MM | M42/M44 | PZ/Z | +/+ | +/- | -/- | -/- | α−3.7/α−3.7 | |
| 4F | +/+ | +/- | -/- | MM | M40/M47 | PZ/PZ | +/+ | -/- | -/- | +/- | αα/αα | |
| 4G | +/+ | -/- | -/- | SS | S12/S12 | PZ/PZ | +/- | -/- | +/- | +/- | αα/α−3.7 | |
| 4H | +/- | -/- | -/- | MM | M42/M42 | PZ/Z | +/- | +/- | -/- | -/- | α−3.7/α−3.7 | |
| 4I | +/+ | -/- | -/- | SS | S12/S12 | PZ/Z | +/+ | -/- | +/- | -/- | αα/αα | |
| 4J | +/- | +/+ | -/- | MS | M42/S12 | PZ/Z | +/+ | -/- | -/- | +/- | αα/αα | |
| 4K | +/- | +/- | +M/- | MS | M41/S12 | Z/Z | +/- | -/- | -/- | -/- | αα/α−3.7 | |
| 4L | +/+ | -/- | -/- | MS | M42/S12 | PZ/Z | +/+ | -/- | +/- | -/- | αα/αα | |
| 4M | -/- | +/- | -/- | MM | M42/M42 | PZ/Z | +/- | +/- | -/- | -/- | αα/αα | |
| 4N | -/- | -/- | -/- | MM | M42/M42 | Z/Z | -/- | -/- | -/- | -/- | αα/αα | IIIf/IIIf |
| 4O | -/- | +/- | -/- | MM | M42/M42 | Z/Z | +/- | -/- | -/- | -/- | αα/αα | |
| 4R | +/- | +/- | -/- | MM | M42/M42 | PZ/Z | +/- | -/- | -/- | -/- | αα/αα | |
| 4S | +/- | +/+ | -/- | MM | M40/M42 | PZ/PZ | +/+ | -/- | -/- * | -/- | α−3.7/α−3.7 | IIc/IIj |
| 4T | -/- | -/- | -/- | MM | M42/M42 | Z/Z | -/- | -/- | -/- | -/- | αα/αα | IIIf/IIIf |
| 4U | +/- | -/- | -/- | MM | M42/M42 | Z/Z | +/- | -/- | -/- | -/- | αα/α−3.7 | |
| 4V | +/+ | +/- | -/- | MS | M40/S12 | PZ/PZ | +/+ | +/- | +/- | +/- | αα/αα | |
| 4W | +/- | +/+ | -/- | MS | M40/S12 | PZ/Z | +/+ | +/- | -/- | -/- | αα/αα | |
| 4X | +/+ | -/- | -/- | MS | M40/S12 | PZ/PZ | +/- | -/- | +/- | +/- | αα/α−3.7 | |
| 4Y | +/+ | -/- | +M/- | MS | M41/S12 | PZ/Z | +/- | -/- | +/- | -/- | αα/αα | |
| 5A | +/+ | +/+ | -/- | MS | M40/S9 | PZ/PZ | +/- | +/- | -/- | -/- | αα/αα | |
| 5B | +/+ | -/- | -/- | MM | M42/46 | PZ/PZ | +/+ | -/- | -/- | -/- | αα/α−3.7 | IIg/IIg |
| 5C | +/- | +/- | -/- | SS | S12/S12 | PZ/PZ | +/+ | -/- | -/- | -/- | αα/αα | |
| 5D | +/- | +/- | -/- | MS | M40/S12 | PZ/PZ | +/- | -/- | -/- | -/- | α−3.7/α−3.7 | |
| 5E | +/- | -/- | -/- | MS | M42/S12 | PZ/PZ | -/- | -/- | +/- | +/- | N/A | |
| 5F | +/+ | -/- | N/A | MS | M42/S12 | PZ/Z | +/+ | -/- | -/- | +/- | αα/α−3.7 | |
| 5I | +/- | -/- | N/A | MS | M42/S12 | PZ/Z | +/- | -/- | -/- | +/- | αα/α−3.7 | |
| 5J | +/- | -/- | -/- | MM | M42/M42 | PZ/PZ | +/- | -/- | +/+ | -/- | αα/αα | |

# Appendix 3: Sequencing Primers Used In This Study

| Region | Primer name | Primer sequence (5'-3') | Strand sequenced | Position in PCR fragment |
|---|---|---|---|---|
| Xba | XSF1 | GCCAGCCAAACAGACATGGACTC | 5'-3' | 11 |
| | XSF2 | GCCTGACCACCTGACCAACGTG | 5'-3' | 191 |
| | XSF3 | GTTAGCCGGGTGGTAGTGGC | 5'-3' | 414 |
| | XSR1 | CCATTACACCAAGTGAAGGACG | 3'-5' | 947 |
| Sac | SSF1 | GAAGGAAGGATGACTGCTGG | 5'-3' | 9 |
| | SSF2 | TGCCTGTAATCCCAGCTAC | 5'-3' | 290 |
| | SSR1 | AAGCCGATGTGGTAGAATG | 3'-5' | 555 |
| IZHVR | IZSF1 | GGACACAGATGGAGGCTAAT | 5'-3' | 4 |
| | IZSR1 | CATGAGGCAGAGGGATGCGT | 3'-5' | 1610 |
| PZ/Z | PZSF1 | GTGGGCAGAAGGGGAGACAG | 5'-3' | 3 |
| | PZSF2 | ATCTCCACGCAGGCCGACAC | 5'-3' | 291 |
| | PZSF(T) | CCCTGGGGGCTGAGCACTACCAAG | 5'-3' | 167 |
| | PZSR1 | GAGGGAAGGCTATCTGTCTCC | 3'-5' | 959 |
| Acc | ASF1 | AGGCTGGTCTCAAACTCCTGACCTC | 5'-3' | 32 |
| | ASF2 | CAGGACAATGCCAGCGGGAAG | 5'-3' | 305 |
| | ASF3 | GGCTAACACCAAGGCGGTCTG | 5'-3' | 559 |
| | ASR1 | CATTGTTTCTTGATCTGGGC | 3'-5' | 821 |
| | ASR1.2 | CATAACTCAGGGCCAAGTGAGAG | 3'-5' | 796 |
| Rsa | RSF1 | CCCCTTCCTTCCTCACCCCACAT | 5'-3' | 30 |
| | RSF2 | ACAGTGAACCACGACCTCTAGGCCA | 5'-3' | 195 |
| | RSF3 | ACTATCGCCAGAGGGAAAGGG | 5'-3' | 408 |
| | RSF4 | GTGTGTTCCCCGATCCCG | 5'-3' | 571 |
| | RSR1 | CGTTGGGGTCGGGCGCTGTC | 3'-5' | 734 |
| | RSR2 | ATCCTGGGGCGGAGAGGAAT | 3'-5' | 536 |
| | RSR3 | GCGTGGGGTGGTGGGTACTG | 3'-5' | 288 |
| α Pst | αPSF1 | GTTGAATGCTCCAGCCGGTTC | 5'-3' | 2 |
| | αPSF2 | GCGTCCTGGTGTTTATTCCT | 5'-3' | 367 |
| | αPSF3 | ACTCGCTCGCCCGTGTGTTC | 5'-3' | 569 |
| | αPSR2 | CCAGGGTTTATGCTTGGGGCG | 3'-5' | 897 |
| θ Pst | P99SF1 | GGTGTCTTCGTTATTTCATC | 5'-3' | 3 |
| | P99SF2 | CGAGACCAGCCTAGCCAACATAGTG | 5'-3' | 276 |
| | P99SF3 | GGAAACCCCATCTCTACTAA | 5'-3' | 596 |
| | P99SR1 | GCTTCATTTCTTTTTGTGGC | 3'-5' | 882 |

Position of primer refers to position in PCR product relative to the 5' terminus of the forward PCR primer.

## Appendix 4: Sequences of α-Globin Regions Investigated

Where appropriate, human and chimp sequences are aligned, with * indicating differences between the sequences and - indicating the introduction of a gap to facilitate the alignment. Refer to Chapter 3 for a diagrammatic representation of the features of each sequence region, and also for details of the positions of polymorphic sites. Where alignments have introduced gaps and therefore altered the position of the polymorphic sites in the human sequence presented here, the new position is indicated.

## A4.1: Xba Region

Sequences of human (bottom) and chimpanzee (top) aligned.

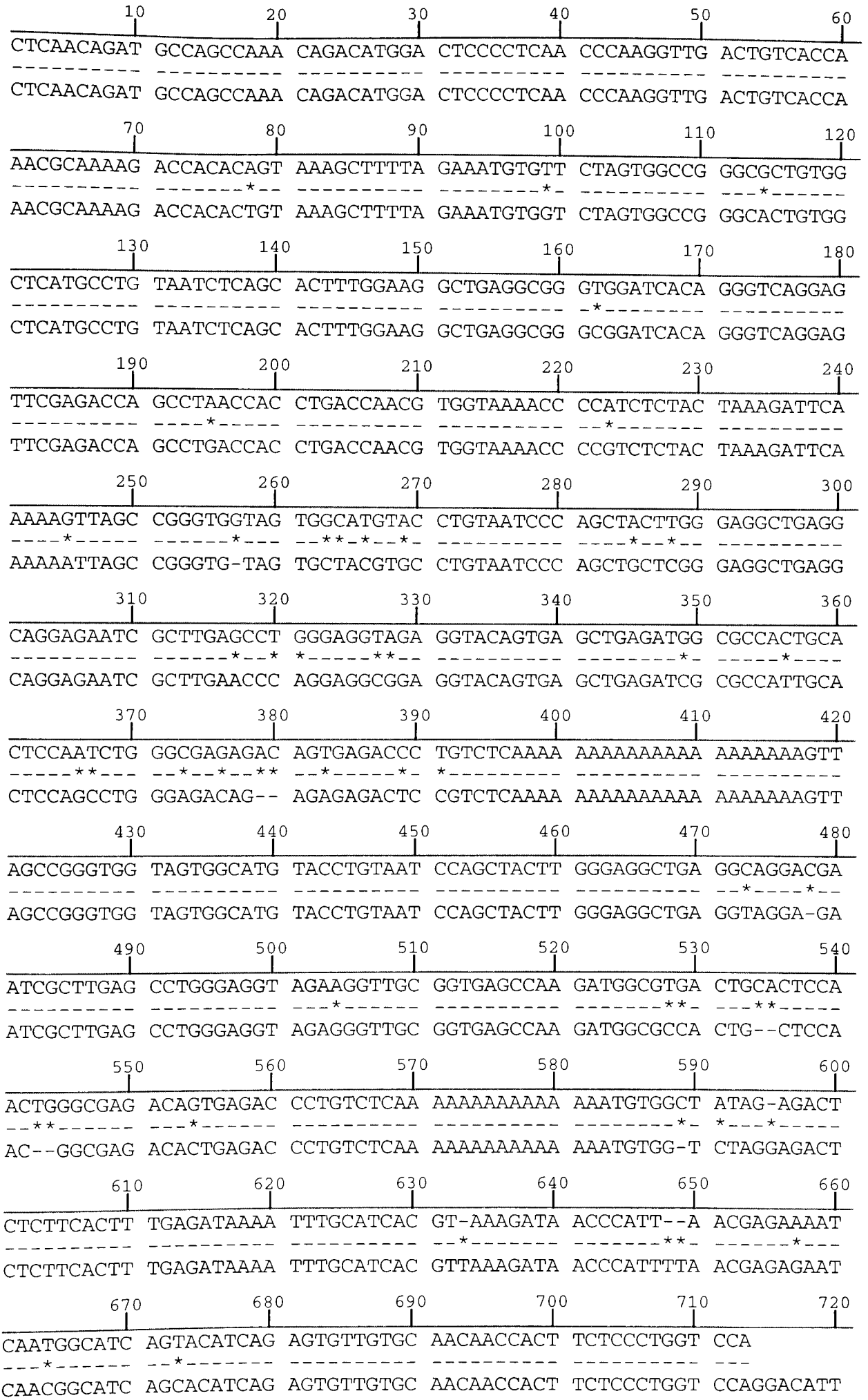
Features

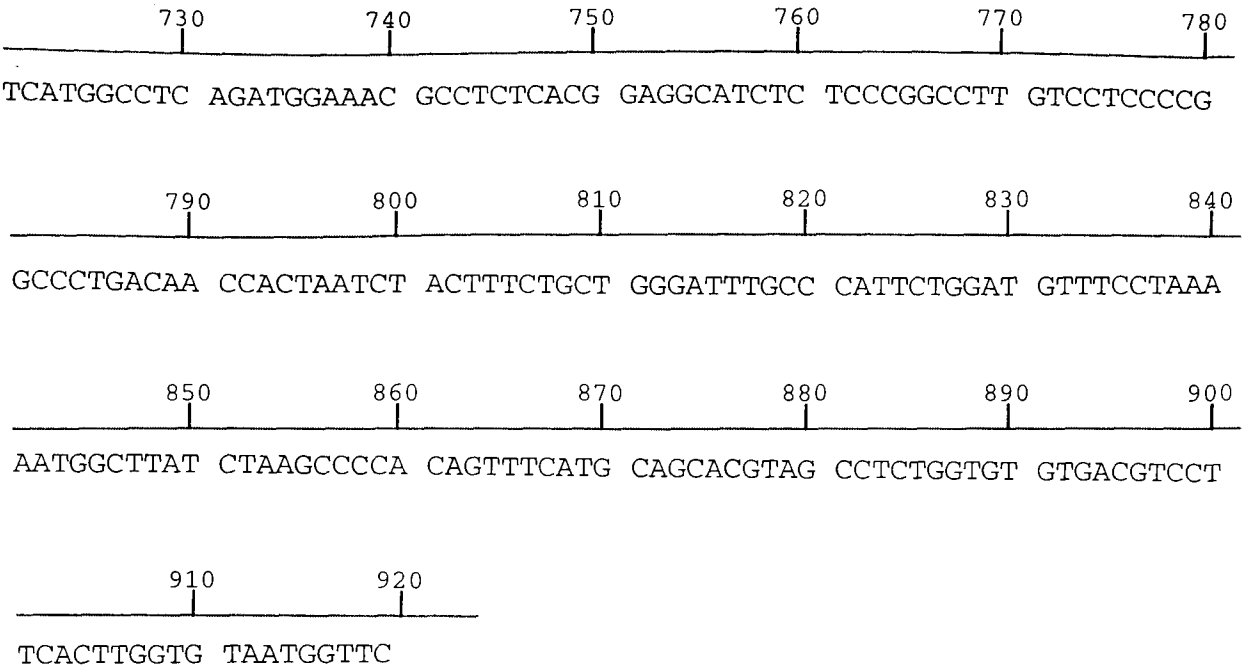
*Alu* repeat 1 from 106-575

Truncated *Alu* repeat 2 from 414-575

**Polymorphisms**

| Position in Human Sequence | Position in Aligned Sequence |
|---|---|
| 532 | 536 |
| 586 | 593 |




730
740
750
760
770
780
TCATGGCCTC AGATGGAAAC GCCTCTCACG GAGGCATCTC TCCCGGCCTT GTCCTCCCCG
790
800
810
820
830
840
GCCCTGACAA CCACTAATCT ACTTTCTGCT GGGATTTGCC CATTCTGGAT GTTTCCTAAA
850
860
870
880
890
900
AATGGCTTAT CTAAGCCCCA CAGTTTCATG CAGCACGTAG CCTCTGGTGT GTGACGTCCT
910
920
TCACTTGGTG TAATGGTTC

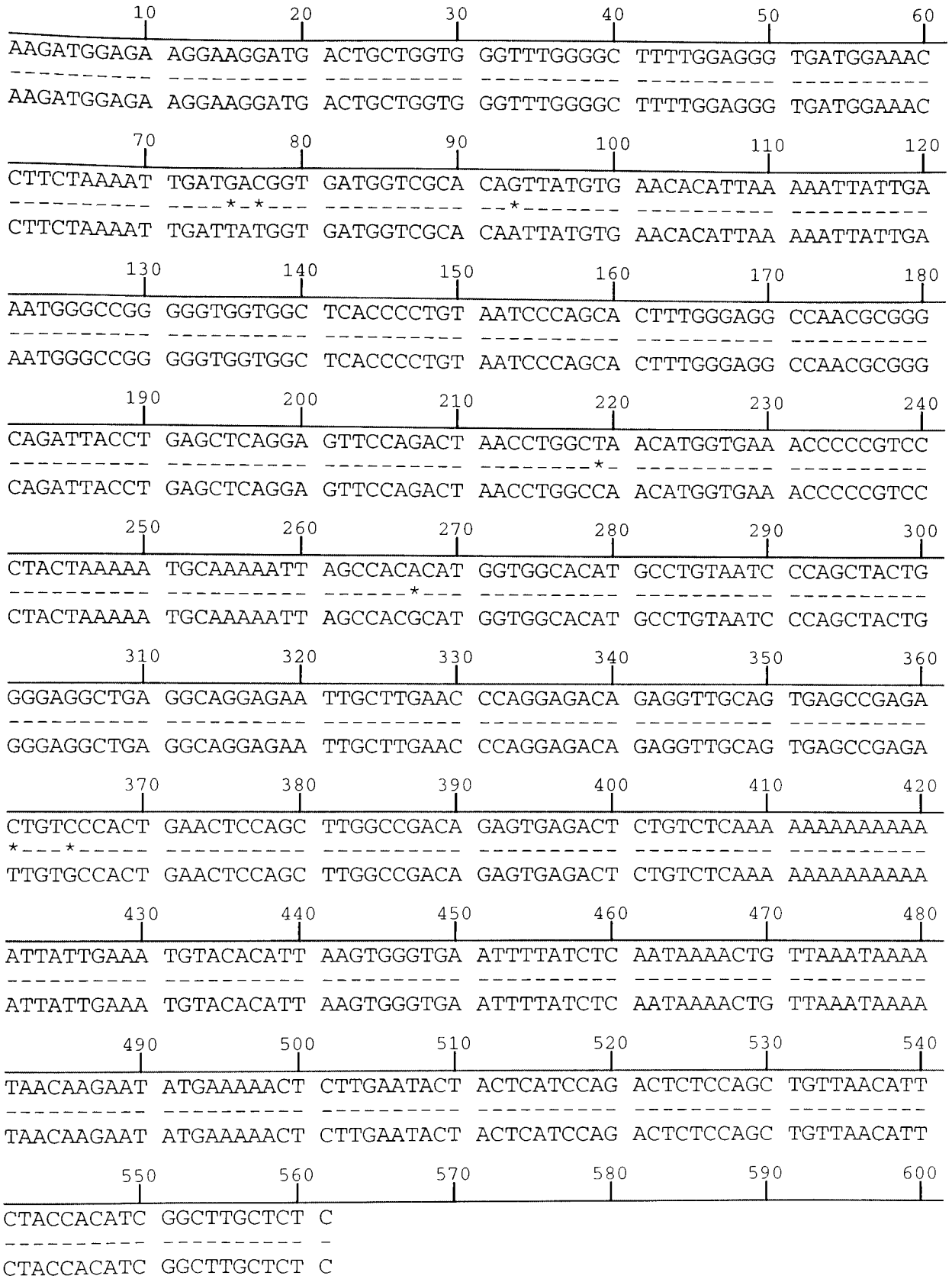
## A4.2: Sac Region

Sequences of human (bottom) and chimpanzee (top) aligned

Features

χ sequence from 24-31

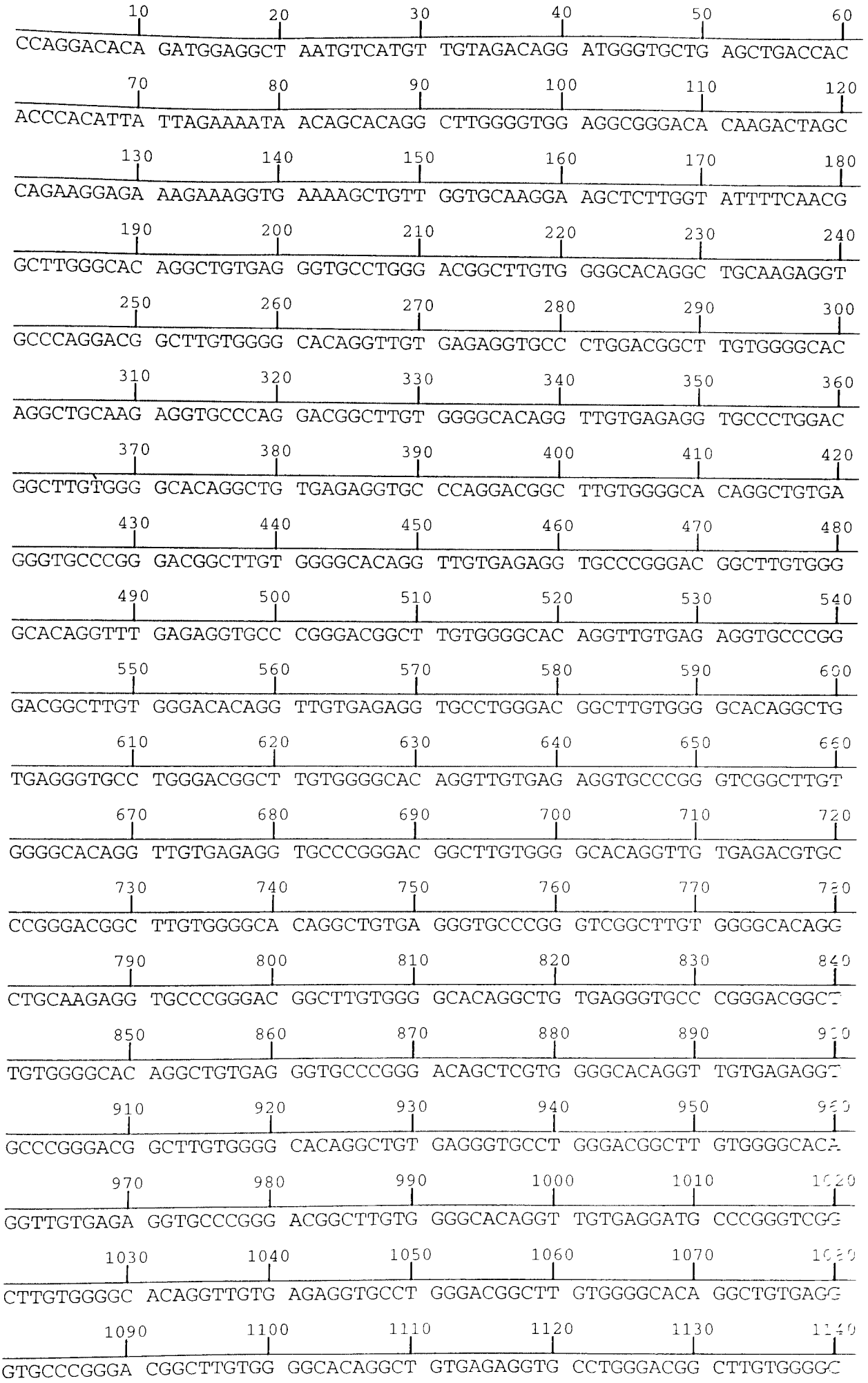
Sac *Alu* repeat from 125-421

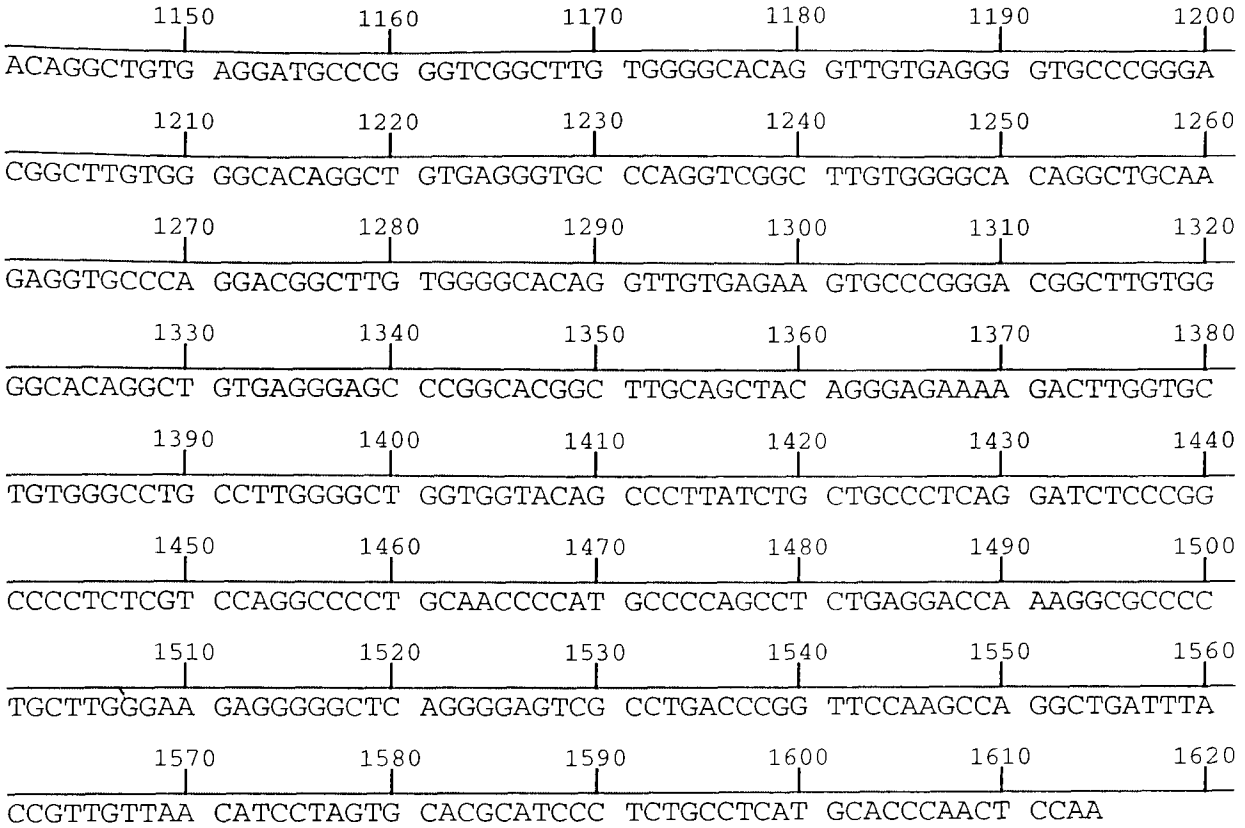


## A4.3: IZHVR Region

Human sequence containing IZHVR.

Repeat array at 184-1354 (repeats are 36-37 bp long; Chapter 4)
This array is of a M class allele (after Goodbourn *et al.*, [1983]).


10 20 30 40 50 60
CCAGGACACA GATGGAGGCT AATGTCATGT TGTAGACAGG ATGGGTGCTG AGCTGACCAC
70 80 90 100 110 120
ACCCACATTA TTAGAAAATA ACAGCACAGG CTTGGGGTGG AGGCGGGACA CAAGACTAGC
130 140 150 160 170 180
CAGAAGGAGA AAGAAAGGTG AAAAGCTGTT GGTGCAAGGA AGCTCTTGGT ATTTTCAACG
190 200 210 220 230 240
GCTTGGGCAC AGGCTGTGAG GGTGCCTGGG ACGGCTTGTG GGGCACAGGC TGCAAGAGGT
250 260 270 280 290 300
GCCCAGGACG GCTTGTGGGG CACAGGTTGT GAGAGGTGCC CTGGACGGCT TGTGGGGCAC
310 320 330 340 350 360
AGGCTGCAAG AGGTGCCCAG GACGGCTTGT GGGGCACAGG TTGTGAGAGG TGCCCTGGAC
370 380 390 400 410 420
GGCTTGTGGG GCACAGGCTG TGAGAGGTGC CCAGGACGGC TTGTGGGGCA CAGGCTGTGA
430 440 450 460 470 480
GGGTGCCCGG GACGGCTTGT GGGGCACAGG TTGTGAGAGG TGCCCGGGAC GGCTTGTGGG
490 500 510 520 530 540
GCACAGGTTT GAGAGGTGCC CGGGACGGCT TGTGGGGCAC AGGTTGTGAG AGGTGCCCGG
550 560 570 580 590 600
GACGGCTTGT GGGACACAGG TTGTGAGAGG TGCCTGGGAC GGCTTGTGGG GCACAGGCTG
610 620 630 640 650 660
TGAGGGTGCC TGGGACGGCT TGTGGGGCAC AGGTTGTGAG AGGTGCCCGG GTCGGCTTGT
670 680 690 700 710 720
GGGGCACAGG TTGTGAGAGG TGCCCGGGAC GGCTTGTGGG GCACAGGTTG TGAGACGTGC
730 740 750 760 770 780
CCGGGACGGC TTGTGGGGCA CAGGCTGTGA GGGTGCCCGG GTCGGCTTGT GGGGCACAGG
790 800 810 820 830 840
CTGCAAGAGG TGCCCGGGAC GGCTTGTGGG GCACAGGCTG TGAGGGTGCC CGGGACGGCT
850 860 870 880 890 900
TGTGGGGCAC AGGCTGTGAG GGTGCCCGGG ACAGCTCGTG GGGCACAGGT TGTGAGAGGT
910 920 930 940 950 960
GCCCGGGACG GCTTGTGGGG CACAGGCTGT GAGGGTGCCT GGGACGGCTT GTGGGGCACA
970 980 990 1000 1010 1020
GGTTGTGAGA GGTGCCCGGG ACGGCTTGTG GGGCACAGGT TGTGAGGATG CCCGGGTCGG
1030 1040 1050 1060 1070 1080
CTTGTGGGGC ACAGGTTGTG AGAGGTGCCT GGGACGGCTT GTGGGGCACA GGCTGTGAGG
1090 1100 1110 1120 1130 1140
GTGCCCGGGA CGGCTTGTGG GGCACAGGCT GTGAGAGGTG CCTGGGACGG CTTGTGGGGC


1150 1160 1170 1180 1190 1200
ACAGGCTGTG AGGATGCCCG GTCGGCTTG TGGGGCACAG GTTGTGAGGG GTGCCCGGGA
1210 1220 1230 1240 1250 1260
CGGCTTGTGG GGCACAGGCT GTGAGGGTGC CCAGGTCGGC TTGTGGGGCA CAGGCTGCAA
1270 1280 1290 1300 1310 1320
GAGGTGCCCA GGACGGCTTG TGGGGCACAG GTTGTGAGAA GTGCCCGGGA CGGCTTGTGG
1330 1340 1350 1360 1370 1380
GGCACAGGCT GTGAGGGAGC CCGGCACGGC TTGCAGCTAC AGGGAGAAAA GACTTGGTGC
1390 1400 1410 1420 1430 1440
TGTGGGCCTG CCTTGGGGCT GGTGGTACAG CCCTTATCTG CTGCCCTCAG GATCTCCCGG
1450 1460 1470 1480 1490 1500
CCCCTCTCGT CCAGGCCCCT GCAACCCCAT GCCCCAGCCT CTGAGGACCA AAGGCGCCCC
1510 1520 1530 1540 1550 1560
TGCTTGGGAA GAGGGGGCTC AGGGGAGTCG CCTGACCCGG TTCCAAGCCA GGCTGATTTA
1570 1580 1590 1600 1610 1620
CCGTTGTTAA CATCCTAGTG CACGCATCCC TCTGCCTCAT GCACCCAACT CCAA

## A4.4: ψζ1 Sequence

Human sequence obtained with PZ.R7 reverse primer. Refer to Proudfoot *et al.*, [1982] for full sequence of ψζ1 gene, and for ψζ1/ζ2 sequence differences (also Chapter 4).

Chimpanzee sequence can be found in Willard *et al.*, [1986]; some information in Chapter 4.

Features

Sequence from 0-41 bp is divergent between ψζ1 and ζ2 genes
ψζ1 exon 1 at 186-336 (initiation codon at 241)
IVS 1 begins at 337
IVS1.HVR from 395-939 (this sequence is of a $PZ^{39}$ allele)
Alternative reverse primer, PZ.R2, at 964-989.

10 20 30 40 50 60

GGGTGGGCAG AAGGGGAGAC AGAAGCGATA GGTTCCTCAG CCCCCAGTCC CACCTGAGGG

70 80 90 100 110 120

CCCCTTTGTC ACTGGATCTG ATAAGAAACA CCACCCCTGC AGCCCCCTCC CCTCACCTGA

130 140 150 160 170 180

CCAATGGCCA CAGCCTGGCT GGGCCCAGCT CCCTGTATAT AAGGGGACCC TGGGGGCTGA

190 200 210 220 230 240

GCACTACCAA GGCCAGTCCT GAGCAGGCCC AACTCCAGTG CAGCTGCCCA CCCTGCCGCC

250 260 270 280 290 300

ATGTCTCTGA CCAAGACTTA GGGGACCATC ATTGTGTCCA TGTGGGCCAA GATCTCCACG

310 320 330 340 350 360

CAGGCCGACA CCATCGGCAC CGAGACTCTG GAGAGGTGAG TGTCAGATGG GACTGCCAGA

370 380 390 400 410 420

GGGACTGGGT GGGAGGCCAG GTATGTGAGT GGGGACAGTG GGGAGCGGGC AGTGGGGAGG

430 440 450 460 470 480

GGACCGTGGG GAGGGGACAG TGAGTAGGGG ACAGTGGGGA GGACAGTGGA GAGGGGACAG

490 500 510 520 530 540

TGAGGAGGGG ACCATGGGAA GGGGACCGTG GAGTGGGGAC AGTGAGGAGG GGACCATAGG

550 560 570 580 590 600

GAGGGGACAG TGGGGAGGGG ACAGTGAGGA GGGGACCGTG GGGAGGGGAC AGTGAGGAGG

610 620 630 640 650 660

GGACCGTGGG GAGGAGACAG TGAGGAGGGG ACCGTAGGGA GGGGACAGTG AGGAGGGGAC

670 680 690 700 710 720

CGTGGGGAGG GGACAGTGAG GAGGGGACCG TGGGGAGGGG ACAGTGAGGA GGGGACCGTG

730 740 750 760 770 780

GGAAGGAGAC AGTGAGGAGG GGACCTTGGG GAGGGGACAG TGAGGAGGGG ACCATGGGGA

790 800 810 820 830 840

GGGGACAGTG AGGAGGGGAC AATGGAGAGG GGACAGTGAG GAGGGGACTG TGGGGAGAGG

850 860 870 880 890 900

ACAGTGAGGA GGGGACCATG GGGAGGGCAC AGTGGGGAGG GGAGAGTGAG GAAGGGACAG

910 920 930 940 950 960

TGAGGAGGGG ACTGTGGGGA GGGGACAGTG GAGACAGATA GCCTTCCCTC TCAGTGAGGA

970 980 990 1000 1010 1020

GGGCAGGGTA AGGAGGGAAC GATTAGGAGT TGCACAACCA TCTGGGCTCG CTGAGACCTG

1030 1040 1050 1060 1070 1080

GGCAGGCACA GGCCCAGGTT CTGACAAGCA GAGGGTGAAA GGTTTCGTTC TAGGCCTGAA

1090 1100 1110 1120 1130 1140

GGGCCTTACA GGGCAGCCAG GGCACTACAG CCTCTAAAGT CCCAGCATCT G

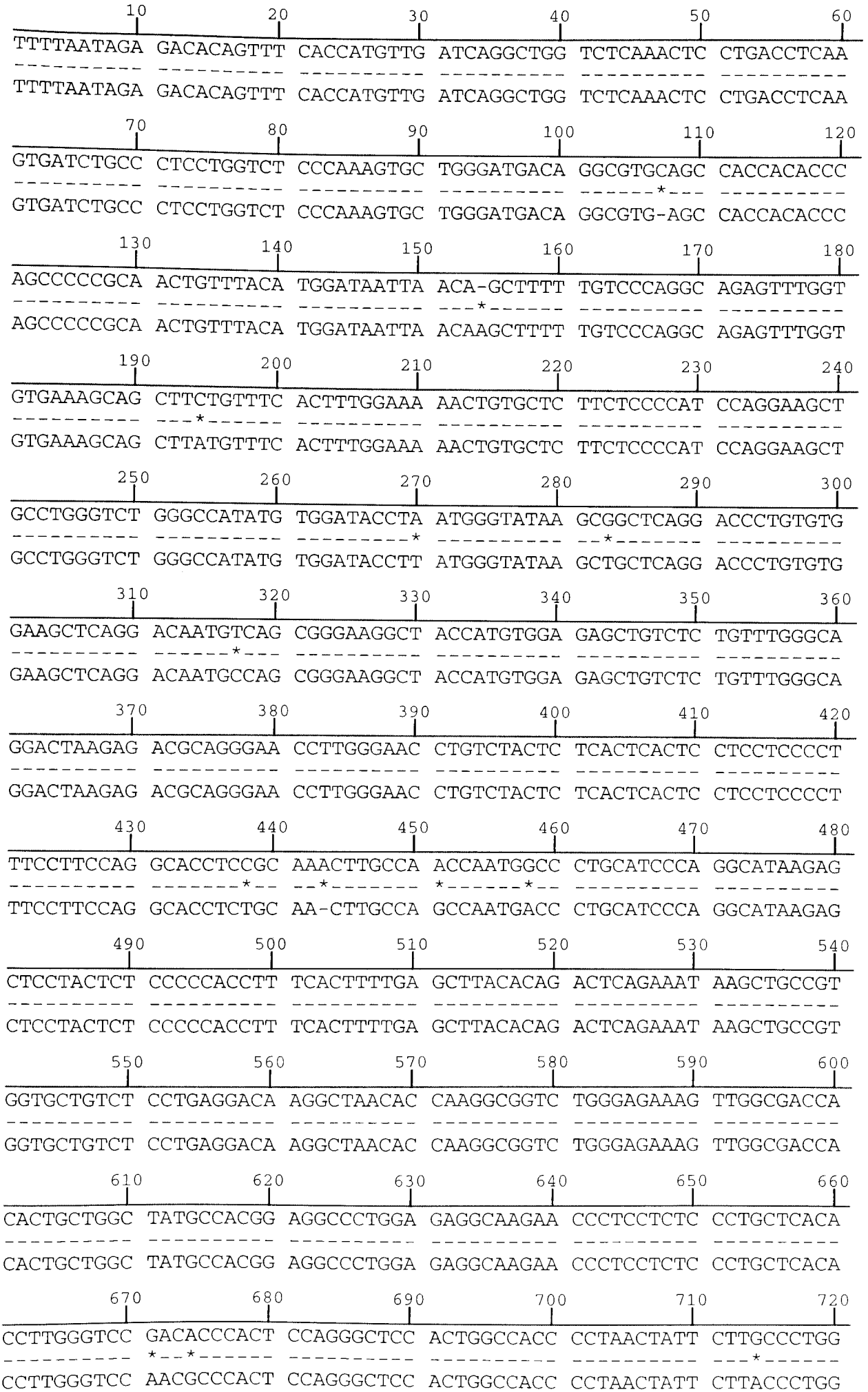
## A4.5: Acc Region

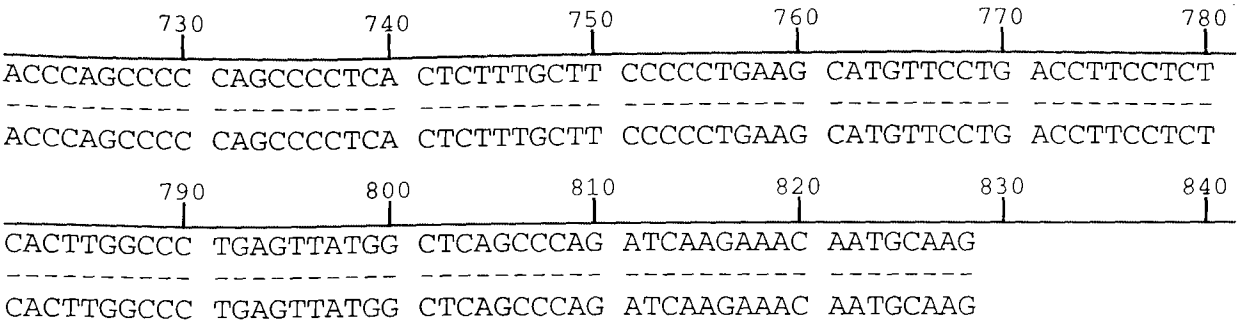
Sequences of human (bottom) and chimpanzee (top) aligned.
Refer to Sawada and Schmid, [1986] for full ψα1 *Alu* sequence.

Features
ψα1 *Alu* repeat ends at 122
ψα2 exon 1 from 539-635
IVS 1 from 636-761
ψα2 exon 1 begins at 762.

**Polymorphisms**

| Position in Human Sequence | Position in Aligned Sequence |
|---|---|
| 395 | 396 |
| 424 | 425 |
| 573 | 575 |
| 594 | 596 |




730 740 750 760 770 780
ACCCAGCCCC CAGCCCCTCA CTCTTTGCTT CCCCCTGAAG CATGTTCCTG ACCTTCCTCT
ACCCAGCCCC CAGCCCCTCA CTCTTTGCTT CCCCCTGAAG CATGTTCCTG ACCTTCCTCT
790 800 810 820 830 840
CACTTGGCCC TGAGTTATGG CTCAGCCCAG ATCAAGAAAC AATGCAAG
CACTTGGCCC TGAGTTATGG CTCAGCCCAG ATCAAGAAAC AATGCAAG

## A4.6: Rsa Region

Human sequence only.

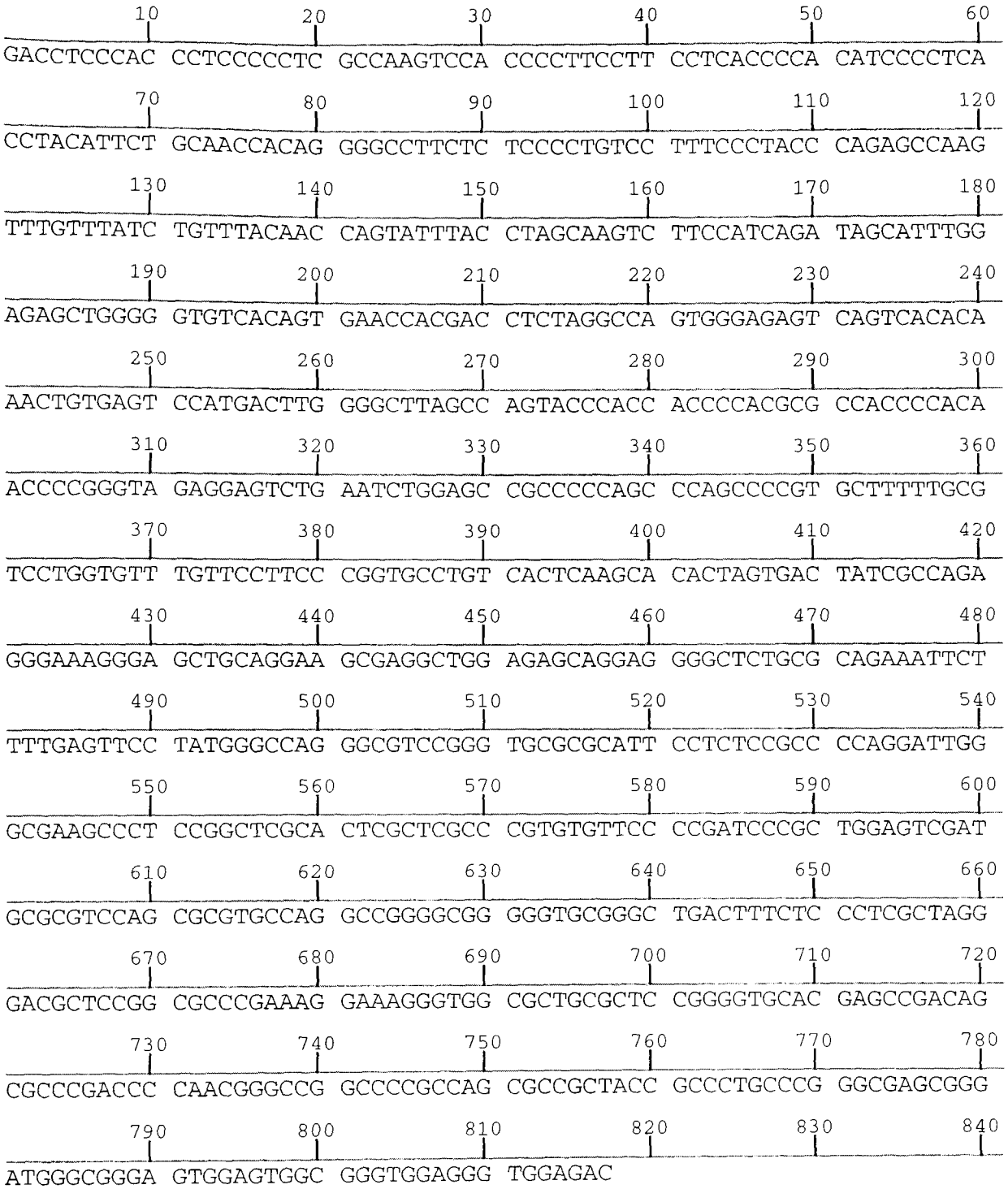
Features

Z box begins at position 138

The end of the Rsa region is 189 bp 3' to the α2 gene Cap site

**Polymorphisms**

| Position relative to Cap site | Position in Human Sequence |
|---|---|
| 734 | 273 |
| 347 | 660 |
| 331-330 | 675-676 |
| 278 | 729 |


10 20 30 40 50 60
GACCTCCCAC CCTCCCCCTC GCCAAGTCCA CCCCTTCCTT CCTCACCCCA CATCCCCTCA
70 80 90 100 110 120
CCTACATTCT GCAACCACAG GGGCCTTCTC TCCCCTGTCC TTTCCCTACC CAGAGCCAAG
130 140 150 160 170 180
TTTGTTTATC TGTTTACAAC CAGTATTTAC CTAGCAAGTC TTCCATCAGA TAGCATTTGG
190 200 210 220 230 240
AGAGCTGGGG GTGTCACAGT GAACCACGAC CTCTAGGCCA GTGGGAGAGT CAGTCACACA
250 260 270 280 290 300
AACTGTGAGT CCATGACTTG GGGCTTAGCC AGTACCCACC ACCCCACGCG CCACCCCACA
310 320 330 340 350 360
ACCCCGGGTA GAGGAGTCTG AATCTGGAGC CGCCCCCAGC CCAGCCCCGT GCTTTTTGCG
370 380 390 400 410 420
TCCTGGTGTT TGTTCCTTCC CGGTGCCTGT CACTCAAGCA CACTAGTGAC TATCGCCAGA
430 440 450 460 470 480
GGGAAAGGGA GCTGCAGGAA GCGAGGCTGG AGAGCAGGAG GGGCTCTGCG CAGAAATTCT
490 500 510 520 530 540
TTTGAGTTCC TATGGGCCAG GGCGTCCGGG TGCGCGCATT CCTCTCCGCC CCAGGATTGG
550 560 570 580 590 600
GCGAAGCCCT CCGGCTCGCA CTCGCTCGCC CGTGTGTTCC CCGATCCCGC TGGAGTCGAT
610 620 630 640 650 660
GCGCGTCCAG CGCGTGCCAG GCCGGGGCGG GGGTGCGGGC TGACTTTCTC CCTCGCTAGG
670 680 690 700 710 720
GACGCTCCGG CGCCCGAAAG GAAAGGGTGG CGCTGCGCTC CGGGGTGCAC GAGCCGACAG
730 740 750 760 770 780
CGCCCGACCC CAACGGGCCG GCCCCGCCAG CGCCGCTACC GCCCTGCCCG GGCGAGCGGG
790 800 810 820 830 840
ATGGGCGGGA GTGGAGTGGC GGGTGGAGGG TGGAGAC

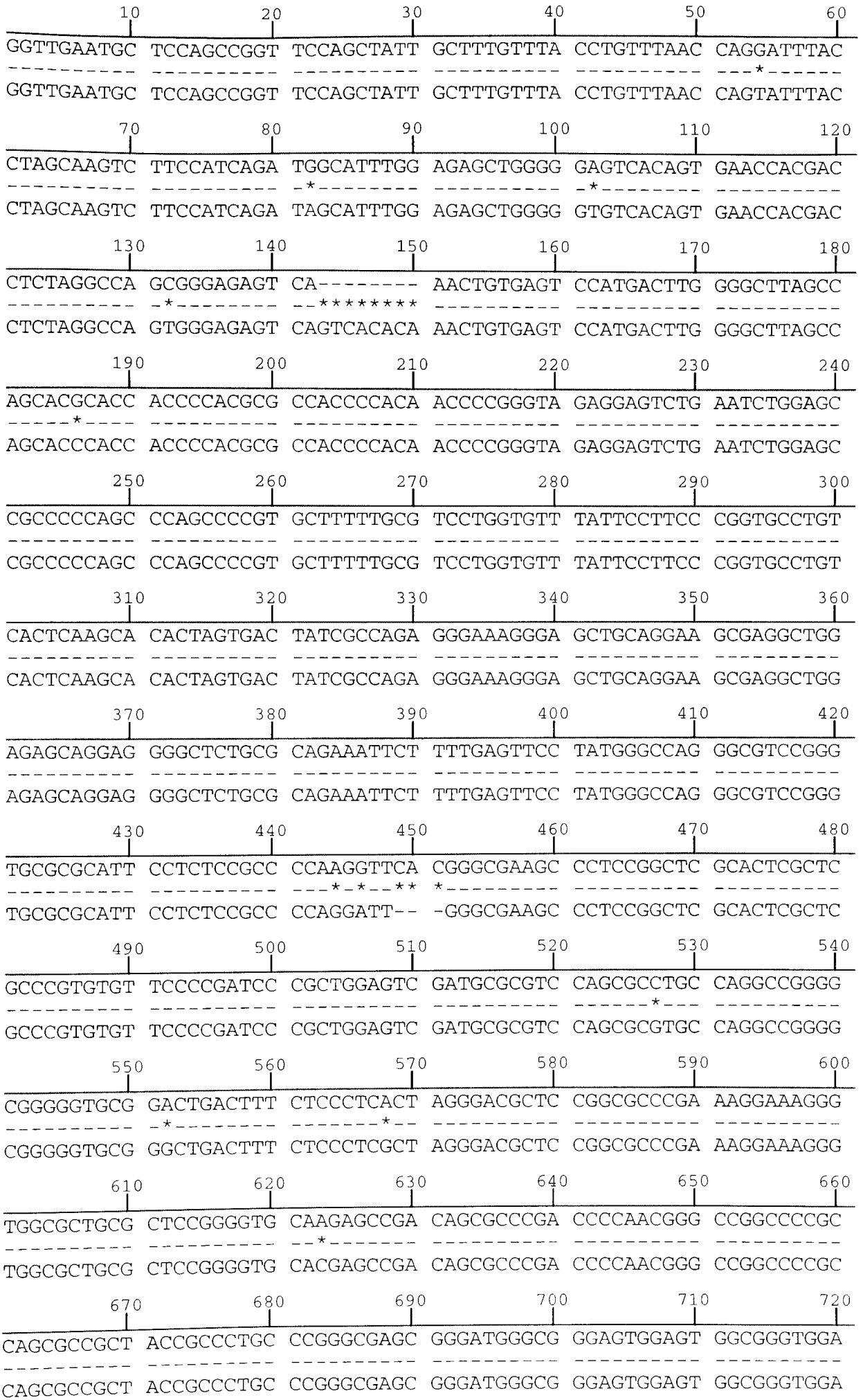
## A4.7: α Pst Region

Sequences of human (bottom) and chimpanzee (top) aligned.

Features
Z box starts at 48
α1 gene Cap site at 919

**Polymorphisms**

| Position relative to Cap Site | Position in Aligned Human Sequence |
|---|---|
| 662 | 258 |
| 635 | 282 |
| 414 | 506 |
| 347 | 573 |
| 278 | 642 |



```
          730        740        750        760        770        780
GGGTGGAGAC GTCCTGGCCC CCGCCCCGCG TGCACCCCCA GGGGAGGCCG AGCCCGCCGC
---------- ---------- ---------- ---------- ---------- ----------
GGGTGGAGAC GTCCTGGCCC CCGCCCCGCG TGCACCCCCA GGGGAGGCCG AGCCCGCCGC

          790        800        810        820        830        840
CCGGCCCCGC GCAGGCCCCG CCCGGGACTC CCCTGCGGTC CAGGCCGCGC CCCGGGCTCC
---------- ---------- ---------- ---------- ---------- ----------
CCGGCCCCGC GCAGGCCCCG CCCGGGACTC CCCTGCGGTC CAGGCCGCGC CCCGGGCTCC

          850        860        870        880        890        900
GCGCCAGCCA ATGAGCGCCG CCCGGCCGGG CGTGCCCCCG CGCCCCAAGC ATAAACCCTG
---------- ---------- ---------- ---------- ---------- ----------
GCGCCAGCCA ATGAGCGCCG CCCGGCCGGG CGTGCCCCCG CGCCCCAAGC ATAAACCCTG

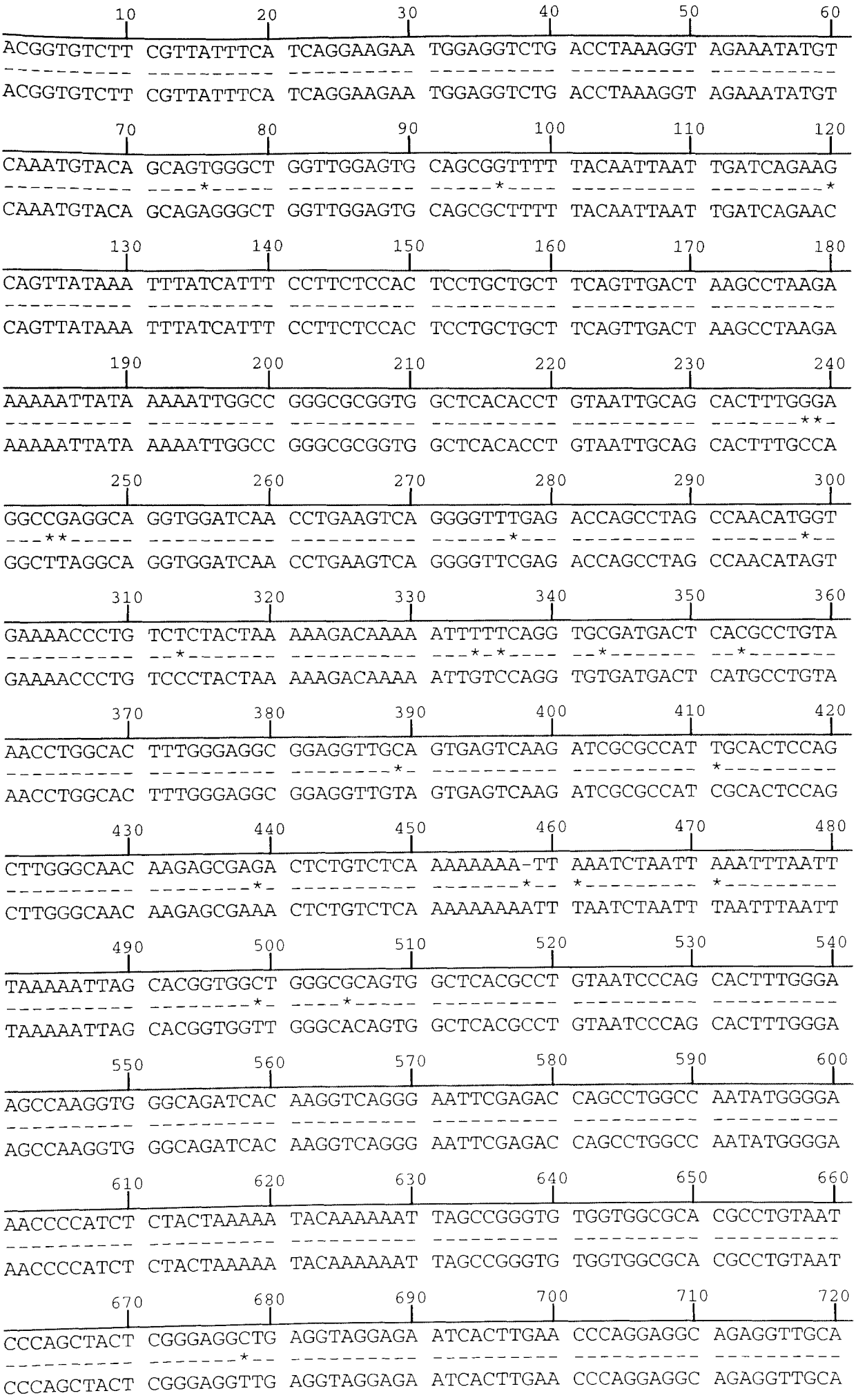
          910        920        930        940        950        960
GCGCGCTCGC GGCCCGGCAC TCTTCTGGTC CCCACAGACT C
---------- ---------- ---------- ---------- -
GCGCGCTCGC GGCCCGGCAC TCTTCTGGTC CCCACAGACT C
```

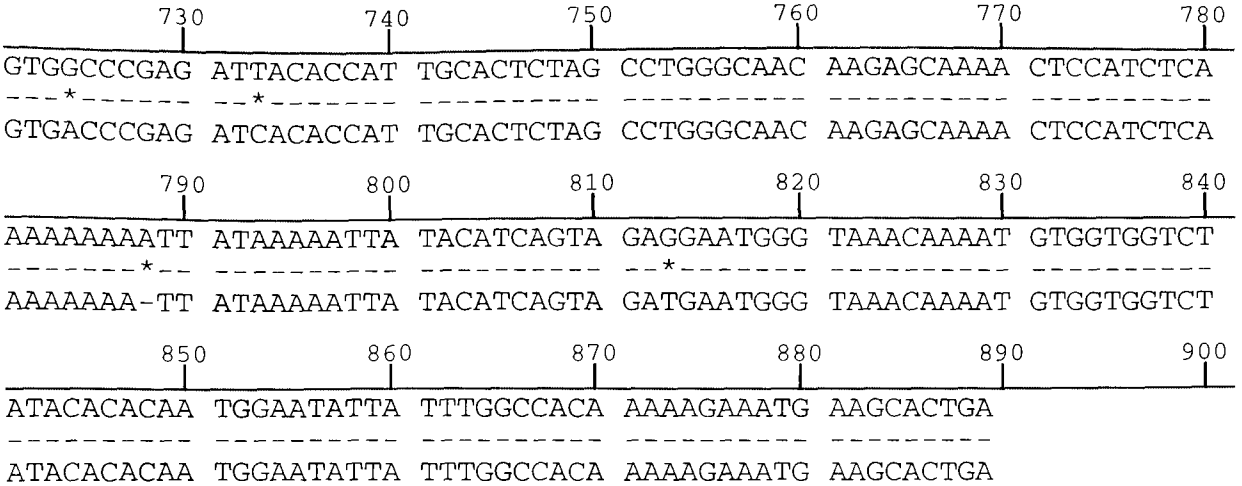
## A4.8: θ Pst Region

Sequences of human (bottom) and chimpanzee (top) aligned

Features

Ro pseudogene from 76-178
*Alu* repeat 1 from 195-458
*Alu* repeat 2 from 496-789




730 740 750 760 770 780
GTGGCCCGAG ATTACACCAT TGCACTCTAG CCTGGGCAAC AAGAGCAAAA CTCCATCTCA
---*------ --*------- ---------- ---------- ---------- ----------
GTGACCCGAG ATCACACCAT TGCACTCTAG CCTGGGCAAC AAGAGCAAAA CTCCATCTCA
790 800 810 820 830 840
AAAAAAAATT ATAAAAATTA TACATCAGTA GAGGAATGGG TAAACAAAAT GTGGTGGTCT
-------*-- ---------- ---------- --*------- ---------- ----------
AAAAAAA-TT ATAAAAATTA TACATCAGTA GATGAATGGG TAAACAAAAT GTGGTGGTCT
850 860 870 880 890 900
ATACACACAA TGGAATATTA TTTGGCCACA AAAAGAAATG AAGCACTGA
---------- ---------- ---------- ---------- ---------
ATACACACAA TGGAATATTA TTTGGCCACA AAAAGAAATG AAGCACTGA

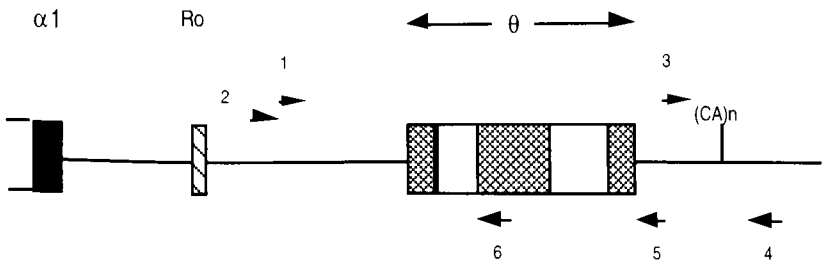
# Appendix 5: Non-Haplotyping PCR Primers And Conditions

Table A5.1: Optimised Non-Haplotyping PCRs

| Region | Primer sequences (5'-3') | Position in α-cluster | Cycling Conditions | | | | Buffer (Mg2+) | Comment |
|---|---|---|---|---|---|---|---|---|
| | | | Denaturation | Annealing | Extension | No. cycles | | |
| 3' θ | Forward: GAATGGGGTGGTAGGTGAGGGC<br>Reverse: CATGTGCCTAGAAGACTGGAACAGC | 29998<br>35123 | 98-2.5<br>98-1.5<br>98-1.5 | 60-1<br>60-1<br>60-1 | 72-1.5<br>72-1.5<br>72-5.0 | 2<br>27<br>1 | 2.5mM | *Sma I* |
| θ gene | Forward: GGTCGGAGCTGTCGGAGATTGAGCG<br>Reverse: CTGCTCCTGGGAGAGGCTTTACTCAAACAC | 29295<br>30156 | 95-3<br>95-1.5<br>95-1.5 | 66-1<br>66-1<br>66-1 | 72-2<br>72-2<br>72-5 | 2<br>27<br>1 | 2.5mM | θ |
| 5' θ | Forward: TAAATGAGAAAACAGACTCGGGCAA<br>Reverse: CGGGGAAAGCCAGGAAGGTC | 28830<br>29661 | 95-3<br>95-1.5<br>95-1.5 | 62-1<br>62-1<br>62-1 | 72-2<br>72-2<br>72-5 | 2<br>27<br>1 | 1.5mM | - |
| 3' ψζ1 gene | Forward: AAGGAGGGAACGATTAGGAG<br>Reverse: CCTCCTTCCCTCCAATACTGAGAA | 12835<br>14275 | 98-1.5<br>98-1<br>98-1 | 60-1.5<br>61-1<br>61-1 | 72-2.5<br>72-2<br>72-5 | 2<br>27<br>1 | 1.5mM | IVSII.HVR |
| ζ2 Gene | Forward: TAAGGCCACAGGAGAGGAACAGGAGTG<br>(Use with PZ.R7; Chapter 2) | -187 | Conditions as described in Chapter 2 for PZ/Z site | | | | | |
| ψζ1/ζ1 Gene | Reverse: CTCCTAATCGTTCCCTCCTTACCCT<br>(Use with PZ.F1; Chapter 2) | 12854 | Conditions as described in Chapter 2 for PZ/Z site | | | | | Alternative to PZ.R7 (Chap 2) |

Cycling conditions given in minutes at each temperature. Buffer and other reaction components are as described for α-haplotyping PCRs in Chapter 2. Refer to Chapter 4 for details of ζ2 PCR and amplification using the PZ.R2 primer, and to Chapter 6 for rationale for the development of 3' ψζ1 gene PCR. Note: the θ gene is located at positions 30,167 (initiation codon) to 30,788 (termination codon) in the α-gene cluster (refer to [Hsu *et al.*, 1988] for θ gene sequence).

Figure A5.1: Primers at 3' end of α-cluster




1 and 5 = θ gene PCR, 2 and 6 = 5' θ gene PCR, and 3 and 4 = 3' θ gene PCR.

Table A5.2: *Sma I* Polymorphism and Haplotype

The 3' θ gene region contains a $(CA)_n$ repeat (Figure A5.1) which was found to be polymorphic in six Melanesian chromosomes (two different alleles). This region also contains a polymorphic *Sma I* site. Table A5.2 shows characterisation of this site in 42 human chromosomes from world populations. The rarer (-) allele is not apparently associated with specific haplotypes.

| Population | *Sma I* | Haplotype |
|---|---|---|
| Senegal | +/+ | Ia/IIc |
| W. Samoa | +/+ | Ia/IVa |
| Britain | +/+ | Ia/IVc |
| Britain | +/+ | Ia/VIa |
| Saudi Arabia | +/+ | Ia/IIa |
| Iceland | +/- | Ia/IIa |
| Brunei | +/+ | IIa/IIIa |
| Senegal | +/+ | IIIa/IIIf |
| Iceland | +/+ | IIIa/IIIb |
| Austr Aborig | +/- | IIIa/IInovel |
| Senegal | +/+ | IIc/VIb |
| Philippines | +/+ | IIc/IIe |
| Senegal | +/+ | IIe/IIId |
| Ponape | +/+ | VInovel/Vc |
| San | +/+ | Vc/IIg |
| New Guinea (H) | +/+ | Vd/IVb |
| Bantu | +/+ | Vc/IIIf |
| Senegal | +/+ | VIIc/VIIf |
| Bantu | +/+ | IVa/VIIi |
| Melanesia | +/+ | IVa/IVa |
| Melanesia | -/- | IIIa/IIIa |

Melanesian chromosomes from Vanuatu. (H)= highlands. Samples as described in Chapter 3.

# References

Alani, E., Reenan, R.A.G. and Kolodner, R.D. (1994). Interaction between mismatch repair and genetic recombination in *Saccharomyces cerevisiae*. Genetics **137**: 19-39.

Allison, A.C. (1964). Polymorphism and natural selection in human populations. Cold Spring Harbor Sympos. Quant. Biol. **29**: 137-149.

Antonarakis, S.E., Kazazian, H.H. and Orkin, S.H. (1985). DNA polymorphisms and molecular pathology of the human globin gene clusters. Hum. Genet. **69**: 1-14.

Aquadro, C.F. (1992). Why is the genome variable? TIG **8**: 355-362.

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L. and Batzer, M.A. (1995). *Alu* repeats: a source for the genesis of primate microsatellites. Genomics **29**: 136-144.

Aris-Brosou, S. and Excoffier, L. (1996). The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. **13**: 494-504.

Armour, J.A.L., Crosier, M. and Jeffreys, A.J. (1992). Human minisatellite alleles detectable only after PCR amplification. Genomics **12**: 116-124.

Armour, J.A.L., Crosier, M. and Jeffreys, A.J. (1996). Distribution of tandem repeat polymorphism within minisatellite MS621 (D5S110). Ann. Hum. Genet. (Lond). **60**: 11-20.

Armour, J.A.L., Harris, P.C. and Jeffreys, A.J. (1993). Allelic diversity at minisatellite MS205 (D16S309): evidence for polarized variability. Hum. Mol. Genet. **2**: 1137-1145.

Armour, J.A.L., Crosier, M., Malcolm, S., Chan, J.C.T. and Jeffreys, A.J. (1995). Human minisatellite loci composed of interspersed GGA-GGT triplet repeats. Proc. Roy. Soc. Lond. (Biol.) **261**: 345-349.

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

Arnheim, N. and Erlich, H. (1992). Polymerase Chain Reaction Strategy. Annu. Rev. Biochem. **61**: 131-156.

Aronow, B., Lattier, D., Silbiger, R., Dusing, M., Hutton, J., Jones, G., Stock, J., McNeish, J., Potter, S., Witte, D. and Wiginton, D. (1989). Evidence for a complex regulatory array in the first intron of the human adenosine deaminase gene. Genes Dev. **3**: 1384-1400.

Arteiler, B. and Porteous, D.J. (1992). Polyacyrlamide-agarose composite gels for high resolution of fingerprints. TIG **8**: 82.

Ashley, T. (1994). Mammalian meiotic recombination: a reexamination. Hum. Genet. **94**: 587-593.

Ayala, F.J., Escalante, A., O'Huigin, C. and Klein, J. (1994). Molecular genetics of speciation and human origins. Proc. Natl. Acad. Sci. USA **91**: 6787-6794.

Ayala, F.J. and Escalante, A.A. (1996). The evolution of human populations: a molecular perspective. Molecular Phylogenetics and Evolution **5**: 188-201.

Baker, M.D. and Read, L.R. (1995). High-frequency gene conversion between repeated Cμ sequences intergrated at the chromosomal immunoglobulin μ locus in mouse hybridoma cells. Mol. Cell. Biol. **15**: 766-771.

Bailey, W.J., Hayasaka, K., Skinner, C.G., Kehoe, S., Sieu, L.C., Slightom, J.L. and Goodman, M. (1992). Reexamination of the African hominoid trichotomy with additional sequences from the primate β-globin gene cluster. Molecular Phylogenetics and Evolution **2**: 97-135.

Ballard, J.W.O. and Kreitman, M. (1995). Is mitochondrial DNA a strictly neutral marker? Trends in Ecology and Evolution **10**: 485-488.

Baltimore, D. (1981). Gene conversion: some implications for immunoglobulin genes. Cell **24**: 592-594.

Batzer, M.A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D.H., Shaikh, T.H., Novick, G.E., Ioannou, P.A., Scheer, W.D., Herrera, R.J. and Deininger, R.L. (1994). African origin of human-specific polymorphic *Alu* insertions. Proc. Natl. Acad. Sci. USA **91**: 12288-12292.

Baysal, E. and Huisman, T.H.J. (1994). Detection of common deletional α-thalassemia-2 determinants by PCR. Am. J. Hematol. **46**: 208-213.

Beck, K.-F., Sathopulos, I., Berninger, M.G. and Schweizer, M. (1993). Overcoming GC compression in nucleotide sequencing. BioTechniques **14**: 375.

Begun, D.J. and Aquadro, C. (1995). Evolution at the tip and base of the X chromosome in an African population of *Drosophila melanogaster*. Mol. Biol. Evol. **12**: 382-390.

Begun, D.J. and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D.melanogaster*. Nature **356**: 519-520.

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

Bellwood, P.S. (1989). The colonization of the Pacific: some current hypotheses. In: The Colonization of the Pacific: A Genetic Trail, Hill, A.V.S. and Serjeantson, S.W. (eds.). Oxford; Oxford University Press.

Bennett, S.T., Lucassen, A.M., Gough, S.C.L., Powell, E.E., Undlien, D.E., Pritchard, L.E., Merriman, M.E., Kawaguchi, Y., Dronsfield, M.J., Pociot, F., Nerup, J., Bouzekri, N., Cambon-Thomsen, A., Ronningen, K.S., Barnett, A.H., Bain, S.C. and Todd, J.A. (1995). Susceptibility to human type I diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nature Genet. **9**: 284-292.

Bernardi, G. (1989). The isochore organisation of the human genome. Annu. Rev. Genet. **23**: 637-661.

Bookstein, R., Lai, C.-C., To, H. and Lee, W.-H. (1990). PCR-based detection of a polymorphic *Bam*H1 site in intron 1 of the human retinoblastoma (RB) gene. Nucl. Acids Res. **18**: 1666.

Borts, R. and Haber, J.E. (1989). Length and distribution of meiotic gene conversion tracts and crossovers in *Saccharomyces cerevisiae*. Genetics **123**: 69-80.

Borts, R.H. and Haber, J.E. (1987). Meiotic recombination in Yeast: alteration by multiple heterozygosities. Science **237**: 1459-1465.

Bowcock, A.M., Kidd, J.R., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd, K.K. and Cavalli-Sforza, L.L. (1991). Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc. Natl. Acad. Sci. USA **88**: 839-843.

Bowden, D.K., Vickers, M.A. and Higgs, D.R. (1992). A PCR strategy to detect the common severe determinants of α-thalassaemia. Brit. J. Haemat. **81**: 104-108.

Britten, R.J., Baron, W.F., Stout, D.B. and Davidson, E.H. (1988). Sources and evolution of human *Alu* repeated sequences. Proc. Natl. Acad. Sci. USA **85**: 4770-4774.

Brookfield, J.F.Y. and Sharp, P.M. (1994). Neutralism and selectionism face up to DNA data. TIG **10**: 109-110.

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

Buckle, V.J., Higgs, D.R., Wilkie, A.O.M., Super, M. and Weatherall, D.J. (1988). Localisation of human α-globin to 16p13.3 to pter. J. Med. Genet. **25**: 847-859.

Cann, R.L., Stoneking, M. and Wilson, A.C. (1987). Mitochondrial DNA and human evolution. Nature **325**: 31-36.

Carpenter, A.T.C. (1987). Gene conversion, recombination nodules and the initiation of meiotic synapsis. BioEssays **6**: 232-236.

Castelloe, J. and Templeton, A.R. (1994). Root probabilities for intraspecific gene trees under neutral coalescent theory. Molecular Phylogenetics and Evolution **3**: 102-113.

Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. and Mountain, J. (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc. Natl. Acad. Sci. USA **85**: 6002-6006.

Chakravarti, A. (1994). The behaviour of meiosis in sperm. Am. J. Hum. Genet. **55**: 421-422.

Chakravarti, A., Buetow, K.H., Antonarakis, S.E., Waber, P.G., Boehm, C.D. and Kazazian, H.H. (1984). Nonuniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. **36**: 1239-1258.

Chakravarti, A., Elbein, S.C. and Permutt, M.A. (1986). Evidence for increased recombination near the human insulin gene: implication for disease association studies. Proc. Natl. Acad. Sci. USA **83**: 1045-1049.

Chakraborty, R. and Jin, L. (1992). Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. Hum. Genet. **88**: 267-272.

Chapman, B.S., Vincent, K.A. and Wilson, A.C. (1986). Persistence or rapid generation of DNA length polymorphism at the ζ-globin locus of humans. Genetics **112**: 79-92.

Charlesworth, B., Morgan, M.T. and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics **134**: 1289-1303.

Charmley, P., Chao, A., Concannon, P., Hood, L. and Gatti, R.A. (1990). Haplotyping the human T-cell receptor β-chain gene complex by use of restriction fragment length polymorphisms. Proc. Natl. Acad. Sci. USA **87**: 4823-4827.

Chen, Y.-S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A.S. and Wallace, D.C. (1995). Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am. J. Hum. Genet. **57**: 133-149.

Cheng, S., Fockler, C., Barnes, W.M. and Higuchi, R. (1994). Effective amplification of long targets from cloned inserts and human genomic DNA. Proc. Natl. Acad. Sci. USA **91**: 5695-5699.

Clark, A.G. (1987). Neutrality tests of highly polymorphic restriction-fragment-length polymorphisms. Am. J. Hum. Genet. **41**: 948-956.

Clegg, J.B. (1987a). Can the product of the θ gene be a real globin? Nature **329**: 465-466.

Clegg, J.B. (1987b). Gene conversions in the horse α-globin gene complex. Mol. Biol. Evol. **4**: 492-503.

Cohen, D.W. and Athieno Odhiambo, E.S. (1989). Siaya. The Historical Anthropology of an African Landscape. Currey.

Collis, P., Antoniou, M. and Grosveld, F. (1990). Definition of the minimal requirements within the human β-globin gene and the dominant control region for high level expression. EMBO J. **9**: 233-240.

Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli*. Nature **274**: 775-780.

Craddock, C.F., Vyas, P., Sharpe, J.A., Ayyub, H., Wood, W.G. and Higgs, D.R. (1995). Contrasting effects of α- and β-globin regulatory elements on chromatin structure may be related to their different chromosomal environments. EMBO J. **14**: 1718-1726.

Craig, J.M. and Bickmore, W.A. (1993). Chromosome bands - flavours to savour. BioEssays **15**: 349-354.

Crandall, K.A. and Templeton, A.R. (1993). Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. Genetics **134**: 959-969.

Cullen, M., Erlich, H., Klitz, W. and Carrington, M. (1995). Molecular mapping of a recombination hotspot located in the second intron of the human TAP2 locus. Am. J. Hum. Genet. **56**: 1350-1358.

Curtis, D. and Bender, W. (1991). Gene conversion in *Drosophila* and the effects of the meiotic mutants mei-9 and mei-218. Genetics **127**: 739-746.

Datta, A., Adjiri, A., New, L., Crouse, G.F. and Jinks-Robertson, S. (1996). Mitotic crossovers between diverged sequences are regulated by mismatch repair proteins in *Saccharomyces cerevisiae*. Mol. Cell. Biol. **16**: 1085-1093.

Decorte, R., Wu, R., Maryne, P. and Cassiman, J.-J. (1994). Identification of internal variation in the pseudoautosomal VNTR DXYS17, with nonrandom distribution of alleles on the X and the Y chromosomes. Am. J. Hum. Genet. **54**: 506-515.

Deka, R., Chakraborty, R. and Ferrell, R.E. (1991). A population genetic study of six VNTR loci in three ethnically defined populations. Genomics **11**: 83-92.

Deka, R., Jin, L., Shriver, M.D., Yu, L.M., DeCroo, S., Hundrieser, J., Bunker, C.H., Ferrell, R.E. and Chakraborty, R. (1995). Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations. Am. J. Hum. Genet. **56**: 461-474.

Desmarais, E., Vigneron, S., Buresi, C., Cambien, F., Cambou, J.P. and Roizes, G. (1993). Variant mapping of the Apo(B) AT rich minisatellite. Dependence on nucleotide sequence of the copy number variations. Instability of the non-canonical alleles. Nucl. Acids Res. **21**: 2179-2184.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J. and Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**: 152-154.

Dillon, N. and Grosveld, F. (1993). Transcriptional regulation of multigene loci: multilevel control. TIG **9**: 134-137.

DiRienzo, A. and Wilson, A.C. (1991). Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc. Natl. Acad. Sci. USA **88**: 1597-1601.

Dodgson, J.B., McCune, K.C., Rusling, D.J., Krust, A. and Engel, J.D. (1981). Adult chicken $\alpha$-globin genes, $\alpha^A$ and $\alpha^D$: no anemic shock $\alpha$-globin exists in domestic chickens. Proc. Natl. Acad. Sci. USA **78**: 5998-6002.

Donnelly, P. and Tavare, S. (1986). The ages of alleles and a coalescent. Advances in Applied Probability **18**: 1-19.

Dorit, R.L., Akashi, H. and Gilbert, W. (1995). Absence of polymorphism at the ZFY locus on the human Y chromosome. Science **268**: 1183-1186.

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

Dover, G.A. (1986). Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. TIG 159-165.

Dover, G.A. (1989). Victims or perpetrators of DNA turnover? Nature **342**: 347-348.

Dover, G.A. (1993). Evolution of genetic redundancy for advanced players. Curr. Opinion Genet. Dev. **3**: 902-910.

Dutton, C.M., Paynton, C. and Sommer, S.S. (1993). General method of amplifying regions of very high G+C content. Nucl. Acids Res. **21**: 2953-2954.

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

Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980). The structure and evolution of the human β-globin gene family. Cell **21**: 653-668.

Eichler, E.E., Holden, J.J.A., Popovich, B.W., Reiss, A.L., Snow, K., Thibodeau, S.N., Richards, C.S., Ward, P.A. and Nelson, D.L. (1994). Length of uninterrupted CGG repeats determines instability in the FMR1 gene. Nature Genet. **8**: 88-94.

Elbein, S.C. (1992). Linkage disequilibrium among RFLPs at the insulin-receptor locus despite intervening *Alu* repeat sequences. Am. J. Hum. Genet. **51**: 1103-1110.

Ellsworth, D.L., Shriver, M.D. and Boerwinkle, E. (1995). Nucleotide sequencing analysis of the apolipoprotein B 3' VNTR. Hum. Mol. Genet. **4**: 937-944.

Erickson, L.M. and Maeda, N. (1994). Parallel evolutionary events in the haptoglobin gene clusters of rhesus monkey and human. Genomics **22**: 579-589.

Erlich, H.A. (1989). PCR Technology. Principles and Applications for DNA Amplification.

Erlich, H.A. and Gyllensten, U.B. (1991). Shared epitopes among HLA class II alleles: gene conversion, common ancestry and balancing selection. Immunol. Today **12**: 411-414.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. Theoretical Population Biology **3**: 87-112.

Ewens, W.J., Spielman, R.S. and Harris, H. (1981). Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc. Natl. Acad. Sci. USA **78**: 3748-3750.

Excoffier, L. (1990). Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. J. Mol. Evol. **30**: 125-139.

Excoffier, L. and Langaney, A. (1989). Origin and differentiation of human mitochondrial DNA. Am. J. Hum. Genet. **44**: 73-85.

Excoffier, L., Pellegrini, B., Sanchez-Mazas, A., Simon, C. and Langaney, A. (1987). Genetics and history of sub-Saharan Africa. Yearbook of Physical Anthropology **30**: 151-194.

Excoffier, L. and Slatkin, M. (in press). Maximum-Likelihood estimation of molecular haplotype frequencies in a diploid population.

Feinberg, A.P. and Vogelstein, B. (1983). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. Annals of Biochemistry **132**: 6-13.

Fischel-Ghodsian, N., Higgs, D.R. and Beyer, E.C. (1987). Function of a new globin gene. Nature **329**: 397.

Fitch, D.H.A., Mainone, C., Goodman, M. and Slightom, J.L. (1990). Molecular history of gene conversions in the primate fetal γ-globin genes. J. Biol. Chem. **265**: 781-793.

Flam, F. (1994). Hints of a Language in Junk DNA. Science **266**: 1320.

Fletcher, H.L. (1994). Possible loss of length conservation and reciprocity during recombination or conversion in tandem arrays. Genetics **138**: 511-518.

Flint, J., Boyce, A.J., Martinson, J.J. and Clegg, J.B. (1989). Population bottlenecks in Polynesia revealed by minisatellites. Human Genetics **83**: 257-263.

Flint, J., Harding, R.M., Clegg, J.B. and Boyce, A.J. (1993). Why are some genetic diseases common? Distinguishing selection from other processes by molecular analysis of globin gene variants. Hum. Genet. **91**: 91-117.

Flint, J., Hill, A.V.S., Bowden, D.K., Oppenheimer, S.J., Sill, P.R., Serjeantson, S.W., Bana-Koiri, J., Bhatia, K., Alpers, M.P., Boyce, A.J., Weatherall, D.J. and Clegg, J.B. (1986). High frequencies of α-thalassaemia are the result of natural selection by malaria. Nature **321**: 744-750.

Fuerst, P.A., Chakraborty, R. and Nei, M. (1977). Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics **86**: 455-483.

Fullerton, S.M. (1995). Sequence Variation at the β-Globin Locus. D.Phil.Thesis. Oxford University.

Fullerton, S.M., Harding, R.M., Boyce, A.J. and Clegg, J.B. (1994). Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. Proc. Natl. Acad. Sci. USA **91**: 1805-1809.

Gelfand, D. (1989). Taq DNA Polymerase. In: PCR Technology. Principles and Applications for DNA Amplification, Erlich, H.A. (eds.). Macmillan publishers Ltd.

Giles, E. and Ambrose, S.H. (1986). Are we all out of Africa? Nature **322**: 21-22.

Gillespie, J.H. (1987). Molecular evolution and the neutral allele theory. Oxford Surveys of Evolutionary Biology **4**: 10-37.

Goldstein, D.B., Linares, A.R., Cavalli-Sforza, L.L. and Feldman, M.W. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92**: 6723-6727.

Goodbourn, S.E.Y. (1983). Analysis of the human α-globin gene complex. D.Phil. Thesis. Oxford University.

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

Goodbourn, S.E.Y., Higgs, D.R., Clegg, J.B. and Weatherall, D.J. (1984). Allelic variation and linkage properties of a highly polymorphic restriction fragment in humans. Mol. Biol. Med. **2**: 223-238.

Graven, L., Passarino, G., Semino, O., Boursot, P., Santachiara-Benerecetti, S., Langaney, A. and Excoffier, L. (1995). Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. Mol. Biol. Evol. **12**: 334-345.

Gray, I.C. and Jeffreys, A.J. (1991). Evolutionary transience of hypervariable minisatellites in man and the primates. Proc. Roy. Soc. Lond. (Biol.) **243**: 241-253.

Green, M. and Krontiris, T.G. (1993). Allelic variation of reporter gene activation by the HRAS1 minisatellite. Genomics **17**: 429-434.

Greenberg, J.H. (1963). The Languages of Africa. The Hague: Mouton.

Grimm, T., Muller, B., Dreier, M., Kind, E., Bettecken, T., Meng, G. and uller, C.R.M. (1989). Hot spot of recombination within DXS174 in the Duchenne muscular dystrophy gene. Am. J. Hum. Genet. **45**: 368-372.

Grompe, M. (1993). The rapid detection of unknown mutations in nucleic acids. Nature Genet. **5**: 111-117.

Harding, R.M., Boyce, A.J. and Clegg, J.B. (1992). The evolution of tandemly repetitive DNA: recombination rules. Genetics **132**: 847-859.

Harding, R.M., Boyce, A.J., Martinson, J.J., Flint, J. and Clegg, J.B. (1993). A computer simulation study of VNTR population genetics: Constrained recombination rules out the infinite alleles model. Genetics **135**: 911-922.

Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S. and Clegg, J.B. (in preparation). Archaic African and Asian lineages in the genetic ancestry of modern humans.

Hardison, R.C. and Gelinas, R.E. (1986). Assignment of orthologous relationships among mammalian α-globin genes by examining flanking regions reveals a rapid rate of evolution. Mol. Biol. Evol. **3**: 243-261.

Hardison, R. and Miller, W. (1993). Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. Mol. Biol. Evol. **10**: 73-102.

Hardison, R.C., Sawada, I., Cheng, J.-F., Shen, C.-K.J. and Schmid, C.W. (1986). A previously undetected pseudogene in the human α-globin gene cluster. Nucl. Acids Res. **14**: 1903-1911.

Harpending, H.C., Sherry, S.T., Rogers, A.R. and Stoneking, M. (1993). The genetic structure of ancient human populations. Current Anthropology **34**: 483-496.

Harvey, C.B., Pratt, W.S., Islam, I., Whitehouse, D.B. and Swallow, D.M. (1995). DNA polymorphisms in the lactase gene. European Journal of Human Genetics **3**: 27-41.

Hawley, R.S. and Arbel, T. (1993). Yeast genetics and the fall of the classical view of meiosis. Cell **72**: 301-303.

Hess, J.F., Fox, M., Schmid, C. and Shen, C.-K.J. (1983). Molecular evolution of the human adult α-globin-like gene region: insertion and deletion of *Alu* family repeats and non-*Alu* DNA sequences. Proc. Natl. Acad. Sci. USA **80**: 5970-5974.

Hess, J.F., Schmid, C.W. and Shen, C.K.J. (1984). A gradient of sequence divergence in the human adult α-globin duplication units. Science **226**: 67-70.

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

Higgs, D.R., Hill, A.V.S., Bowden, D.K., Weatherall, D.J. and Clegg, J.B. (1984). Independent recombination events between the duplicated human α-globin genes; implications for their concerted evolution. Nucl. Acids Res. **12**: 6965-6977.

Higgs, D.R., Vickers, M.A., Wilkie, A.O.M., Pretorius, I.M., Jarman, A.P. and Weatherall, D.J. (1989). A review of the molecular genetics of the human α-globin gene cluster. Blood **73**: 1081-1104.

Higgs, D.R., Wainscoat, J.S., Flint, J., Hill, A.V.S., Thein, S.L., Nicholls, R.D., Teal, H., Ayyub, H., Peto, T.E.A., Falusi, A.G., Jarman, A.P., Clegg, J.B. and Weatherall, D.J. (1986). Analysis of the human α-globin cluster reveals a highly informative genetic locus. Proc. Natl. Acad. Sci. USA **83**: 5165-5169.

Higuchi, R.G. and Ochman, H. (1989). Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucl. Acids Res. **17**: 5865.

Hill, A.V.S., Nicholls, R.D., Thein, S.L. and Higgs, D.R. (1985). Recombination within the human embryonic ζ-globin locus: a common ζ-ζ chromosome produced by gene conversion of the ψζ gene. Cell **42**: 809-819.

Hill, A.V.S. and Wainscoat, J.S. (1986). The evolution of the α- and β-globin gene clusters in human populations. Hum. Genet. **74**: 16-23.

Hilliker, A.J., Harauz, G., Reaume, A.G., Gray, M., Clark, S.H. and Chovnick, A. (1994). Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. Genetics **137**: 1019-1026.

Hobbs, H.H., Russell, D.W., Brown, M.S. and Golding, J.L. (1990). The LDL receptor locus in familial hypercholesterolemia: mutational analysis of a membrane protein. Annu. Rev. Genet. **24**: 133-170.

Holmquist, G.P. (1992). Chromosome bands, their chromatin flavors, and their functional features. Am. J. Hum. Genet. **51**: 17-37.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. and Takahata, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. USA **92**: 532-536.

Hsu, S.-L., Marks, J., Shaw, J.-P., Tam, M., Higgs, D.R., Shen, C.C. and Shen, C.-K.J. (1988). Structure and expression of the human θ1 globin gene. Nature **331**: 94-96.

Hubby, J.L. and Lewontin, R.C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. Genetics **54**: 577-594.

Hubert, R., MacDonald, M., Gusella, J. and Arnheim, N. (1994). High resolution localization of recombination hot spots using sperm typing. Nature Genet. **7**: 420-424.

Hudson, R.R. (1990). Gene genealogies and the coalescent process. In: Oxford Surveys in Evolutionary Biology, Futuyma, D. and Antonovics, J. (eds.). Oxford; Oxford University Press.

Hudson, R.R., Kreitman, M. and Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics **116**: 153-159.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. Proc. Roy. Soc. Lond. (Biol.) **256**: 119-124.

Ikemura, T. and Aota, S.-I. (1988). Global variation in G+C content along vertebrate genome DNA. Possible correlation with chromosome band structures. J. Mol. Biol. **203**: 1-13.

Innis, M.A. and Gelfand, D.H. (1990). Optimization of PCRs. In: PCR Protocols. A Guide to Methods and Applications, Innis, M.A., Gelfand, D.H., Sninsky, J.J. and White, T.J. (eds.).

Innis, M.A., Gelfand, D.H., Sninsky, J.J. and White, T.J. (1990). PCR Protocols. A Guide to Methods and Applications. San Diego; Academic Press INC.

Jarman, A.P. and Higgs, D.R. (1988). A New Hypervariable Marker for the Human α-Globin Gene Cluster. Am. J. Hum. Genet. **43**: 249-256.

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

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

Jeffreys, A.J. (1979). DNA sequence variants in the $^{G}\gamma$, $^{A}\gamma$, δ and β-globin genes of man. Cell **18**: 1-10.

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

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

Jeffreys, A.J., Tamaka, K., MacLeod, A., Monkton, D.G., Neil, D.L. and Armour, J.A.L. (1994). Complex gene conversion events in germline mutation at human minisatellites. Nature Genet. **6**: 136-145.

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

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

Jeffreys, A.J., Wilson, V., Neumann, R. and Keyte, J.(1988b). Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. Nucl. Acids Res. **23**: 10953-10971.

Jobling, M.A. and Tyler-Smith, C. (1995). Fathers and sons: the Y chromosome and human evolution. TIG **11**: 449-456.

Jones, J.S. and Rouhani, S. (1986). How small was the bottleneck? Nature **319**: 449-450.

Jorde, L.B. (1995). Linkage diseqililbrium as a gene-mapping tool. Am. J. Hum. Genet. **56**: 11-14.

Jorde, L.B., Banshad, M.J., Watkins, W.S., Zenger, R., Fraley, A.E., Krakowiak, P.A., Carpenter, K.D., Soodyall, H., Jenkins, T. and Rogers, A.R. (1995). Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. **57**: 523-538.

Jurka, J. and Smith, T. (1988). A fundamental division in the *Alu* family of repeated sequences. Proc. Natl. Acad. Sci. USA **85**: 4775-4778.

Jurka, J., Smith, T.F. and Labuda, D. (1988). Small cytoplasmic Ro pseudogene and an *Alu* repeat in the human α1 globin gene. Nucl. Acids Res. **16**: 766.

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

Kaplan, N.L., Hudson, R.R. and Langley, C.H. (1989). The "hitchhiking effect" revisited. Genetics **123**: 887-899.

Kazazian, H.H., Chakravarti, A., Orkin, S.H. and Antonarakis, S.E. (1983). DNA polymorphisms in the human β-globin gene cluster. In: Evolution of Genes and Proteins, Nei, M. and Koehn, R.K. (eds.). Sunderland, MA; Sinauer.

Ki-Zerbo, J. (1990). I. Methodology and African Prehistory. General History of Africa. UNESCO.

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

Kimura, M. and Ohta, T. (1978). Stepwise mutation model and distribution of allele frequencies in a finite population. Proc. Natl. Acad. Sci. USA **75**: 2868-2872.

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

Kreitman, M. (1987). Molecular population genetics. Oxford Surveys of Evolutionary Biology **4**: 38-60.

Krontiris, T.G. (1995). Minisatellites and human disease. Science **269**: 1682-1683.

Kunkel, T. (1993). Slippery DNA and diseases. Nature **365**: 207-208.

Kwiatkowski, D.J., Dib, C., Slaugenhaupt, S.A., Povey, S., Gusella, J.F. and Haines, J.L. (1993). An index marker map of chromosome 9 provides strong evidence for positive interference. Am. J. Hum. Genet. **53**: 1279-1288.

Lahr, M.M. and Foley, R. (1994). Multiple dispersals and modern human origins. Evolutionary Anthropology **3**: 48-60.

Lauer, J., Shen, C.-K.J. and Maniatis, T. (1980). The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. Cell **20**: 119-130.

Laurie, C.C. and Stam, L.F. (1994). The effect of an intronic polymorphism on alcohol dehydrogenase expression in *Drosophila melanogaster*. Genetics **138**: 379-385.

Leung, S.-O., Proudfoot, N.J. and Whitelaw, E. (1987). The gene for θ-globin is transcribed in human fetal erythroid tissues. Nature **329**: 551-554.

Levinson, G. and Gutman, G.A. (1987). Slipped-strand mispairing: a major mechansim for DNA sequence evolution. Mol. Biol. Evol. **4**: 203-221.

Lewontin, R.C. (1972). The apportionment of human diversity. Evolutionary Biology **6**: 381-398.

Lewontin, R.C. (1991). Twenty-Five years ago in genetics: Electrophoresis in the development of evolutionary genetics: Milestone or Millstone? Genetics **128**: 657-662.

Lewontin, R.C. and Hubby, J.L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations in *Drosophila pseudoobscura*. Genetics **54**: 595-609.

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

Li, W.-H. and Graur, D. (1991). Fundamentals of Molecular Evolution. Sunderland, Massachusetts; Sinauer Associates, Inc.

Li, W.-H. and Sadler, L.A. (1991). Low nucleotide diversity in man. Genetics **129**: 513-523.

Lichten, M. and Goldman, A.S.H. (1995). Meiotic recombination hotspots. Annu. Rev. Genetics **29**: 423-444.

Liebhaber, S.A. and Begley, K.A. (1983). Structural and evolutionary analysis of the two chimpanzee α-globin mRNAs. Nucleic Acids Rsearch **11**: 8915-8929.

Liebhaber, S.A., Rappaport, E.F., Cash, F.E., Ballas, S.K., Schwartz, E. and Surrey, S. (1984). Hemoglobin I mutation encoded at both α-globin loci on the same chromosome: Concerted evolution in the human genome. Science **226**: 1449-.

Liu, J., Wu, T.-C. and Lichten, M. (1995). The location and structure of double-strand DNA breaks induced during yeast meiosis: evidence for a covalently linked DNA-protein intermediate. EMBO J. **14**: 4599-4608.

Lu, Y.H. and Negre, S. (1993). Use of glycerol for enhanced efficiency and specificity of PCR amplification. TIG **9**: 297.

Lucassen, A.M., Julier, C., Beresi, J.-P., Boitard, C., Froguel, P., Lathrop, M. and Bell, J.I. (1993). Susceptibility to insulin dependent diabetes mellitus maps to a 4.1kb segment of DNA spanning the insulin gene and associated VNTR. Nature Genet. **4**: 305-310.

Lucassen, A.M., Screaton, G.R., Julier, C., Elliot, T.J., Lathrop, M. and Bell, J.I. (1995). Regulation of insulin gene expression by the IDDM associated, insulin locus haplotype. Hum. Mol. Genet. **4**: 501-506.

Maeda, N., Bliska, J.B. and Smithies, O. (1983). Recombination and balanced chromosome polymorphisms suggested by DNA sequences 5' to the human δ-globin gene. Proc. Natl. Acad. Sci. USA **80**: 5012-5016.

Malone, R.E., Kim, S., Bullard, S.A., Lundquist, S., Hutchings-Crow, L., Cramton, S., Lutfiyya, L. and Lee, J. (1994). Analysis of a recombination hotspot for gene conversion occurring at the HIS2 gene of *Saccharomyces cerevisiae*. Genetics **137**: 5-18.

Marks, J., Shaw, J.-P. and Shen, C.-K.J. (1986). Sequence organization and genomic complexity of primate θ1 globin gene, a novel α-globin-like gene. Nature **321**: 785-788.

Martin, M., Mann, D. and Carrington, M. (1995). Recombination rates across the HLA complex: use of microsatellites as a rapid screen for recombinant chromosomes. Hum. Mol. Genet. **4**: 423-428.

Martinson, J.J. (1991). Genetic Variation in South Pacific Islanders. D.Phil. Thesis. Oxford University.

Martinson, J.J., Boyce, A.J. and Clegg, J.B. (1994). VNTR alleles associated with the α-globin locus are haplotype and population related. Am. J. Hum. Genet. **55**: 513-525.

Martinson, J.J., Excoffier, L., Swinburn, C., Boyce, A.J., Harding, R.M., Langaney, A. and Clegg, J.B. (1995). High diversity of α-globin haplotypes in a Senegalese population, including many previously-unreported variants. Am. J. Hum. Genet. **57**: 1186-1198.

Martinson, J.J., Harding, R.M., Philippon, G., Saint-Marie, F.F., Roux, J., Boyce, A.J. and Clegg, J.B. (1993). Demographic reductions and genetic bottlenecks in humans: minisatellite allele distributions in Oceania. Hum. Genet. **91**: 445-450.

Massy, B.D., Rocco, V. and Nicolas, A. (1995). The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in *Saccharomyces cerevisiae*. EMBO J. **14**: 4589-4598.

Maynard-Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genetical Research **23**: 23-35.

Mayr, E. (1963). Animal Species and Evolution. Cambridge, MA; Harvard University Press.

McDonald, J.H. and Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351**: 652-654.

McGinnis, R.E. and Spielman, R.S. (1994). Linkage Disequilibrium in the Insulin Gene Region: Size Variation at the 5' Flanking Polymorphism and Bimodality among "Class 1" Alleles. Am. J. Hum. Genet. **55**: 526-532.

Meselson, M.S. and Radding, C.M. (1975). A general model for genetic recombination. Proc. Natl. Acad. Sci. USA **72**: 358-361.

Michelson, A.M. and Orkin, S. (1983). Boundaries of gene conversion within the duplicated human α-globin genes. J. Biol. Chem. **258**: 15245-15454.

Mokhtar, G. (1990). II. Ancient Civilizations of Africa. General History of Africa. UNESCO.

Monckton, D.G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A. and Jeffreys, A.J. (1994). Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. Nature Genet. **8**: 162-169.

Moriyama, E.N. and Powell, J.R. (1996). Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13**: 261-277.

Murti, J.R., Bumbulis, M. and Schimenti, J.C. (1992). High-frequency germ line gene conversion in transgenic mice. Mol. Cell. Biol. **12**: 2545-2552.

Nagel, R.L. and Ranney, H.M. (1990). Genetic epidemiology of structural mutations of the β-globin gene. Sem. Hematol. **27**: 342-359.

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

Nei, M. (1987). Molecular Evolutionary Genetics. New York; Columbia University Press.

Nei, M. and Roychoudhury, A.K. (1982). Genetic relationship and evolution of human races. Evolutionary Biology **14**: 1-59.

Nei, M. and Roychoudhury, A.K. (1993). Evolutionary Relationships of Human Populations on a Global Scale. Mol. Biol. Evol. **10**: 927-943.

Nei, M. and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. Genetics **97**: 145-163.

Nei, M. and Takezaki, N. (1996). The root of the phylogenetic tree of human populations. Mol. Biol. Evol. **13**: 170-177.

Newton, C.R., Graham, A., Heptinstall, L.E., Powell, S.J., Summers, C., Kalsheker, N., Smith, J.C. and Markham, A.F. (1989). Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucl. Acids Res. **17**: 2503-2516.

Nicholls, R.D., Fischel-Ghodsian, N. and Higgs, D.R. (1987). Recombination at the human α-globin gene cluster: sequence features and topological constraints. Cell **49**: 369-378.

Nicolas, A., Treco, D., Scultes, N.P. and Szostak, J.W. (1989). An initiation site for meiotic gene conversion in the yeast *Saccharomyces cerevisiae*. Nature **338**: 35-39.

Ohno, S. (1970). Evolution by gene duplication. New York; Springer-Verlag.

Ohta, T. (1995). Gene conversion vs point mutation in generating variability at the antigen recognition site of major histocompatibility complex loci. J. Mol. Evol. **41**: 115-119.

Ojwang, P.J., Ogada, T., Beris, P., Hattori, Y., Lanclos, K.D., Kutlar, A., Kutlar, F. and Huisman, T.H.J. (1987). Haplotypes and α–globin gene anlyses in sickle cell anaemia patients from Kenya. Brit. J. Haemat. **65**: 211-215.

Olerup, O., Troye-Blomberg, M., Schreuider, G.M.T. and Riley, E.M. (1988). HLA-DR and -DQ gene polymorphism in West Africans is twice as extensive as in North European Caucasions: evolutionary implications. Proc. Natl. Acad. Sci. USA **88**: 8480-8484.

Oliver, R. (1991). The African Experience. Weidenfeld and Nicolson.

Orita, M., Sekiya, T. and Hayashi, K. (1990). DNA sequence polymorphisms in *Alu* repeats. Genomics **8**: 271-278.

Orkin, S.H. and Goff, S.C. (1981). The duplicated human α-globin genes: their relative expression as measured by RNA analysis. Cell **24**: 345-351.

O'Shaughnessy, D.F., Hill, A.V.S., Bowden, D.K., Weatherall, D.J. and Clegg, J.B. (1990). Globin genes in Micronesia: origins and affinities of Pacific island peoples. Am. J. Hum. Genet. **46**: 144-155.

Paldi, A., Gyapay, G. and Jami, J. (1995). Imprinted chromosomal regions of the human genome display sex-specific meiotic recombination frequencies. Curr. Biol. **5**: 1030-1035.

Peacock, A.C. and Dingman, C.W. (1968). Molecular weight estimation and separation of ribonucleic acid by electrophoresis in agarose-acrylamide composite gels. Biochemistry **7**: 668-674.

Perutz, M.F. (1990). Frequency of abnormal human haemoglobins caused by C to T transitions in CpG dinucleotides. J. Mol. Biol. **213**: 203-206.

Peterson, A.C., DiRienzo, A., Lehesjoki, A.-E., Chapelle, A.D.L., Slatkin, M. and Freimer, N.B. (1995). The distribution of linkage disequilibrium over anonymous genome regions. Hum. Mol. Genet. **4**: 887-894.

Petes, T.D., Malone, R.E. and Symington, L.S. (1991). Recombination in Yeast. In: The Molecular and Cellular Biology of the Yeast *Saccharomyces*: Genome Dynamics, Protein synthesis and energetics., (eds.). Cold Spring Harbor Laboratory Press.

Poloni, E.S., Excoffier, L., Mountain, J.L., Langaney, A. and Cavalli-Sforza, L.L. (1995). Nuclear DNA polymorphism in a Mandenka population from Senegal: comparison with eight other human populations. Ann. Hum. Genet. (Lond). **59**: 43-61.

Ponticelli, A. and Smith, G.R. (1992). Chromosomal context dependence of a eukaryotic recombinational hot spot. Proc. Natl. Acad. Sci. USA **89**: 227-231.

Powers, P., Altay, C., Huisman, T.H.J. and Smithies, O. (1984). Two novel arrangements of human fetal globin genes: $^{G}\gamma$-$^{G}\gamma$ and $^{A}\gamma$-$^{A}\gamma$. Nucl. Acids Res. **12**: 7023-7034.

Powers, P.A. and Smithies, O. (1986). Short gene conversions in the human fetal globin gene region: a by-product of chromosome pairing during meiosis? Genetics **112**: 343-358.

Pressley, L., Higgs, D.R., Clegg, J.B. and Weatherall, D.J. (1980). Gene deletions in $\alpha$-thalassemia prove that the 5' $\zeta$ locus is functional. Proc. Natl. Acad. Sci. USA **77**: 3586-3589.

Proudfoot, N.J., Rutherford, T.R. and Partington, G.A. (1984). Transcriptional analysis of human $\zeta$-globin genes. EMBO J. **3**: 1533-1540.

Proudfoot, N.J.A., Gil, A. and Maniatis, T. (1982). The structure of the human $\zeta$-globin gene and a closely linked, nearly identical, pseudogene. Cell **31**: 553-563.

Ramsay, M. and Jenkins, T. (1988). $\alpha$-Globin gene cluster haplotypes in the Kalahari San and southern African Bantu-speaking Blacks. Am. J. Hum. Genet. **43**: 527-533.

Renges, H.H., Peacock, R., Dunning, A.M., Talmud, P. and Humphries, S.E. (1992). Genetic relationship between the 3'HVR and diallelic apolipoprotein B gene polymorphisms. Ann. Hum. Genet. (Lond). **56**: 11-13.

Richards, R.I. and Sutherland, G.R. (1994). Simple repeat DNA is not replicated simply. Nature Genet. **6**: 114-116.

Ristaldi, M.S., Casula, S., Rando, A. and Vestri, R. (1995). Sheep $\alpha$-globin gene sequences: implications for their concerted evolution and for the down-regulation of the 3' genes. J. Mol. Evol. **40**: 349-353.

Roberts-Thomson, J., Martinson, J.J., Norwich, J., Harding, R., Boecher, B. and Clegg, J.B. (in press). An ancient common origin of Aboriginal Australians and New Guinea highlanders is supported by $\alpha$-globin haplotype analysis. Am. J. Hum. Genet. **58**.

Rogers, A.R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9**: 552-569.

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

Ruano, G., Kidd, K.K. and Stephens, J.C. (1990). Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. Proc. Natl. Acad. Sci. USA **87**: 6296-6300.

Rudiger, N.S., Gregersen, N. and Kielland-Brandt, M.C. (1995). One short well conserved region of *Alu*-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. Nucl. Acids Res. **23**: 256-260.

Ruvolo, M. (1996). A new approach to studying modern human origins: hypothesis testing with coalescence time distributions. Molecular Phylogenetics and Evolution **5**: 202-219.

Ruvolo, M., Zehr, S., Dornum, M.V., Pan, D., Chang, B. and Lin, J. (1993). Mitochondrial COII sequences and modern human origins. Mol. Biol. Evol. **10**: 1115-1135.

Saccone, S., Sario, A.D., Valle, G.D. and Bernardi, G. (1992). The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. Proc. Natl. Acad. Sci. USA **89**: 4913-4917.

Saiki, R.K. (1989). The Design and Optimization of the PCR. In: PCR Protocols. A Guide to Methods and Applications, Innis, M.A., Gelfand, D.H., Sninsky, J.J. and White, T.J. (eds.).

Saiki, R.K. (1990). Amplification of Genomic DNA. In: PCR Protocols. A Guide to Methods and Applications, Innis, M.A., Gelfand, D.H., Sninsky, J.J. and White, T.J. (eds.). San Diego; Academic Press INC.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988). Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. Science **239**: 487-491.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985). Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anaemia. Science **230**: 1350-1354.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). Molecular Cloning. A laboratory manual. Cold Spring Harbor Laboratory Press.

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

Sarich, V.M. and Wilson, A.C. (1967). Immunological time scale for hominid evolution. Science **158**: 1200-1203.

Savatier, P., Trabuchet, G., Faure, C., Chelboune, Y., Gouy, M., Verdier, G. and Nigon, V.M. (1985). Evolution of the primate β-globin gene region. High rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee. J. Mol. Biol. **182**: 21-29.

Sawada, I. and Schmid, C.W. (1986). Primate evolution of the α-globin gene cluster and its *Alu*-like repeats. J. Mol. Biol. **192**: 693-709.

Sawada, I., Willard, C., Shen, C.-K.J., Chapman, B., Wilson, A. and Schmid, C.W. (1985). Evolution of *Alu* Family Repeats since the divergence of human and chimpanzee. J. Mol. Evol. **22**: 316-322.

Schaeffer, S.W. and Miller, E.L. (1993). Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehyrogenase region of *Drosophila pseudoobscura*. Genetics **135**: 541-552.

Schon, E.A., Wernke, S.M. and Lingrel, J.B. (1982). Gene conversion of two functional goat α-globin genes preserves only minimal flanking sequences. J. Biol. Chem. **257**: 6825-6835.

Schultes, N.P. and Szostak, J.W. (1990). Decreasing gradients of gene conversion on both sides of the initiation site for meiotic recombination at the ARG4 locus in yeast. Genetics **126**: 813-822.

Schwacha, A. and Kleckner, N. (1995). Identification of double Holliday Junctions as intermediates in meiotic recombination. Cell **83**: 783-791.

Schwarz, K., Hansen-Hagge, T. and Bartram, C. (1990). Improved yields of long PCR products using gene 32 protein. Nucl. Acids Res. **18**: 1079.

Sclotterer, C. and Tautz, D. (1992). Slippage synthesis of simple sequence DNA. Nucl. Acids Res. **20**: 211-215.

Sharpe, J.A., Chan-Thomas, P.S., Lida, J., Ayyub, H., Wood, W.G. and Higgs, D.R. (1992). Analysis of the human α-globin upstream regulatory element (HS-40) in transgenic mice. EMBO J. **11**: 4565-4572.

Shaw, J.-P., Marks, J. and Shen, C.-K.J. (1987). Evidence that the recently discovered θ1 globin gene is functional in higher primates. Nature **326**: 717-720.

Shen, W.-H. and Hohn, B. (1992). DMSO improves PCR amplification of DNA with complex secondary structure. TIG **8**: 227.

Sibley, C.G. and Ahlquist, J.E. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J. Mol. Evol. **20**: 2-15.

Slightom, J.L., Blechl, A.E. and Smithies, O. (1980). Human foetal $^{A}\gamma$ and $^{G}\gamma$ globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated gene. Cell **21**: 627-638.

Siniscalco, M., Bernini, L., Filippi, G., Latte, B., Khan, P.M., Piomelli, S. and Rattazzi, M. (1966). Population genetics of haemoglobin variants, thalassaemia and glucose-6-phosphate dehydrogenase deficiency, with particular reference to the malaria hypothesis. Bull. WHO **34**: 379-393.

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

Smith, G.R. and Stahl, F.W. (1984). Homologous recombination promoted by chi sites and RecBC Enzyme of *Escherichia coli*. BioEssays **2**: 224-249.

Smithies, O. and Powers, P.A. (1986). Gene conversions and their relation to homologous chromosome pairing. Phil. Trans. R. Soc. Lond. B. **312**: 291-302.

Spurdle, A.B., Hammer, M.F. and Jenkins, T. (1994). The Y *Alu* polymorphism in Southern African populations and its relationship to other Y-specific polymorphisms. Am. J. Hum. Genet. **54**: 319-330.

Stahl, F.W. (1994). The Holliday junction on its thirtieth anniversary. Genetics **138**: 241-246.

Steen, V.M., Molven, A., Aarskog, N.K. and Gulbrandsen, A.-K. (1995). Homologous unequal cross-over involving a 2.8 kb direct repeat as a mechanism for the generation of allelic variants of the human cytochrome P450 CYP2D6 gene. Hum. Mol. Genet. **4**: 2251-2257.

Steinberger, E.M., Thompson, L.D. and Hartmann, J.M. (1993). On the use of excess homozygosity for subpopulation detection. Am. J. Hum. Genet. **52**: 1275-1277.

Steinmetz, M., Stephan, D. and Lindahl, K.F. (1986). Gene organization and recombinational hotspots in the murine major histocompatibility complex. Cell **44**: 895-904.

Stoeckert, C.J., Collins, F.S. and Weissman, S.M. (1984). Human foetal globin DNA sequences suggest novel conversion event. Nucl. Acids Res. **12**: 4469-4479.

Stoltzfus, A. and Doolittle, W.F. (1993). Slippery introns and globin gene evolution. Curr. Biol. **3**: 215-217.

Strachan, T. (1992). The Human Genome. Oxford; BIOS Scientific Publishers.

Strand, M., Prolla, T.A., Liskay, R.M. and Petes, T.D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature **365**: 274-276.

Stringer, C.B. and Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. Science **239**: 1263-1268.

Sun, H., Treco, D., Schultes, N.P. and Szostak, J.W. (1989). Double-strand breaks at an initiation site for meiotic gene conversion. Nature **338**: 87-90.

Sun, H., Treco, D. and Szostak, J.W. (1991). Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-stand breaks at the ARG4 recombination initiation site. Cell **64**: 1155-1161.

Swisher, C.C., Curtis, G.H., Jacob, T., Getty, A.H., Suprijo, A. and Widiasmoro. (1994). Age of the earliest known hominids in Java, Indonesia. Science **263**: 1118-1121.

Szostak, J.W., Orr-Weaver, T.L., Rothstein, R.J. and Stahl, F.W. (1983). The double-strand-break repair model for recombination. Cell **33**: 25-35.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437-460.

Takahata, N. (1993). Allelic genealogy and human evolution. Mol. Biol. Evol. **10**: 2-22.

Takahata, N. (1994). Repeated failures that led to the eventual success in human evolution. Mol. Biol. Evol. **11**: 803-805.

Templeton, A.R. (1993). The "Eve" hypothesis: a genetic critique and reanalysis. American Anthropologist **95**: 51-72.

Thein, S.L. and Hinton, J. (1991). A simple and rapid method of direct sequencing using Dynabeads. Brit. J. Haemat. **79**: 113-115.

Thomas, B.J. and Rothstein, R. (1989). Elevated recombination rates in transcriptionally active DNA. Cell **56**: 619-630.

Thomas, G.R., Roberts, E.A., Walshe, J.M. and Cox, D.W. (1995). Haplotypes and mutations in Wilson Disease. Am. J. Hum. Genet. **56**: 1315-1319.

Thomas, M.G., Miller, K.W.P., Cook, C.E. and Hagelberg, E. (1994). A method for avoiding mis-priming when sequencing with Dynabeads. Nucl. Acids Res. **22**: 3243-3244.

Tiemei, C., Quan, Y. and En, W. (1994). Antiquity of *Homo sapiens* in China. Nature **368**: 55-56.

Treco, D., Thomas, B. and Arnheim, N. (1985). Recombination hot spot in the human β-globin gene cluster: meiotic recombination of human DNA fragments in Saccharomyces cerevisiae. Mol. Cell. Biol. **5**: 2029-2038.

Trent, R.J., Buchanan, J.G., Webb, A., Goundar, R.P.S., Seruvatu, L.M. and Mickleson, K.N.P. (1988). Globin Genes are useful markers to identify genetic similarities between Fijians and Pacific Islanders from Polynesia and Melanesia. Am. J. Hum. Genet. **42**: 601-607.

Ullu, E. and Tschudi, C. (1984). *Alu* sequences are processed 7SL RNA genes. Nature **312**: 171-172.

Valdes, A.M., Slatkin, M. and Freimer, N.B. (1992). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133**: 737-749.

Varadaraj, K. and Skinner, D.M. (1994). Denaturants or cosolvents improve the specificty of PCR amplification of a G+C rich DNA using genetically engineered DNA polymerases. Gene **140**: 1-5.

Vyas, P., Vickers, M.A., Simmons, D.L., Ayyub, H., Craddock, C.F. and Higgs, D.R. (1992). Cis-acting sequences regulating expression of the human α-globin cluster lie within constitutively open chromatin. Cell **69**: 781-793.

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

Wahls, W.P., Wallace, L.J. and Moore, P.D. (1990b). The Z-DNA motif $d(TG)_{30}$ promotes reception of information during gene conversion events while stimulating homolgous recombination in human cells in culture. Mol. Cell. Biol. **10**: 785-793.

Wainscoat, J.S., Hill, A.V.S., Boyce, A.L., Flint, J., Hernandez, M., Thein, S.L., Old, J.M., Lynch, J.R., Falusi, A.G., Weatherall, D.J. and Clegg, J.B. (1986). Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature **319**: 491-493.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theoretical and Population Biology **7**: 256-276.

Watterson, G.A. (1978). The homozygosity test of neutrality. Genetics **88**: 405-417.

Weatherall, D.J. (1991). The New Genetics and Clinical Practice. New York; Oxford University Press.

Weiner, A.M., Deiningner, P.L. and Efstratiadis, A. (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. **55**: 631-661.

Weissenbach, J., Gynapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. and Lathrop, M. (1992). A second-generation linkage map of the human genome. Nature **359**: 794-801.

Weller, P., Jeffreys, A.J., Wilson, V. and Blanchetot, A. (1984). Organization of the human myoglobin gene. EMBO J. **3**: 439-446.

Willard, C., Wong, E., Hess, J.F., Shen, C.-K.J., Chapman, B., Wilson, A.C. and Schmid, C.W. (1986). Comparison of human and chimpanzee ζ1 globin genes. J. Mol. Evol. **22**: 49-.

Wolfe, K.H., Sharp, P.M. and Li, W.-H. (1989). Mutation rates differ among regions of the mammalian genome. Nature **337**: 283-285.

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

Wolff, R.K., Plaetke, R., Jeffreys, A.J. and White, R. (1989). Unequal crossing-over between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. Genomics **5**: 382-384.

Wolpoff, M.H., Xinzhi, W. and Thorne, A.G. (1984). Modern *Homo sapiens* origins: a general theory of hominid evolution involving the fossil evidence from East Asia. In: The Origins of Modern Humans: A World Survey of the Fossil Evidence, Smith, F.H. and Spencer, F. (eds.). New York; A R Liss.

Wu, D.Y., Ugozzoli, L., Pal, B.K. and Wallace, R.B. (1989). Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. Proc. Natl. Acad. Sci. USA **86**: 2757-2760.

Wu, T.-C. and Lichten, M. (1994). Meiosis-induced double-strand break sites determined by yeast chromatin structure. Science **263**: 515-518.

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

Xiong, W., Li, W.-H., Posber, I., Yamamura, T., Yamaoto, A., Gotto, A.M. and Chan, L. (1991). No severe bottleneck during human evolution: evidence from two apolipoprotein C-II deficiency alleles. Am. J. Hum. Genet. **48**: 383-389.

Zago, M.A., Santos, E.J.M., Clegg, J.B., Guerreiro, J.F., Martinson, J.J., Norwich, J. and Figueiredo, M.S. (1995). α-Globin gene haplotypes in South American Indians. Human Biology **67**: 535-546.

Zahn-Zabal, M., Lehmann, E. and Kogli, J. (1995). Hot spots of recombination in fission yeast: inactivation of the M26 hot spot by deletion of the *ade6* promoter and the novel hotspot *ura4*-aim. Genetics **140**: 469-478.

Zangenberg, G., Huang, M.-M., Arnheim, N. and Erlich, H. (1995). New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. Nature Genet. **10**: 407-414.

Zimmer, E.A., Martin, S.L., Beverley, S.M., Kan, Y.W. and Wilson, A.C. (1980). Rapid duplication and loss of genes coding for the α-chains of haemoglobin. Proc. Natl. Acad. Sci. USA **77**: 2158-2162.