GLOBIN GENES OF THE SOUTH AFRICAN INDIAN POPULATION

AMANDA KRAUSE

A dissertation submitted to the Faculty of Medicine, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy

Johannesburg 1994

ABSTRACT

Most South Africans with clinically significant haemoglobinopathies, mainly thalassaemia major, are of Asian Indian origin. This study was undertaken to characterise the normal and pathological variation around the α and ß globin clusters, both at the haematological and molecular level in the major religious and language subgroups of South African Indians. These data were used to assess the need for a screening programme for the major haemoglobinopathies and to evaluate the local prenatal diagnostic service. The variation was also used to assess the genetic relationships between the subgroups of South African Indians, to compare them to Indians studied elsewhere and to other populations.

The subjects of this study were all South African Indians. They consisted of 638 individuals collected in a school survey, 60 normal nuclear families, 31 families with at least one child with a major haemoglobinopathy, four families 'at-risk' for having a child with thalassaemia major, and seven families, with at least one abnormal globin gene, ascertained through our genetic counselling clinic. Families were studied using standard haematological techniques, as well as with the DNA-based techniques of 'Southern blotting' and the polymerase chain reaction.

Microcytosis and hypochromia are extremely common in South African Indians. This is mainly due to a high frequency of the -α chromosome, but also to a high rate of iron deficiency. Although heterozygote rates for ß thalassaemia are relatively low, about 1-2% in the major groups, it appears that a greater number of thalassaemia major patients are ascertained, perhaps because of high rates of consanguinity and intra-caste marriage or pockets of higher frequency in the minor groups, like the Moslem Memon.

At the molecular level, α thalassaemia in South African Indians is mainly of the $-\alpha^{3.7\text{II}}$ type, though the $-\alpha^{3.7\text{III}}$ and $-\alpha^{4.2}$ chromosomes are present, the latter at relatively high frequency in the Hindu Hindi. The -- chromosome was not detected, and one family with a non-deletion HbH disease was ascertained. HbH and Hb Bart's hydrops fetalis are thus rare in this population. ß thalassaemia is heterogeneous, with 10 mutations described on 67 chromosomes. The mutation at IVS1nt5 is the commonest in all the major groups.

Together with the mutations at codon 41/42 and IVS1nt1 and the 619bp deletion, it accounts for 84% of all chromosomes. Two mutations not previously described in Asian Indians, frameshift codon 44(-C) and Poly-A (T→C), were found in this cohort.

The costs of screening the South African Indian population for ß thalassaemia, and the other ß-globin variants, is low compared to the costs of treatment for the existing patients. A culturally-appropriate screening programme targeted at young couples of child-bearing age should be initiated as soon as possible. Detection of raised HbA_2 levels, using haemoglobin electrophoresis, appears to be the most cost-effective primary screen.

Whereas linked marker analysis only provided a full diagnosis in 64.5% of families, the availability of mutation analysis, together with linked marker analysis, means that close to 100% of families can now be offered a full prenatal diagnosis. At present, couples requesting prenatal diagnosis are still predominantly those with a previously affected child, although there is an increasing number of couples shown to be 'at-risk' on haematological screening. For the latter mutation analysis is particularly useful. The availability of CVS has increased the acceptability of the prenatal diagnostic service, although it is still not being used by all 'at-risk' couples.

Studies of normal variation in the α and ß globin clusters, including gene rearrangements, RFLP and haplotype frequencies, show that South African Indians share features with Indians studied elsewhere. A number of the religious and language subgroups, particularly the Hindu Gujerati and Hindu Hindi, have distinctive features due to geographical isolation and genetic drift. The Indians, although predominantly Caucasoid, also have Mongoloid markers, reflecting common ancestry as well as more recent gene flow, through political, economic and military interaction. The geographical distributions of the ß thalassaemia mutations, together with studies of the associated haplotypes, provide evidence for a limited number of origins prior to population divergence, with subsequent limited gene conversion and recombination events leading to the observed haplotype diversity.

DECLARATION

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

I declare that this work has been approved by the Ethics Committee for Research on Human Subjects of the University of the Witwatersrand. The certificate number is 6/8/87.

AMANDA KRAUSE

29th day of January 1994

This work is dedicated to my parents for their encouragement throughout my academic career, and to my husband, Mark Heilbrunn, for his endless support and patience.

ACKNOWLEDGEMENTS

I am grateful to the many individuals who donated blood samples without whom this study would have been impossible. I am also indebted to Professor A Wadee and Mr D Dunn who collected many of the samples in the Lenasia school survey and assisted in follow-up collection of repeat samples; to Dr S Candy who assisted me with follow-up and genetic counselling of individuals with haemoglobinopathies detected during the Lenasia school survey; to Dr H Roode and Dr J Gorvy who allowed me to attend their thalassaemia clinics and make contact with the haemoglobinopathy patients and their families; to Sister Namitha Chabilal who facilitated contact with thalassaemia patients in Natal; and to Dr Yusuf Patel, Dr Himla Soodyall, Linda Pounasamoy and Ashok Boodhoo whose personal contacts facilitated ascertainment and collection of blood from random healthy Indian families.

I also wish to thank Dr JS Wainscoat and Dr J Old, John Radcliffe Institute, Oxford for supplying us with the ß globin cluster probes, Dr J Old for the oligonucleotide sequences for analysis of the ß globin cluster RFLPs and ß thalassaemia mutations using PCR; and Dr DR Higgs, John Radcliffe Institute, Oxford who supplied us with the α globin cluster probes. The research detailed in this thesis has also benefited from generous personal communications with these colleagues. Further Dr J Old has assisted us in a number of difficult diagnostic cases.

My special thanks to the following:

- My colleagues in the Department of Human Genetics for their support, friendship and helpful discussion, especially to Dr Michele Ramsay, Dr Amanda Spurdle and Professor Jennifer Kromberg.
- The individuals in the Department of Haematology at the SAIMR who allowed me ready access to their Coulter Counter and haemoglobin electrophoresis equipment, so that I could complete the haematological part of this study.
-The Department of Chemical Pathology, SAIMR, where all the iron studies were carried out.
-The SAIMR Photographic Unit who processed the photographic plates in this thesis.

ACKNOWLEDGEMENTS

I am grateful to the many individuals who donated blood samples without whom this study would have been impossible. I am also indebted to Professor A Wadee and Mr D Dunn who collected many of the samples in the Lenasia school survey and assisted in follow-up collection of repeat samples; to Dr S Candy who assisted me with follow-up and genetic counselling of individuals with haemoglobinopathies detected during the Lenasia school survey; to Dr H Roode and Dr J Gorvy who allowed me to attend their thalassaemia clinics and make contact with the haemoglobinopathy patients and their families; to Sister Namitha Chabilal who facilitated contact with thalassaemia patients in Natal; and to Dr Yusuf Patel, Dr Himla Soodyall, Linda Pounasamoy and Ashok Boodhoo whose personal contacts facilitated ascertainment and collection of blood from random healthy Indian families.

I also wish to thank Dr JS Wainscoat and Dr J Old, John Radcliffe Institute, Oxford for supplying us with the ß globin cluster probes, Dr J Old for the oligonucleotide sequences for analysis of the ß globin cluster RFLPs and ß thalassaemia mutations using PCR; and Dr DR Higgs, John Radcliffe Institute, Oxford who supplied us with the α globin cluster probes. The research detailed in this thesis has also benefited from generous personal communications with these colleagues. Further Dr J Old has assisted us in a number of difficult diagnostic cases.

My special thanks to the following:

- My colleagues in the Department of Human Genetics for their support, friendship and helpful discussion, especially to Dr Michele Ramsay, Dr Amanda Spurdle and Professor Jennifer Kromberg.
- The individuals in the Department of Haematology at the SAIMR who allowed me ready access to their Coulter Counter and haemoglobin electrophoresis equipment, so that I could complete the haematological part of this study.
-The Department of Chemical Pathology, SAIMR, where all the iron studies were carried out.
-The SAIMR Photographic Unit who processed the photographic plates in this thesis.

-My husband, Mark Heilbrunn, for his help with the computer, as well as for many years of patience, understanding and support.
-My mother for many hours spent helping me check data and preparing my thesis.

I gratefully acknowledge the financial support which I have received during my studies, in particular from the MRC, the University of the Witwatersrand, the Freda Lawenski Trust, and the Stella and Paul Loewenstein Charitable and Educational Trust. My appreciation also to the Director of the SAIMR who made facilities available in the Department of Human Genetics for the research described in this thesis.

I would like to express special thanks and appreciation to my supervisor Professor Trefor Jenkins for his guidance, support and patience throughout this study.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF FIGURES

CHAPTER 6

CHAPTER 7

ABBREVIATIONS AND SYMBOLS

α	Alpha
ß	Beta
γ	Gamma
δ	Delta
ϵ	Epsilon
ζ	Zeta
θ	Theta
λ	Lambda
χ	Chi
ψ	Pseudo
μg	Microgram
$\mu \ell$	Microlitre
μm	Micrometre
A	Adenine
ACD	Acid citrate dextrose
ARMS	Amplification refractory mutation system
bp	Base pair
BRL	Bethesda Research Laboratories
C	Cytosine
°C	Degrees centigrade
CMC	Carboxy methyl cellulose
CVS	Chorionic villus sampling
dATP	Deoxyadenosine-5′-triphosphate
dCTP	Deoxycytidine-5′-triphosphate
dGTP	Deoxyguanosine-5′-triphosphate
$d\ell$	Decilitre
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxyribonucleotide-5′-triphosphate
dTTP	Deoxythymidine-5′-triphosphate

E.	*Escherichia*
EDTA	Ethylene-diamine-tetra-acetic acid
FBC	Full blood count
FEP	Free erythrocyte protoporphyrin
fℓ	Femtolitre
Fw	Framework
g	Gram
G	Guanine
Hb	Haemoglobin
Hct	Haematocrit
HLA	Human leucocyte antigen
HPFH	Hereditary persistence of fetal haemoglobin
HPLC	High pressure liquid chromatography
HVR	Hypervariable region
IVS	Intervening sequence
IZHVR	Interzeta hypervariable region
kb	Kilobases
ℓ	Litre
LA	Luria agar
LB	Luria broth (Luria Bertani medium)
LCR	Locus control region
m	Metre
M	Molar (moles/litre)
mA	Milliampere
Mb	Megabase
MCH	Mean cell haemoglobin
MCV	Mean cell volume
mg	Milligram
mℓ	Millilitre
mM	Millimolar (millimoles/litre)
mm	Millimetre
mRNA	Messenger RNA

N	Normal
nm	Nanometre
nt	Nucleotide
OD	Optical density
P.	*Plasmodium*
PCR	Polymerase chain reaction
pg	Picogram
pM	Picomolar
Poly-A	Polyadenylation
RBC	Red blood cell
RCC	Red cell count
RDW	Red cell distribution width
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
rpm	Revolutions per minute
SAIMR	South African Institute for Medical Research
SDS	Sodium dodecyl sulphate
T	Thymine
TBE	Tris borate EDTA
TE	Tris EDTA
Tris	Tris(hydroxymethyl) aminomethan
USSR	Union of Soviet Socialist Republics
UV	Ultraviolet
V	Volts
VNTR	Variable number tandem repeat
WHO	World Health Organisation
YAC	Yeast artificial chromosome

CHAPTER 1 - INTRODUCTION

The human globin gene family has been considered a paradigm for studying differential gene activity and the molecular basis of genetic disorders and gene expression. It has frequently set the pace and established the precedents for novel discoveries in the area of normal gene structure and function as well as mechanisms of abnormal gene expression leading to specific genetic diseases.

The search for the molecular basis of the thalassaemias reflects the recent history of the advances in cellular biology and biochemistry. The thalassaemias provided the first testing ground for the application of recombinant DNA technology to human disease. The human globin genes were among the first mammalian genes to be cloned into plasmids and analysed at the nucleotide level. The identification of mutations in globin genes in inherited disorders of haemoglobin synthesis provided the first near complete description of a disease at the molecular level and a basis for DNA-based prenatal diagnosis. The haemoglobinopathies were thus the first inherited disorders for which DNA analysis was used in prenatal testing.

Analysis of naturally-occurring mutations and experimental systems has added much to our understanding of the mechanisms underlying globin gene expression and the important regions in globin regulation. Molecular advances in the study of the genes have assisted in relating clinical phenotypes to underlying defects, in establishing prenatal diagnosis for the disorders, and also in population genetics studies. The globin genes have served as a model system and a point of reference for studies on the normal biology and molecular pathology of other gene systems and their associated disorders in man.

Disorders of haemoglobin synthesis (the haemoglobinopathies) are not only important as a research tool but also constitute a major disease load in some parts of the world. They result from the synthesis of structurally abnormal haemoglobin chains (the haemoglobin variants) or reduced numbers of globin chains (the thalassaemia syndromes). The haemoglobinopathies are among the commonest inherited disorders, generally having an autosomal recessive pattern of inheritance. There are estimated to be at least 190 million carriers worldwide and 240 000 infants are born annually with a major

haemoglobinopathy, sickle cell anaemia, ß or α thalassaemia. They have reached their high frequencies because the carrier state confers a selective advantage against *Plasmodium falciparum* malaria and their distribution therefore mirrors that of endemic *Plasmodium falciparum* malaria.

In South Africa, clinically significant haemoglobinopathies occur mainly in individuals of Mediterranean or Asian Indian origin, with the major disease load falling on the latter group. Thus this study was undertaken to characterise the normal and pathological variation around the α and ß clusters, both at the haematological and the molecular levels in South African Asian Indians. It was hoped that this information would help assess the need for a screening programme for the haemoglobinopathies, refine the techniques used in our prenatal diagnostic service and perhaps provide some insights into the origins of the different subgroups of the local Indian population.

In the introduction below the main historical developments, the structure and function of haemoglobin, the molecular organisation of the globin genes, the associated clinical diseases and their molecular bases and the ways in which this knowledge can be applied to prenatal diagnosis and treatment are reviewed. The geographical distribution of the haemoglobinopathies is also discussed. Finally the history of the South African Asian Indians is dealt with and the aims of the study outlined.

As the literature and knowledge of haemoglobins is so extensive, it has not always been possible to reference all the original literature and, instead, comprehensive reviews have been cited in some sections.

1.1 An historical overview

The advances in the field of haemoglobin and thalassaemia are too numerous to discuss individually. They have been divided into four phases and reviewed by Weatherall and Clegg 1981.

In the first phase, between 1925 and 1940, the original descriptions of the clinical features of the homozygous and heterozygous states for different types of thalassaemia appeared.

Although it is possible that Hippocrates referred to thalassaemia or sickle cell anaemia, the first clinical description of what was almost certainly homozygous ß thalassaemia was that of Cooley and Ley in 1925 in their paper entitled "A Series of Cases of Splenomegaly in Children with Anaemia and Peculiar Bone Changes". The early cases were all of Mediterranean background and the condition was thus named thalassaemia, by Whipple and Bradford in 1932, from θαλασσα, the Greek for 'sea'.

From 1940-1949, the second phase, the true genetic basis of the disorder was recognised. Caminopteros in Greece and Angelini in Italy had noted abnormalities in the red cells of relatives with thalassaemia as early as 1936. During the 1940's steady progress was made in elucidating the genetic basis of thalassaemia and by 1949 it was clear that Cooley's anaemia was the homozygous state of a "partially dominant Mendelian gene", the heterozygous state being present in parents of affected individuals and termed 'thalassaemia minor'.

Similarly, as early as 1910 Herrick had recognised that the cells of certain individuals had the peculiar property of undergoing reversible alterations in shape when the partial pressure of oxygen was lowered. Most people whose cells showed 'sickling' were healthy and were said to have the 'sickle cell trait', while others had severe fatal childhood anaemia, called 'sickle cell anaemia'. The disease was shown to be inherited and the pedigrees of families were most simply explained by the hypothesis that individuals with sickle cell trait and sickle cell anaemia were heterozygotes and homozygotes, respectively, for a particular abnormal gene (Neel 1949).

During the third phase, from 1949-1960, much was learnt about the structure and genetic control of haemoglobin, the different haemoglobins and their relative compositions in thalassaemia. Pauling and his colleagues showed that the haemoglobin present in the red cells of patients with sickle cell anaemia could be separated from normal adult haemoglobin by electrophoresis. They showed that the proteins differed in physical properties and presumably in structure. In addition, they demonstrated that haemoglobin from individuals with sickle cell trait contained a mixture of normal (HbA) and abnormal (HbS) haemoglobin, thus demonstrating a direct correlation between the proposed genetic basis and the haemoglobins synthesised (Pauling *et al.* 1949). The data also provided

evidence for a pair of genes controlling haemoglobin synthesis. The new electrophoresis technique was subsequently used to describe a large number of haemoglobin variants.

Further, Ingram demonstrated by 'protein fingerprinting' a position in the amino acid chain which was occupied by glutamic acid in HbA and valine in HbS, providing the first example of a genetically determined variant in which the structural abnormality was identified. This finding also demonstrated that different alleles code for different primary protein sequences (Ingram 1956, 1958).

During this period two non-allelic globin loci were shown to exist. The presence of the minor adult haemoglobin, HbA_2, was also demonstrated. Simultaneously the chemical structure of haemoglobin was being elucidated. The haemoglobin molecule was shown to have two identical halves, each consisting of an α and ß polypeptide chain. Fetal haemoglobin was shown to have γ in place of ß chains and by the early 1960's the complete amino acid sequences of the α, ß and γ chains had been determined.

It was shown that ß thalassaemia caused a reduction in HbA synthesis. In addition, a form of thalassaemia 'non-allelic' with sickle cell anaemia was recognised, in which the unusual haemoglobin variants, HbH and Hb Bart s, that lack α chains, were demonstrated.

In 1959 Ingram and Stretton proposed their model describing the genetic basis of thalassaemia. It incorporated Itano's structure rate hypothesis, which suggested that the primary structure of a haemoglobin determined its rate of synthesis, and Pauling's idea that thalassaemia was due to a low rate of production from the thalassaemia gene. Ingram and Stretton proposed two major classes of thalassaemia: α in which there is defective production of α chains and ß with defective ß chain production. They also explained the origin of HbH. They described the reduced rate of α or ß chain synthesis as due to a "hidden amino acid substitution" which was not visible on electrophoresis (Ingram and Stretton 1959). This paper according to Weatherall and Clegg (1981) was "a landmark which became the basis for all future work on the genetics and biosynthesis of haemoglobins in the thalassaemias and for the later elucidation of the molecular defect in some of these disorders".

During the fourth phase from 1960-1970, steady progress was made in elucidating the biochemical and molecular nature of the thalassaemias. Thalassaemia was shown to be a heterogeneous group of genetic disorders with a worldwide distribution. The disorders were found to be common, and the α and ß thalassaemias were distinguished more clearly. Globin chain synthesis, a technique developed by Clegg and Weatherall, aided the studies of the thalassaemias enormously; furthermore ineffective erythropoiesis was demonstrated to be the pathophysiological basis. The abnormalities in red cell maturation were ascribed to unbalanced globin production and precipitation of the chains produced in excess on the cell membrane, thus interfering with cell membrane function.

The 1970's marked the advent of the techniques of recombinant DNA technology, which has resulted in an exponential increase in the knowledge of globin gene structure and function. Some of the detailed molecular knowledge gained from the 1970's onwards will be reviewed in the sections that follow.

1.2 The protein structure and biosynthesis of haemoglobin

Haemoglobin, the major oxygen transport protein, is synthesised in adults in nucleated red cell precursors (normoblasts) which are located in the bone marrow. Globin synthesis persists for 24-48 hours in the anucleate circulating red blood cell but is absent from the mature red blood cell. The 5-7 day period of erythroid cell differentiation is marked by progressive specialisation of the cell for synthesis and accumulation of large amounts of haemoglobin, so that by the time the reticulocyte stage is reached, haemoglobin synthesis accounts for over 95% of total cellular protein produced. The high haemoglobin concentration in red blood cells and their easy availability is an important reason for the major role haemoglobin has played in protein, genetic and molecular studies.

All human haemoglobins are tetrameric, consisting of two identical dimers each with an α-like and a ß-like globin polypeptide chain. Each of the four globin polypeptide chains is associated with a haem molecule, which lies in a non-polar cleft in the chain and binds oxygen. The monomers are linked together by electrostatic and hydrophobic interactions. Haemoglobin is an allosteric protein, which changes conformation as it binds oxygen, such that oxygen binding at one site facilitates binding at the three other sites.

The globin chains are members of two developmentally regulated families: the α-like globins with 141 amino acids and the ß-like globins with 146 amino acids. The α-like chains are either ζ or α, and the ß-like chains are ϵ, γ, δ or ß. The synthesis of α-like and non α-like chains appears to be regulated throughout erythropoiesis so that nearly equal amounts of each are produced. Consequently, nearly all newly synthesised subunits are rapidly incorporated into soluble tetramers, with only minute amounts of unpaired chains accumulating, and these are removed rapidly by proteolysis. This precision of synthesis is physiologically important because unpaired globin chains are insoluble and tend to precipitate in erythroid cells.

The protein structure and biosynthesis of haemoglobin is reviewed in Benz and Forget 1975, Weatherall and Clegg 1981.

1.2.1 The ontogeny of human haemoglobins

Human haemoglobin is heterogeneous at all stages of development. Haemoglobins Gower 1 ($\zeta_2\epsilon_2$), Gower 2 ($\alpha_2\epsilon_2$) and Portland ($\zeta_2\gamma_2$) predominate in the early embryo. By 8-10 weeks of gestation HbF ($\alpha_2\gamma_2$) constitutes over 90% of the circulating haemoglobin. HbF may be of two types, as there are two types of γ chain, differing by one amino acid at position 136, where they have either glycine ($^G\gamma$) or alanine ($^A\gamma$) (Schroeder *et al.* 1968). The $^G\gamma$ and $^A\gamma$ chains are produced in a ratio of 3:1 during fetal life. This alters to the adult ratio of 2:3 within a few months of birth (Huisman *et al.* 1977). From 36 weeks onwards there is a gradual decrease in HbF and a concomitant increase in adult HbA ($\alpha_2\beta_2$) production. In the normal adult about 97% of haemoglobin is HbA, while about 2.5% is HbA_2 ($\alpha_2\delta_2$) and about 0.5% is HbF.

The switches between the different types of haemoglobin correlate broadly with the sites of erythropoiesis. Early erythropoiesis, up to six to seven weeks of gestation, occurs in the yolk sac where the ϵ genes, as well as α and predominantly ζ genes, are expressed in primitive erythroblasts (Peschle *et al.* 1985). The liver is the site of erythropoiesis from five weeks gestation to term, during which time definitive line erythroblasts in the liver synthesise α and γ globin predominantly (Peschle *et al.* 1985, Albitar *et al.* 1992). Hepatic haemopoiesis begins to decline at 20 weeks gestation as the bone marrow begins

to take over as the main site of blood production. α and ß globin become the predominant chains, though δ and a small amount of γ chains are also produced. Although ζ and ϵ mRNA have been detected in normal adults (Albitar *et al.* 1989), and low levels of ζ globin chains have been detected in cord blood and in adults with some types of α thalassaemia, no ζ globin has been detected in normal adults (Chung *et al.* 1984, Hill *et al.* 1985a, Chui *et al.* 1986, Yagi *et al.* 1986, Luo *et al.* 1988, Tang *et al.* 1992).

The developmental expression of the globin genes, together with the sites of haemopoiesis, is shown schematically in Figure 1.1.

1.3 The organisation and structure of the globin genes

On the basis of evidence from protein and DNA sequence data, it appears that all the globin genes have a common ancestor. The ancestral α-like and ß-like genes were separated by a translocation or transpositional event approximately 500 million years ago, so that the gene clusters are now present on different chromosomes (Efstratiadis *et al.* 1980). Each cluster has evolved by gene duplication, and the genes have been modified by various processes, including divergence, deletion, gene conversion and retroposition. Both clusters have pseudogenes, thought to be due to ancestral duplication followed by sequence alteration in the coding or regulatory regions leading to their inactivation.

The human globin genes, which have all been sequenced fully, have a basic common structure conserved within and between species. They are compact, 1-2kb, with three exons and two introns, in homologous positions relative to the coding sequence, but of variable length and sequence. In the α-like genes the introns occur between codons 31 and 32 and between codons 99 and 100, whereas in the ß-like genes they occur between codons 30 and 31 and between codons 104 and 105. The coding sequences of the genes are conserved so that exon 2 encodes the regions of haem-binding and $\alpha_1\beta_2$ interaction, while exon 3 encodes the region of $\alpha_1\beta_1$ contact. The codon/intervening sequence splice junctions conform to the Chambon rule (5'-GT/AG-3') that is a prerequisite for normal splicing of nuclear mRNA. Conserved consensus sequences, for the donor site (namely the last three nucleotides of the exon and first six nucleotides of the intron) and for the

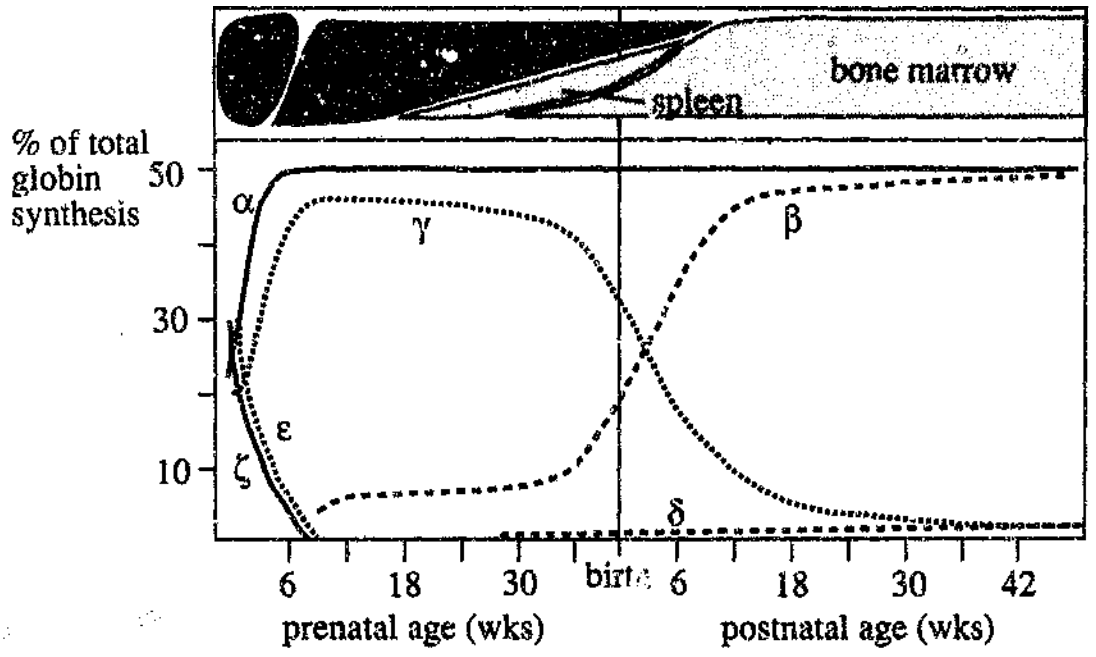


FIGURE 1.1 The developmental expression of the globin genes, together with the sites of haemopoiesis (from Grosveld *et al.* 1993).

acceptor site (the last 10 nucleotides of the intron and first nuclectide of the exon) are also required for normal splicing.

Conserved sequences essential for expression in eukaryotic genes surround the globin genes. The TATA box serves to locate transcription initiation at the cap site, about 30 bases downstream from it, and influences the rate of transcription. There are two upstream promoters required for optimal transcription: the CCAAT box and a GC-rich region with a sequence that can be inverted or duplicated, located 70-80bp and 80-100bp upstream of the cap site, respectively. The conserved polyadenylation signal site, AATAAA, is situated 3′ to the gene and is important in cutting the RNA transcript and in the addition of adenosines to form a poly-A tract, necessary for nuclear to cytoplasmic transport and mRNA stability in the cytoplasm.

All the globin genes are expressed only in erythroid cells, with different members in each family expressed sequentially from 5′ to 3′ during development (Fritsch *et al.* 1980, Lauer *et al.* 1980). This arrangement occurs in several species and is thought to be important in regulation.

The organisation and structure of the globin genes is reviewed in Orkin and Nathan 1981a, Collins and Weissman 1984, Nienhuis *et al.* 1984, Orkin and Kazazian 1984, Higgs *et al.* 1989, Kazazian 1990.

1.3.1 The α globin gene cluster

The α globin gene cluster, initially localised to chromosome 16 using somatic cell hybrids (Deisseroth *et al.* 1977), has been finely mapped to 16p13.3→pter (Buckle *et al.* 1988), lying within a few hundred kilobases of the telomere (Wilkie *et al.* 1991a). It spans 26kb and consists, from 5′ to 3′, of the embryonic ζ2 gene, the ψζ1 gene (which may sometimes have the form of a functional ζ gene), ψα2, ψα1, the two duplicated adult α genes (α2 and α1), and the θ1 gene (Lauer *et al.* 1980, Pressley *et al.* 1980, Proudfoot and Maniatis 1980, Proudfoot *et al.* 1982, Hill *et al.* 1985a, Hardison *et al.* 1986, Marks *et al.* 1986a, Hsu *et al.* 1988, Higgs *et al.* 1989).

The α globin gene cluster is situated in a GC-rich isochore, greater than 300kb in length and it, like the surrounding DNA, has features typical of such a region. It is also GC-rich, early-replicating, associated with non-methylated CpG islands and contains *Alu* family repeats (Fischel-Ghodsian *et al.* 1987a, Fischel-Ghodsian *et al.* 1987b). Such regions of the genome normally contain a high proportion of housekeeping genes and the tissue-specific α globin gene cluster is an interesting exception (Higgs *et al.* 1989).

1.3.1.1 The ζ2 and ψζ1 genes

The amino acid sequence of the ζ globin chain corresponds to the nucleotide sequence of ζ2, the functional ζ gene (Pressley *et al.* 1980, Proudfoot *et al.* 1982). The ζ2 and ψζ1 genes are each situated in a duplicated 2 000bp region and are highly homologous. Except for some repeat sequence variation in the introns, the two genes differ by only six base pairs, one of which introduces a stop mutation into codon 6 of ψζ1, converting it into a pseudogene. The pseudogene cannot direct protein synthesis or transcription, even when this mutation is restored to a normal codon through gene conversion by ζ2, suggesting that other sequences in the divergent 3′ non-coding region are important for expression (Proudfoot *et al.* 1982, Hill *et al.* 1985a). The downstream ζ-like gene thus exists in two forms: a pseudogene or a ζ1 form, the frequencies of which vary between populations (Hill *et al.* 1985a). The introns in ζ genes are much longer than in other α-like genes and, unlike other globin genes, IVS1 is larger than IVS2. Both introns of the ζ genes contain tandem repeats (Proudfoot *et al.* 1982, Goodbourn *et al.* 1983).

1.3.1.2 The ψα1 and ψα2 genes

The ψα1 gene is 73% homologous to the α2 gene, but has transcriptional control mutations, an initiation codon mutation, an altered reading frame resulting in premature stop codons and a non-functional polyadenylation site (Proudfoot and Maniatis 1980, Whitelaw and Proudfoot 1983).

The ψα2 gene lacks a promoter region and is inactivated by frameshift mutations and transcriptional control mutations (Hardison *et al.* 1986).

The α globin gene cluster is situated in a GC-rich isochore, greater than 300kb in length and it, like the surrounding DNA, has features typical of such a region. It is also GC-rich, early-replicating, associated with non-methylated CpG islands and contains *Alu* family repeats (Fischel-Ghodsian *et al.* 1987a, Fischel-Ghodsian *et al.* 1987b). Such regions of the genome normally contain a high proportion of housekeeping genes and the tissue-specific α globin gene cluster is an interesting exception (Higgs *et al.* 1989).

1.3.1.1 The ζ2 and ψζ1 genes

The amino acid sequence of the ζ globin chain corresponds to the nucleotide sequence of ζ2, the functional ζ gene (Pressley *et al.* 1980, Proudfoot *et al.* 1982). The ζ2 and ψζ1 genes are each situated in a duplicated 2 000bp region and are highly homologous. Except for some repeat sequence variation in the introns, the two genes differ by only six base pairs, one of which introduces a stop mutation into codon 6 of ψζ1, converting it into a pseudogene. The pseudogene cannot direct protein synthesis or transcription, even when this mutation is restored to a normal codon through gene conversion by ζ2, suggesting that other sequences in the divergent 3′ non-coding region are important for expression (Proudfoot *et al.* 1982, Hill *et al.* 1985a). The downstream ζ-like gene thus exists in two forms: a pseudogene or a ζ1 form, the frequencies of which vary between populations (Hill *et al.* 1985a). The introns in ζ genes are much longer than in other α-like genes and, unlike other globin genes, IVS1 is larger than IVS2. Both introns of the ζ genes contain tandem repeats (Proudfoot *et al.* 1982, Goodbourn *et al.* 1983).

1.3.1.2 The ψα1 and ψα2 genes

The ψα1 gene is 73% homologous to the α2 gene, but has transcriptional control mutations, an initiation codon mutation, an altered reading frame resulting in premature stop codons and a non-functional polyadenylation site (Proudfoot and Maniatis 1980, Whitelaw and Proudfoot 1983).

The ψα2 gene lacks a promoter region and is inactivated by frameshift mutations and transcriptional control mutations (Hardison *et al.* 1986).

1.3.1.3 The α1 and α2 genes

Family studies of specific α globin mutations had suggested that the α genes were duplicated in humans but it was thought that the normal number of genes could be polymorphic (Hollán *et al.* 1972). Solution hybridisation studies showed unequivocally that normal individuals had four α genes (Kan *et al.* 1975a). The α genes are duplicated in many species, suggesting that this event took place prior to their divergence, 300 million years ago (Zimmer *et al.* 1980).

The α2 and α1 genes are located 3.7kb apart, within a region of homology of approximately 4kb, interrupted by two small non-homologous regions. The homologous segments are thought to be the result of an ancient duplication event, while the non-homologous sequences may have arisen later by insertion. Thus, the sequences within and flanking the two α genes are virtually identical. The exons and IVS1 of the two genes are identical, while IVS2 of α1 is 9bp longer and differs by 3bp from the same region of α2. The two genes diverge considerably in their 3′ untranslated sequence, 13 bases beyond the TAA stop codon (Orkin 1978, Lauer *et al.* 1980, Liebhaber *et al.* 1980, Michelson and Orkin 1980, Liebhaber *et al.* 1981, Michelson and Orkin 1983, Hess *et al.* 1984). The high homology between the two genes is important in unequal crossover.

Although the two α genes encode identical polypeptide chains (Földi *et al.* 1980), it has been shown using the 3′ sequence divergence that the α2 gene produces two to three times as much mRNA as the α1 gene (Liebhaber and Kan 1981, Orkin and Goff 1981, Liebhaber and Cash 1985) and two to threefold more protein than α1 (Liebhaber *et al.* 1986). Thus, the human α cluster was said to contain a major and a minor α locus (Liebhaber *et al.* 1986). It has been shown recently that the ratio of α2:α1 protein is 1.19:1 [Molchanova *et al.* (1994) *Br J Haematol* 88, 300-6].

The dominant pattern of α2 is established after eight weeks gestation and maintained at all stages of development (Orkin and Goff 1981). Prior to eight weeks of development α2 and α1 are equally expressed (Albitar *et al.* 1992).

1.3.1.4 The θ1 gene

The θ1 gene is highly conserved in different species. It appears to have no inactivating mutations, although the promoter sequences, CCAAT and TATA boxes, are displaced from the coding sequence by the insertion of a 200bp GC-rich sequence (Marks *et al.* 1986a,b, Clegg 1987, Shaw *et al.* 1987, Hsu *et al.* 1988).

θ mRNA has been found in embryonic and fetal erythroid tissue as well as in trace amounts in adult reticulocytes, but there is no evidence that the gene is expressed as a stable protein (Leung *et al.* 1987, Hsu *et al.* 1988, Albitar *et al.* 1989, Mamalaki *et al.* 1990). The θ gene shows a pattern of developmental control reciprocal to ζ and parallel to α, such that it appears to represent a poorly expressed fetal/adult gene (Albitar *et al.* 1989). Its position in the α globin cluster does not appear to reflect its developmental expression. Individuals with deletions of the gene have no recognisable phenotype (Fischel-Ghodsian *et al.* 1987c, Fei *et al.* 1988a).

A truncated processed copy of the θ1 gene (ψθ2) has been found on chromosome 22 (Shaw *et al.* 1987).

1.3.2 The ß globin gene cluster

Early studies of Hb Lepore and Kenya had suggested that the ß, γ and δ genes resided on a single chromosome (Baglioni 1962, Huisman *et al.* 1972). The ß globin gene cluster was initially localised to chromosome 11p using somatic cell hybrids (Deisseroth *et al.* 1978, Jeffreys *et al.* 1979) and has been mapped to 11p15 (Magenis *et al.* 1985). The cluster spans a region of over 60kb and consists from 5′ to 3′, in order of developmental expression, of the embryonic ε gene, the two fetal γ genes (^{G}γ and ^{A}γ), the ψß1 gene and the adult δ and ß genes (Flavell *et al.* 1978, Mears *et al.* 1978, Lawn *et al.* 1978, Bernards *et al.* 1979, Little *et al.* 1979, Tuan *et al.* 1979, Fritsch *et al.* 1980).

The ß-like genes are generally longer than the α-like genes because of a larger IVS2. With the exception of the γ genes, there is little sequence homology between IVS2 in the ß-like genes (Efstratiadis *et al.* 1980). In contrast to the α cluster, the genes of the ß

cluster, have many similarities to other tissue-specific genes in their structure (Efstratiadis *et al.* 1980), methylation profile (Van der Ploeg and Flavell 1980) and pattern of replication timing (Holmquist 1987, Epner *et al.* 1988). The cluster also contains *Alu* and *Kpn* family repeats (Duncan *et al.* 1981, Shafit-Zagardo *et al.* 1982).

1.3.2.1 The ϵ, $^{G}\gamma$ and $^{A}\gamma$ genes

The ϵ gene, located 13.3kb 5′ to $^{G}\gamma$, was shown to be a globin-related sequence (Baralle *et al.* 1980). The two γ genes, situated 3.5kb apart, are thought to have arisen by gene duplication and are situated in a 5kb region of homology (Little *et al.* 1979, Tuan *et al.* 1979, Shen *et al.* 1981). The 5′ flanking, 5′ untranslated, coding, the IVS1 and part of the IVS2 sequences are virtually identical, homology being maintained by frequent gene conversion events. There is only one nucleotide sequence difference in the coding region, namely that in codon 136 - GGA coding for glycine in $^{G}\gamma$ and GCA coding for alanine in $^{A}\gamma$. The genes diverge considerably in the 3′ untranslated region (Cavallesco *et al.* 1980, Slightom *et al.* 1980, Starck *et al.* 1990).

1.3.2.2 The $\psi\beta 1$ gene

This true pseudogene is located between the $^{A}\gamma$ and δ genes. A sequence 6-9kb 5′ to ϵ, thought initially to represent a pseudogene, was shown not to be a globin-related sequence (Fritsch *et al.* 1980, Shen and Smithies 1982).

1.3.2.3 The δ and β genes

The δ and β globin chains differ by only 10 out of 146 amino acids. The genes reflect this homology which is maintained from the 5′ flanking region to 3′ to the polyadenylation sequence, with the exception of IVS2 and the 3′ untranslated region (Efstratiadis *et al.* 1980).

1.4 Normal variation in the globin gene clusters

Apart from monozygotic twins, no two individuals are genetically identical. Alterations

in DNA sequence in certain parts of the genome, particularly in coding and regulatory sequences, may cause pathology, but most variation is phenotypically silent and occurs in the flanking sequences of genes, between genes, and in introns. Such 'normal' variation in DNA is extensive and any particular variant, if common, constitutes a DNA polymorphism.

DNA polymorphism may be due to nucleotide substitution, or insertion or deletion of a single-copy segment of DNA (Botstein *et al.* 1980). Variation is also caused by the insertion / deletion of varying numbers of tandemly-repeated short sequences (VNTRs) (Nakamura *et al.* 1987), variation in the number of GT or AC repeats at a locus (Weber and May 1989), and length variation at the 3' end of *Alu* sequences (Economou *et al.* 1989).

Nucleotide substitution, or insertion or deletion, may result in two different alleles at a particular locus. In some cases, the substitution may remove or create a cleavage site for a restriction endonuclease, which can be used to detect it. Insertions or deletions may alter the size of a fragment generated by a restriction endonuclease which cuts outside the region. These mutations were traditionally called 'restriction fragment length polymorphisms' (RFLPs). VNTRs, also a form of RFLP, usually have many alleles, as the number of repeat sequences is altered by unequal exchange at mitosis and meiosis or DNA slippage during replication. They are inherited in Mendelian fashion and the length polymorphism can also be detected by restriction enzymes which cut outside the variable region or by PCR amplification of the region. They are thought to have a relatively high rate of new mutation (1/20-1/250 meioses) (Jarman *et al.* 1986, Jeffreys *et al.* 1988). Many new techniques have been developed to analyse DNA polymorphism. Polymorphisms are useful in population genetics, for mapping disease genes and for prenatal diagnosis and carrier detection.

The pattern of polymorphisms along a chromosome constitutes a haplotype, a concept first applied to DNA analysis by Antonarakis *et al.* (1982a). Haplotypes, which generally require family studies for their determination, are potentially useful in evolutionary and anthropological studies, for epistatic effects and also to study the history of mutations and gene flow (Orkin *et al.* 1982a, Antonarakis *et al.* 1985, Wainscoat *et al.* 1986a, Nagel

and Ranney 1990). Newer techniques using single molecules allow for haplotyping without families (Ruano and Kidd 1989, Ruano *et al.* 1990).

1.4.1. The α globin cluster

On average one in every 80bp in the α globin gene cluster varies between individuals and heterozygosity at the locus is estimated at 0.93 (Higgs *et al.* 1986). Some of the variants have been identified as RFLPs within and flanking the cluster. The rare variants are often population-specific, whereas the more common ones occur in most populations, with the frequency of the rarer allele being greater than 0.05 (Higgs *et al.* 1986).

Sixteen dimorphic RFLPs have been described in the α globin cluster (Higgs *et al.* 1986, Fodde *et al.* 1990) as well as a number of hypervariable regions, consisting of varying numbers of tandem arrays of short, apparently related GC-rich sequences (Higgs *et al.* 1989). The hypervariable regions are situated: 5′ to the cluster 70kb upstream of ζ2 (5′HVR) (Jarman and Higgs 1988); in IVS1 and IVS2 of both the ζ2 and ψζ1 genes (Proudfoot *et al.* 1982); between the ζ2 and ψζ1 genes (interζ HVR) (Goodbourn *et al.* 1983, 1984); and 3′ to the cluster (3′HVR) (Higgs *et al.* 1981, Jarman *et al.* 1986). Such hypervariable regions may be particularly frequent at telomeres (Higgs *et al.* 1989).

Further polymorphism has been described in and around the α globin cluster including a multiallelic locus at -122kb (Higgs *et al.* 1989), a dinucleotide repeat in ψα1 (Fougerousse *et al.* 1992), and a major length polymorphism at the telomere such that the α globin cluster is 170kb, 350kb or 430kb from the telomere, the latter being due to a large imperfect (CA)n repeat (Wilkie *et al.* 1991a, Wilkie and Higgs 1992)

Nine of the polymorphic markers surrounding the α globin cluster have been used to construct α globin haplotypes. The markers show marked linkage disequilibrium, so that within a population there are a few common haplotypes and several others which occur at low frequency. The haplotypes are classed according to their 3′ ends and, in general, haplotypes Ia to IVa appear to be the most frequent in most populations. Africans appear to have a wider array of haplotypes, including some rarely seen in Eurasians (Higgs *et al.* 1986). Haplotype analysis of the α globin cluster has provided evidence for multiple

origins of the single α gene deletion or -α chromosome (Higgs *et al.* 1986, 1989).

Numerous gene rearrangements have been described in the α globin cluster resulting in chromosomes with between zero and four α globin genes and one to four ζ globin genes (Embury *et al.* 1979, Goossens *et al.* 1980, Higgs *et al.* 1980, Winichagoon *et al.* 1982, Rappaport *et al.* 1984, Felice *et al.* 1986, Gu *et al.* 1987, Titus and Hunt 1987). These are thought to be due to the highly homologous nature of adjacent genes, which promotes mispairing and unequal crossover. Several of the α gene rearrangements are of clinical significance.

A number of insertions and deletions have also been described in the α globin cluster, including deletions involving the θ1 (Fei *et al.* 1988a), an insertion between α2 and α1 (Nakatsuji *et al.* 1986) and an insertion/deletion polymorphism in the 5′ flanking region of the α complex that involves the *Alu* family of repeats (Higgs *et al.* 1989).

The α globin cluster, in contrast to the ß cluster, does not show different rates of recombination in different regions (Higgs *et al.* 1986). It is highly polymorphic and the variants are useful as genetic markers for the α globin gene complex and the telomeric region of chromosome 16p. It is thought that studies of these markers may help differentiate important regions of the cluster and illustrate important evolutionary mechanisms (Higgs *et al.* 1989).

1.4.2. The ß globin cluster

The ß globin cluster is thought to be less polymorphic than the α cluster, yet approximately one in every 100 nucleotides of the ß globin cluster was shown to be polymorphic (Jeffreys 1979).

Eighteen single point RFLPs, spanning 63kb, have been defined in the ß globin cluster (reviewed in Boehm and Kazazian 1989). The first, a *Pst*I site located in IVS2 of the δ gene, was detected during a study of cloned DNA from the region of the δ and ß genes, but was found to be a relatively rare polymorphism (Lawn *et al.* 1978). The first useful polymorphism, discovered simultaneously, was the *Hpa*I site 3′ to the ß gene. The site

was frequently absent downstream from the β^S gene and present on β^A chromosomes (Kan and Dozy 1978a). Of the 18 RFLPs described, the majority are found in non-coding DNA. Most sites are public, the frequency of the rarer allele being greater than 5% in most populations, while a few are only polymorphic in Blacks (reviewed in Orkin and Kazazian 1984).

A simple tandem repeat sequence has been described in the 5′ flanking region of the gene. Individuals have between four and seven repeat units and may have point mutations in the repeat sequence (Spritz 1981, Takahashi *et al.* 1991). A second repeat of the form $(AT)_xT_y$ is found 5′ to the ß globin gene (Moschonas *et al.* 1982, Poncz *et al.* 1983, Chebloune *et al.* 1984, Semenza *et al.* 1984a).

Sequence polymorphism has been observed within the ß globin gene itself, so that in most populations three different normal ß globin genes exist, although they occur at different frequencies. These different forms are termed 'frameworks'. Frameworks 1 and 2 differ by a single nucleotide, and are found in all groups studied to date. Framework 3 has four additional substitutions in Mediterraneans but only three in Asian Indians, south-east Asians and Blacks. The frameworks can be identified by direct sequencing or by RFLP analysis as they are in linkage disequilibrium with the *Ava*II/ß and *Bam*HI/ß polymorphisms (Antonarakis *et al.* 1982b, Orkin *et al.* 1982a). A new framework variant has recently been described in Africa (Carestia *et al.* 1991).

A hypervariable *Alu* polymorphism (Economou *et al.* 1989) and an $(AC)_n$ repeat (Weber and May 1989) have been described in the ß globin cluster.

The ß globin haplotype has been considered in three domains spanning 63kb: on the 5′ side, a 34kb region from the *Hin*cII site 5′ to the ε gene to the *Taq*I site 5′ to the δ gene; and on the 3′ side, from the 5′ end of the ß gene, extending 19kb in a 3′ direction and a region between the 5′ and 3′ clusters, spanning about 9kb. In the 5′ and 3′ domains there is non-random association of polymorphic sites, while in the region between the two there is random association with either segment and a relatively higher rate of recombination may occur. The precise boundaries of this central region are unknown. It has been estimated that 75% of the recombination in the ß cluster occurs in

this region and that recombination in this segment is about 3-30 times the average expected rate for the length of DNA and is thus a hotspot for meiotic recombination. The mechanism for high recombination is uncertain, though a repeat sequence may be important (Miesfeld *et al.* 1981, Antonarakis *et al.* 1982a, Chakravarti *et al.* 1984).

As in the α cluster, polymorphic sites are not randomly associated, thus only seven to nine of the sites have typically been used to construct ß globin haplotypes and only a limited number of the possible haplotypes have been observed (Antonarakis *et al.* 1982a).

Five to 10 ß globin haplotypes have been observed among normal individuals in most ethnic groups studied, with three or four being most frequent in each group. Most are also represented on ß thalassaemia chromosomes. Three 5′ haplotypes account for 94% of chromosomes in Caucasians and Orientals while four of the 3′ haplotypes account for 90% of chromosomes. The frameworks are invariably associated with the 3′ haplotypes, two haplotypes are associated with framework 1, a third with framework 2 and a fourth with framework 3 (Antonarakis *et al.* 1982b, Orkin *et al.* 1982a).

The ß globin haplotypes are similar in most populations though they differ markedly in Africans, suggesting early emergence of a small population from Africa with subsequent racial divergence of Orientals and Caucasoids (Wainscoat *et al.* 1986a). The ß globin haplotypes have also assisted in the rapid characterisation of ß thalassaemia alleles. It was predicted that a mutation present on one haplotype would differ from one on a second (Orkin *et al.* 1982a). A systematic strategy to identify mutations was thus established, although the association was not absolute.

Rearrangements have also been demonstrated in the ß cluster, the most common involving the two γ genes. Chromosomes with one to five γ genes have been described (Trent *et al.* 1981a, Sukumaran *et al.* 1983, Harano *et al.* 1985a,b Hattori *et al.* 1986a, Hill *et al.* 1986, Shimasaki and Iuchi 1986, Fei *et al.* 1988b, Liu *et al.* 1988). While the mechanism generating the -γ chromosomes is uncertain, it seems that the $\gamma\gamma\gamma$ and $\gamma\gamma\gamma\gamma$ chromosomes have been generated by unequal sister chromatid exchange (Hill *et al.* 1986). In addition, chromosomes with only two $^{A}\gamma$ or two $^{G}\gamma$ genes exist (Powers *et al.* 1984).

The Lepore haemoglobins are also due to misalignment and crossover between homologous regions of the δ and ß genes (Baglioni 1962).

1.5 Regulation of globin gene expression

Despite the achievements in the study of haemoglobins, many questions remain concerning the control of globin gene expression, though there have been significant recent advances Any model that is proposed must explain how haemopoietic multipotent stem cells are directed to erythroid development, the erythroid specificity of globin gene expression and the expression of different genes during fetal development, together with a switching mechanism.

Expression appears to be mediated mainly at the level of transcription, with some fine tuning during translation and globin-subunit association Despite the overall similarity of the α and ß globin gene clusters, several lines of evidence suggest that the processes underlying α and ß globin expression may be somewhat different and that the higher order structure of the two clusters may be different (Higgs *et al.* 1990a). This poses interesting problems as to how the co-ordinated regulation of the α and ß clusters may occur. It appears that once the clusters are activated during development, no interplay in subsequent regulation occurs (Higgs *et al.* 1989).

In general, regulation of globin gene expression is the result of the interaction of ubiquitous transcription factors, erythroid-specific factors and stage-specific factors (reviewed in Grosveld *et al.* 1993, Higgs and Wood 1993).

Current models suggest that tissue-specific genes are arranged in discrete, independently controlled segments of chromatin, termed regulatory domains, in which transition from a closed to an open regulatory structure may be an important step in regulation of gene expression. These domains contain the structural genes and all *cis*-acting elements required for their expression. The domain is sensitive to DNase I digestion and important regulatory elements are associated with tissue-specific DNase I hypersensitive sites (reviewed in Vyas *et al.* 1992).

Both proximal and distal regulatory sequences interact with *trans*-acting cellular factors in mediating globin gene regulation. The distal regulatory sequences include the ßLCR and HS40 regions in the ß and α clusters respectively. In addition to distal regulation, all globin genes have a promoter region, with three positive acting elements (TATA, CCAAT and CACCC). These do not confer erythroid specificity as the proteins that bind them are ubiquitous transcription factors, although tissue- and/or developmental stage-restricted factors which bind these sequences may exist. Potential binding sites for additional transcription factors have been identified in the globin promoters. Enhancers and negatively acting elements have also been identified in the α and ß clusters (reviewed in Evans *et al.* 1990, Grosveld *et al.* 1993). It seems as though a few erythroid-specific factors co-operate with a small cadre of stage-specific proteins and ubiquitous factors to achieve gene activation and fine control.

Eight lineage-specific factors are known which may be involved in regulating erythroid-specificity of globin gene expression. The best characterised of these is GATA-1, while NF-E2 also appears to be important (reviewed in Grosveld *et al.* 1993, Higgs and Wood 1993).

GATA-1 appears to have a central role, with binding sites in the promoters and enhancers of erythroid expressed genes, including nearly all the globin genes, except α and in the active subregions of the αLCR and ßLCR. GATA elements appear to co-operate with CACCC boxes to direct specificity (reviewed in Evans *et al.* 1990, Orkin 1990, Grosveld *et al.* 1993). GATA-1 may bind to the γ promoter and extinguish expression late in development (Berry *et al.* 1992).

NF-E2 also appears to be important, with binding sites in the ßLCR and HS40 regions (Higgs and Wood 1993). Mutations in the mouse NF-E2 gene cause severe microcytic hypochromic anaemia, as a result of failure to regulate globin production and iron metabolism (Peters *et al.* 1993).

In the adult environment, a stage-specific DNA binding protein, NF-E4, is hypothesised to direct association of the ß promoter with the enhancer and thereby simultaneously reduce ε transcription (reviewed in Crossley and Orkin 1993).

The role of methylation in gene regulation has not been clarified. It appears to be involved in regulation, but is not a major mechanism (Van der Ploeg and Flavell 1980, Mavilio *et al.* 1983, Loo and Cauchi 1992).

1.5.1 The ß locus control region (ßLCR)

The ß globin cluster has been shown to lie in a regulatory domain and is thus consistent with the model of tissue-specific control. In erythroid cells, but not in other cell types, the chromatin extending for 120-200kb around these genes is early replicating, DNase I sensitive and transcriptionally active (Forrester *et al.* 1986, Epner *et al.* 1988, Dhar *et al.* 1989). It has been proposed that the ß globin locus control region (ßLCR), situated 6-18kb upstream of the ϵ gene and defined by four sites that are hypersensitive to DNase I cleavage in erythroid cells, plays a key part in regulating transcription, replication timing and chromatin structure of the ß domain (Tuan *et al.* 1985, Forrester *et al.* 1987, Grosveld *et al.* 1987, Forrester *et al.* 1989, Orkin 1990). The regions around DNase I hypersensitive sites appear to bind *trans*-acting factors (Talbot *et al.* 1990, Lowrey *et al.* 1992). The region of the second hypersensitive site, HS2, seems to account for at least 50% of total LCR activity (Collis *et al.* 1990, Philipsen *et al.* 1990). The sites each consist of 200-300bp segments of DNA containing binding sites for erythroid and ubiquitous transcription factors (reviewed in Grosveld *et al.* 1993). In addition, hypersensitive sites are located 3′ to expressed genes (Tuan *et al.* 1985).

The ßLCR region appears to establish and maintain an open chromatin structure which may permit *trans*-acting regulatory factors to gain access to individual genes. It confers a high level of expression on genes to which it is linked, independent of chromosomal position, although it is partly copy-number dependent (Grosveld *et al.* 1987, Talbot *et al.* 1989, Van Assenfeldt *et al.* 1989). The deletion of ßLCR in some thalassaemias inactivates the entire ß complex, as evidenced by DNase insensitivity and late replication over a distance extending at least 100kb 3′ of the ß globin gene (reviewed in Epner *et al.* 1992). As the ßLCR is in an active chromatin configuration prior to erythroid commitment, it may have a significant role for selective gene repression in lineage specification (Jimenez *et al.* 1992).

In addition, the ßLCR is thought to have a role in switching. It has been proposed, based on the model of chicken globin switching (Choi and Engel 1988), that interaction between the LCR and gene promoter elements is required for gene expression and that competition between the promoters is important. Competition is dependent on proximity to the LCR, achieved by loop formation and stage-specific regulatory factors, acting on sequences flanking the genes and perhaps the LCR. Genes influence each other through a competitive mechanism with the LCR. The mechanism involves spatial distribution of the genes relative to LCR and the stability of interactions between individual promoters and the different hypersensitive sites, although the relative contributions of these are unknown (Grosveld *et al.* 1993). Thus it appears that the inactivity of the ß globin gene early in development depends on competition from the 5′ genes, which interact preferentially with LCR as they are closer. The ε and γ globin genes are autonomously silenced during development by *trans*-acting factors, a mechanism which cannot be overcome by the LCR. Thus, ε silencing results in γ/LCR interaction and expression, and later γ silencing results in ß/LCR interaction and expression (reviewed in Epner *et al.* 1992, Crossley and Orkin 1993, Grosveld *et al.* 1993). The LCR can simultaneously interact with more than one promoter, however (Furukawa *et al.* 1994). Sequences that confer stage-specific silencing have been observed 5′ to the ε gene (Cao *et al.* 1989a, Raich *et al.* 1992) and a protein which binds to the silencer has been isolated (Wada-Kiyama *et al.* 1992). A stage-specific protein which bound to DNA sequences in the ßLCR, γ globin promoter and ^{A}γ enhancer (Lavelle *et al.* 1991) has been shown to be a product of the homeobox gene, HOXB2 (Sengupta *et al.* 1994). However, multiple control elements may be required for complete regulation (Gumucio *et al.* 1993).

Further, the possibility has been raised that subdomains of the ßLCR may selectively interact with the promoters and contribute to stage-specific expression (Crossley and Orkin 1993).

Nearly all the studies of LCR function in globin gene regulation have employed simple constructs that do not mimic the proper arrangement of elements in normal chromatin. A YAC containing the ßLCR, the ß globin cluster and the 3′ flanking sequences was isolated (Gaensler *et al.* 1991) and integrated into the germ line of transgenic mice (Gaensler *et al.* 1993) and into mouse erythroleukaemia cells where the globin genes were

regulated like the chromosomal globin genes (Peterson *et al.* 1993). A cosmid ligation approach that permits the introduction of the complete ß gene complex into transgenic mice has also been described (Strouboulis *et al.* 1992). Both offer potential for future studies on regulation of expression.

1.5.2 The α globin locus control region (HS40)

The regulation of the α globin cluster is rather more complex than that of the ß globin cluster. It lies within an early replicating GC-rich isochore in which there is a high density of constitutively expressed genes as well as the tissue-specific globin genes, with at least eight genes in 180kb (Vyas *et al.* 1992). The genes in the α globin complex are only expressed in erythroid cells and, as with the ß globin cluster, expression is dependent on a remote regulatory sequence located 40kb upstream of the ζ2 cap site, HS40 or αLCR, associated with an erythroid-specific DNase I hypersensitive site (Higgs *et al.* 1990a, Jarman *et al.* 1991). This region is similar in position, structure and function to the ßLCR and is copy-number dependent in part (Higgs *et al.* 1990a). There is a 350bp core fragment with a cluster of binding sites for GATA-1, NF-E2 and CACC box proteins (reviewed in Grosveld *et al.* 1993).

The α-like genes lie adjacent to at least four widely expressed genes, with HS40 lying in an intron of one of these genes, whose transcription is towards the telomere, in the opposite direction to that of the α cluster (Vyas *et al.* 1992). The 30kb segment containing the ζ-α cluster is sensitive to DNase I in erythroid cells, with erythroid-specific sites 5′ to ζ2, α1 and α2 (Yagi *et al.* 1986, Vyas *et al.* 1992). Thus, the *cis*-acting sequences of the α cluster extend over 65kb, the minimal extent of the erythroid domain, though no boundaries of the domain like those in the ß cluster have been defined, presumably as the region exists in an open chromatin conformation in all cells (Vyas *et al.* 1992). This is reinforced by finding no alteration in the replication timing of the α complex in erythroid and non-erythroid cells (Holmquist 1987, Vyas *et al.* 1992). The α genes themselves are not typical of tissue-specific genes as discussed in Section 1.3.1. Thus, the α regulatory region may not need to subserve all functions of the ßLCR, specifically that of 'domain-opening' (Vyas *et al.* 1992).

The αLCR may have a role in the co-ordinate regulation of ζ and α, which is probably partly dependent on gene order, and perhaps positive and negative *trans*-acting factors are important (Orkin 1990).

1.6. The molecular basis of disease

The genetic disorders of haemoglobin are divided into structural variants, thalassaemias and hereditary persistence of fetal haemoglobin (HPFH). The thalassaemias are characterised by a decreased rate of synthesis of one or more globin chains and are classified as α, ß, δ-ß and γ-δ-ß, depending on which chains are deficient. Ramirez *et al.* (1975) reported that α thalassaemia was due to deletions, while ß thalassaemia was due to abnormal processing of mRNA. Though this concept has been modified somewhat, the general principle remains true in the majority of cases. In the structural variants, structurally abnormal globin chains are produced, while HPFH constitutes a group of conditions in which the fetal to adult globin switch is disturbed.

In general, most thalassaemias and structural variants are mutations of exons, critical regions of introns, or regulatory sequences in or near the promoters. They are thus *cis*-acting and affect gene structure rather than molecules involved in modulation of gene expression, though there are some exceptions.

1.6.1 α thalassaemia

Although the molecular basis of the α thalassaemias is heterogeneous, deletions of part of the α globin gene cluster on chromosome 16 are the cause of the majority (more than 95%). Solution hybridisation studies first demonstrated that α thalassaemia is due to the inheritance of three, two, one or zero α genes (Ottolenghi *et al.* 1974, Taylor *et al.* 1974, Kan *et al.* 1975a). Deletions can be broadly classified into those with one α gene deleted (α^+) and those with both α genes deleted on the same chromosome (α^0).

1.6.1.1 Single α gene deletions (-α) (α thalassaemia 2)

Deletions of a single α gene are the commonest molecular defects causing α thalassaemia.

Two main types, $-\alpha^{3.7}$ and $-\alpha^{4.2}$, have been described (Embury *et al.* 1980a). The two α globin genes are embedded in two highly homologous segments of DNA, each about 4kb in length. These regions are divided into homologous subsegments (X, Y and Z) interrupted by non-homologous elements (Lauer *et al.* 1980, Michelson and Orkin 1983, Hess *et al.* 1983, 1984). These homologous segments can misalign and crossover resulting in chromosomes with either a single α ($-\alpha$) or three α genes ($\alpha\alpha\alpha$). This is shown schematically in Figure 1.2.

The homologous X boxes are located 4.2kb apart, 3' to $\psi\alpha1$ and $\alpha2$, respectively. Thus deletions involving misalignment of the X boxes (or 'leftward deletions') remove 4.2kb of DNA, including the $\alpha2$ gene, and a chromosome with only the $\alpha1$ gene remains, the so-called '$-\alpha^{4.2}$' chromosome. The homologous Z boxes, located 3.7kb apart, contain the $\alpha2$ and $\alpha1$ genes. The single α gene that results from the 3.7kb deletion (or 'rightward deletion'), the so-called '$-\alpha^{3.7}$' chromosome, has been classified into three subtypes, I, II and III, depending on the exact position in the Z box where the recombination event has occurred. The resultant α gene is commonly a Lepore-like fusion gene with the 5' end of $\alpha2$ and the 3' end of $\alpha1$, which produces an $\alpha1$-like mRNA (Higgs *et al.* 1984). There is, however, a compensatory increase in the mRNA produced by the single $\alpha1$-like gene (Liebhaber *et al.* 1985).

Apart from the chromosomes with an α gene deleted, a reciprocal chromosome also results, namely one with three α genes, though this is of little clinical significance (Goossens *et al.* 1980, Higgs *et al.* 1980). This can be of the $\alpha\alpha\alpha^{anti3.7}$ or $\alpha\alpha\alpha^{anti4.2}$ type, providing evidence that these chromosomes have originated from interchromosomal crossover, rather than intrachromosomal events (Lie-Injo *et al.* 1981, Trent *et al.* 1981b). Similarly $\alpha\alpha\alpha^{anti3.7}$ chromosomes corresponding to the different $-\alpha^{3.7}$ subtypes have been described. Further recombination events can give rise to quadruplicated α genes or other rearrangements (Nakatsuji *et al.* 1986, Gu *et al.* 1987, De Angioletti *et al.* 1992).

Recombination in the α cluster is thought to be common, as evidenced by the presence of the $-\alpha$ and $\alpha\alpha\alpha$ chromosomes in many population groups and the association of the $-\alpha$ chromosome with many haemoglobin variants, HVR alleles and haplotypes (Goodbourn *et al.* 1984, Winichagoon *et al.* 1984, Flint *et al.* 1986). The frequencies of the different

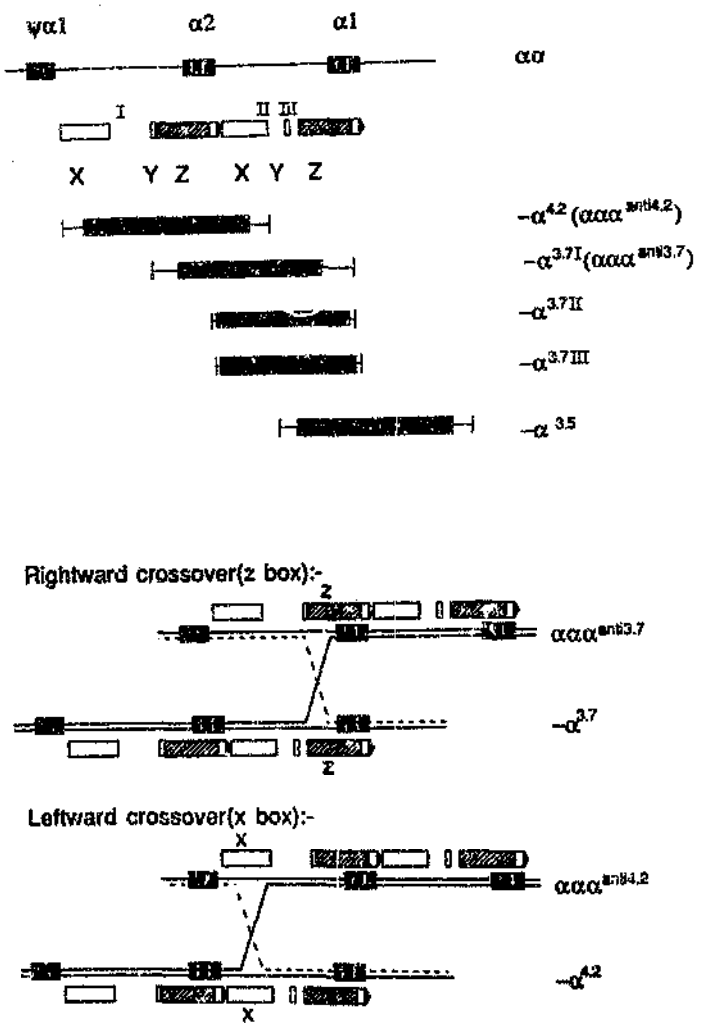


FIGURE 1.2 The α globin genes showing the duplication units divided into X, Y and Z boxes, separated by regions of non-homology (I, II and III) (after Higgs *et al.* 1989).

The extent of the common deletions are shown by black bars. Below, the mechanisms of the rightward and leftward crossovers, and their derivative chromosomes, are indicated.

types of -α chromosome correlate roughly with the length of the homologous segment that serves as a target area for crossover [-$\alpha^{3.7I}$ (1436bp) > -$\alpha^{4.2}$ (1339bp) > -$\alpha^{3.7II}$ (171bp) > -$\alpha^{3.7III}$ (46bp)]. The world frequencies not only reflect the recombination rate but also selection and other factors (Higgs *et al.* 1989). The $\alpha\alpha\alpha^{anti3.7}$ is the more common type of $\alpha\alpha\alpha$ chromosome, strengthening the argument for different rates of crossover (Higgs and Weatherall 1983).

A rare -α chromosome has been described in Asian Indians. It is due to a 3.5kb deletion of α1 and its flanking DNA, though the mechanism is uncertain (Kulozik *et al.* 1988a). Further, a 5.3kb deletion in an Italian family removes the 5′ end of α2 and results in a -α chromosome. It is thought to result from the formation of a multi-stem loop between homologous *Alu* sequences (Lacerra *et al.* 1991).

1.6.1.2. Double α gene deletions (--) (α thalassaemia 1)

Chromosomes on which both α genes are deleted are of more limited geographical distribution than the -α types, but are more important clinically since they can cause HbH disease and Hb Bart's hydrops fetalis.

A number of deletions have been reported which completely or partially remove both α genes on one chromosome so that no α synthesis is possible. At least 23 types exist, the most common being the south-east Asian and Mediterranean types. Some deletions remove the ζ genes as well (reviewed in Higgs *et al.* 1990b, Higgs 1993a).

The mechanism of occurrence of these deletions remains uncertain. Often one breakpoint of the deletion lies in or close to an *Alu* sequence. In addition, palindromes, direct repeat sequences, regions of weak homology and a GAGG motif have been found at the breakpoints. Most of the 3′ breakpoints lie in a 6-8kb segment between α1 and 3′ HVR, a hypomethylated region of the α cluster that may be associated with an unusual chromatin structure that predisposes to recombination (Nicholls *et al.* 1987, Higgs *et al.* 1989).

1.6.1.3 Other deletions causing α thalassaemia

Deletions of the αLCR region may cause thalassaemia, even though the α genes are still present (Hatton *et al.* 1990, Liebhaber *et al.* 1990, Romao *et al.* 1991).

Large terminal deletions of 16p13.3, which remove the entire α cluster, are associated with α thalassaemia as well as mental retardation and dysmorphic features (Wilkie *et al.* 1990a).

1.6.1.4 Non-deletion α thalassaemia mutations

Non-deletion defects in α thalassaemia are relatively rare, though at least 19 have now been characterised (reviewed in Higgs *et al.* 1990b, Higgs 1993b). The first non-deletion α thalassaemia was described by Kan *et al.* (1977), though the mutation was not identified.

Almost all the mutations described to date are in the α2 gene. This is not unexpected since α2 is the 'major locus' (Liebhaber *et al.* 1986) and thus mutations in α2 would be expected to have a greater effect on the phenotype.

Mutations affect gene functioning at all levels by interfering with RNA processing (splice junction and poly-A mutations), RNA translation (initiation codon, nonsense mutation, premature termination and termination codon mutations) or post-translational stability (missense mutations) (reviewed in Higgs *et al.* 1990b, Higgs 1993b).

1.6.2 ß thalassaemia

The ß thalassaemias are extremely heterogeneous at the molecular level. In contrast to the α globin cluster, they are mostly caused by point mutations or small nucleotide insertions or deletions and over 140 different mutations have been described, affecting almost every stage of gene expression. Each ethnic group appears to have only a subset of these mutations, with a small number (usually less than 7) accounting for 90-93% of the ß thalassaemia chromosomes present. A larger number of rarer alleles are observed,

however, in each group (reviewed in Kazazian 1990, Thein 1993, Quaife *et al.* 1994).

It is estimated that over 99% of ß thalassaemia genes in the world population have been defined (Kazazian and Boehm 1988). The mutations are characterised by deficient ($ß^+$) or absent ($ß^0$) ß globin chain production. It is hoped that detailed studies of the different mutations will help in more accurate prediction of the clinical phenotype, though this is extremely complex.

1.6.2.1 Non-deletion ß thalassaemia mutations

The point mutations have been classified into those affecting transcription, RNA processing and modification, translation, and those producing unstable globins. They are reviewed by Kazazian (1990), Kazazian *et al.* (1990) and Thein (1993).

Nearly 40% of the known mutations affect RNA translation. They may be nonsense mutations which terminate translation prematurely, frameshift mutations due to the insertion or deletion of one to seven nucleotides, or initiation codon mutations.

RNA processing mutants affect splicing, resulting in abnormal mRNA. Alterations of the critical GT and AT nucleotides, located at the 5′ and 3′ end of the intron respectively, result in $ß^0$ thalassaemias. Mutations in the consensus sequences produce both $ß^0$ and $ß^+$ thalassaemias, by reducing the efficiency of splicing. Mutations can create new consensus sequences in either an exon or intron. These may affect splicing by producing a new donor site or by activating a cryptic splice site, a sequence which mimics the consensus site and is not used under normal circumstance. Their severity depends on the proportion of normal mRNA produced.

The transcription mutations are concentrated in the 'TATA' box, located 30bp upstream of the cap site, or in the proximal and distal 'CACACCC' sequences at -90 and -105 nucleotides upstream of the ß globin gene. They are generally associated with reduced transcription beginning appropriately at the cap site, though the phenotype is very variable. No mutations have been described in the CAAT box, which is thought to be important in transcription regulation.

RNA modification mutations include those at the cap site and those affecting RNA cleavage and polyadenylation. Although the cap site is the start of transcription, the mutation at this site appears to have an extremely mild effect. In combination with a severe mutation it can cause severe disease (Wong *et al.* 1987). The AATAAA sequence is located 10-20bp 5′ to the site of cleavage of the RNA transcript and addition of the poly-A tail. Mutations in this signal sequence result in only a small percentage of transcripts being appropriately cleaved, with most transcripts not being cleaved until transcription reache another AATAAA signal 1-3kb 3′ to the gene. These abnormally elongated transcripts are unstable, and these mutations thus result in ß$^+$ thalassaemia (Orkin *et al.* 1985, Rund *et al.* 1992a).

1.6.2.2. Deletions of the ß globin genes

Seven deletions of the ß globin gene have been described (reviewed in Kazazian 1990, Thein 1993). All are rare, except one which is found in A' 'n Indians and removes 619bp of the 3′ end of the ß globin gene, from IVS2 onwards (Orkin *et al.* 1979a, 1980, Spritz and Orkin 1982, Kazazian *et al.* 1984a).

1.6.2.3 Other deletions causing ß thalassaemia

A number of deletions which leave the ß globin gene intact yet silence its expression, have now been shown to delete the ßLCR. They may also silence γ and δ expression (Van der Ploeg *et al.* 1980, Curtin *et al.* 1985, Curtin and Kan 1988, Driscoll *et al.* 1989).

1.6.2.4 Dominant ß thalassaemia mutations

A number of rare mutations have been described which cause severe disease in the heterozygous state. These mutations can be divided into: those resulting in an unstable ß variant as a result of a single base substitution or deletion of i tact codons, those with a truncated ß chain due to a premature termination codon, and those with an elongated ß chain with an altered carboxy-terminal as a result of frameshift mutation. Many of the mutations are located in exon 3 and it thus appears that a longer abnormal chain is more problematic than non-synthesis or degradation of a short chain (reviewed in Thein 1992).

1.6.2.5 ß thalassaemia due to mutations not linked to the ß globin cluster

An Albanian and an Italian family has been reported in which ß thalassaemia is unlinked to the ß cluster (Semenza *et al.* 1984b, Murru *et al.* 1992). A few other individuals are reported in whom sequencing of the entire ß globin gene and critical parts of the ßLCR have not revealed a mutation (Kazazian *et al.* 1990, Thein *et al.* 1993).

1.6.3 δß thalassaemia and HPFH

The $(\delta\beta)^0$ thalassaemias are generally due to relatively small deletions which remove or inactivate the δ and ß genes. However, deletions that cause $(^A\gamma\delta\beta)^0$ thalassaemia, extend into or beyond the $^A\gamma$ gene in the 5′ direction and beyond the ß gene in the 3′ direction. One form of Indian $(\delta\beta)^0$ thalassaemia is a complex rearrangement with two deletions and an inversion. The deletional types of HPFH form a continuum with the $(\delta\beta)^0$ thalassaemias in phenotype and extent of deletions, and differ only in their output of γ chains (reviewed in Kutlar and Lanclos 1987a,b, Cao and Murru 1989, Weatherall 1991a). At present there is no adequate theory to explain the phenotypic differences, though the deletion of a negative regulatory region, juxtaposition of enhancer sequences or the modification of chromatin structure have been suggested (Ottolenghi and Giglioni 1982, Ottolenghi *et al.* 1982, Tuan *et al.* 1983, Camaschella *et al.* 1987, 1990a).

δß thalassaemia is also associated with Lepore haemoglobins, which, after misalignment and recombination between a δ and a ß gene, result in a fusion gene with a low globin protein output (Baglioni 1962, Ottolenghi *et al.* 1979). Sardinian δß thalassaemia is due to two point mutations on the same chromosome, one in the ß gene and the second in the $^A\gamma$ promoter (Ottolenghi *et al.* 1988a).

In non-deletion HPFH, either the $^A\gamma$ or $^G\gamma$ gene is over-expressed, resulting in high HbF levels in adult life. This is in contrast to the deletional types in which both γ genes are over-expressed. The non-deletion HPFH disorders have been associated with point mutations clustered in three regions upstream of the transcription initiation site (reviewed in Forget 1990). The mutations may alter the binding of *trans*-acting factors to the

promoter regions and prevent postnatal suppression of γ expression, or they could prevent binding of negative regulatory factors, or enhance the binding of positive ones (Fucharoen *et al.* 1990a, Gumucio *et al.* 1990, Berry *et al.* 1992, Gumucio *et al.* 1991).

A C→T mutation at -158 to the $^{G}\gamma$ is associated with a type of $^{G}\gamma$ HPFH (Labie *et al.* 1985a), particularly when it occurs in association with the 5′ (-+--++) haplotype (Hattori *et al.* 1986b) and $(AT)_9T_5$ 5′ to the ß gene. Increased $^{G}\gamma$ expression occurs under conditions of erythropoietic stress although the mutation appears to have no effect on a normal chromosome (Thein *et al.* 1987a, Ragusa *et al.* 1992).

HPFH not linked to the ß cluster has also been reported (Cao and Murru 1989). One type has been linked to Xp22.2 (Dover *et al.* 1992), an interesting region as the erythroid-specific factor (NF-E1) has been localised to a similar area, Xp11.23 (Miyoshi *et al.* 1988, Zon *et al.* 1990, Cailulo *et al.* 1991). A second type has been localised to chromosome 7q36 (Sampietro and Thein 1991).

1.6.4 Structural variants

At least 621 structural haemoglobin variants have been described (reviewed in Huisman 1993). Such variants may arise by point mutations, deletions, insertions, frameshifts, chain termination mutations, or by unequal crossing over with the production of fusion genes.

The $ß^S$ allele is caused by an A→T mutation in codon 6 of the ß globin gene, which causes a glutamic acid to valine substitution in the polypeptide chain. Genetic data (based on haplotype studies, variation in an area of $(AT)_x(T)_y$ repeats 5′ to the ß globin gene and sequence variation 5′ to the ß globin gene) have defined five independent origins of the sickle mutation, four in Africa and one somewhere between the Horn of Oman and southern India (Wainscoat *et al.* 1983a, Antonarakis *et al.* 1984, Pagnier *et al.* 1984, Kulozik *et al.* 1986, Chebloune *et al.* 1988, Trabuchet *et al.* 1991a, Lapoumeroulie *et al.* 1992, Nagel and Fleming 1992). The five different haplotypes, termed Senegal, Benin, Central African Republic (or Bantu), Cameroon and Arab-Indian are important in determining the severity of the disease associated with the $ß^S$ mutation.

The β^E allele is caused by a G→A mutation at codon 26, resulting in an amino acid change from glutamic acid to lysine. This mutation reduces mRNA processing by activating a cryptic donor splice site, resulting in an aberrant mRNA which cannot be translated, in addition to a low level of normally spliced mRNA (Traeger *et al.* 1980, Orkin *et al.* 1982b).

β^C is the result of a G→A mutation in codon 6, resulting in an amino acid change from glutamic acid to lysine.

The first structural α variant described at the amino acid level was Hb Constant Spring, a mutation in the termination codon of $\alpha 2$, which is frequent in south-east Asia (Clegg *et al.* 1971, Milner *et al.* 1971).

1.7 Clinical features of the haemoglobinopathies

The thalassaemias are a diverse group of hereditary microcytic, hypochromic anaemias characterised by defective synthesis of one or more of the globin subunits. Continued synthesis of the unaffected chain at normal rates leads to the accumulation of relatively insoluble aggregates of unpaired globin chains, which precipitate and form inclusion bodies that damage the cell membrane and cause premature destruction of the red blood cell. In the structural variants, clinical features depend on the physical properties of the abnormal globin chain produced. Only the major features of the common types of thalassaemia and β^S variants will be discussed. The HPFH conditions are of little clinical significance but are important in studies of gene regulation. They are discussed further in Section 5.2.4.

1.7.1. The α thalassaemias

The α thalassaemias, due to impaired synthesis of α globin chains, are heterogeneous at the molecular level. There are potentially several hundred interactions that could take place between the different α thalassaemia determinants. The clinical severity depends ultimately on the resultant number of functional α genes and thus four clinical phenotypes can be distinguished from the normal: the silent carrier, α thalassaemia trait, HbH disease

and Hb Bart's hydrops fetalis. As the number of functional α genes decreases, the α/ß synthetic ratio in reticulocytes declines from the normal of 1 to approximately 0.9, 0.7, 0.4 and 0 respectively (Kan *et al.* 1968). Since production of non-α globin continues in fetal and postnatal life, non-α chains are produced in excess, forming γ_4 (Hb Bart's) in the fetus and $ß_4$ (HbH) in the adult.

Normal individuals have four α genes active in erythroid cells usually with an $\alpha\alpha/\alpha\alpha$ genotype. Individuals with $-\alpha/\alpha\alpha\alpha$ or $\alpha\alpha/\alpha\alpha\alpha$ are generally indistinguishable from normal $\alpha\alpha/\alpha\alpha$ individuals haematologically, though individuals with five α genes produce a slightly increased amount of $\alpha 2$ mRNA and α globin (Liebhaber and Kan 1981, Higgs *et al.* 1980, Galanello *et al.* 1983, Trent *et al.* 1985).

The α thalassaemias have been extensively reviewed by Wasi *et al.* 1974, Weatherall and Clegg 1981, Higgs and Weatherall 1983, Higgs *et al.* 1989, Kazazian 1990, Weatherall 1991a.

1.7.1.1 'Silent carrier' state

The mildest phenotype, or 'silent carrier' state, which is often difficult to distinguish from normal, occurs commonly in individuals with a single α gene deletion ($-\alpha/\alpha\alpha$). They have normal haematological parameters, except for a mean MCV of 78-80fℓ and a borderline low MCH. Some 'silent carriers' detected by α gene analysis have MCVs in the 80-85fℓ range. It appears as if the $-\alpha^{4.2}/\alpha\alpha$ genotype may produce a marginally more severe phenotype than the $-\alpha^{3.7}/\alpha\alpha$ type as reflected in a higher level of Hb Bart's (γ_4) at birth in the former type (Bowden *et al.* 1987). The mild phenotype is in part accounted for by a compensatory increase in $\alpha 1$ expression from the $-\alpha^{3.7}$ chromosome, perhaps because it has a 5′ $\alpha 2$ region (Liebhaber *et al.* 1985).

1.7.1.2 α thalassaemia trait

α thalassaemia trait includes individuals with mild haematological abnormalities, but no clinical symptoms. It is commonly due to either the $-\alpha/-\alpha$ or $--/\alpha\alpha$ genotype. Individuals have an MCV in the 70-75fℓ range, an MCH of 24-27pg, and a normal or low HbA_2

level. Some individuals, particularly children, may have mild anaemia which may lead to an erroneous diagnosis of iron deficiency anaemia. Although in general the correlation of phenotype and genotype is good, the genotype in a single individual cannot be predicted as there is overlap of haematological indices, globin chain synthetic ratios and levels of Hb Bart's in cord blood between the genotypes of α thalassaemia trait and also the silent carrier state.

1.7.1.3 HbH disease

HbH disease is most commonly due to the --/-α genotype, but can be due to homozygosity for non-deletion $\alpha 2$ mutations.

Although an extremely variable disease, there is a strong overall correlation between severity and degree of α chain deficiency. The disease is characterised by chronic haemolytic anaemia, with an Hb level between 2.6 and 13.3g/dℓ, jaundice and hepatosplenomegaly. Some patients may require regular blood transfusions.

The β_4 tetramers are soluble and only precipitate as the red cells age, forming inclusion bodies. The damaged cells are destroyed by the spleen, causing anaemia and ultimately resulting in the commonest complication, hypersplenism. Individuals may also get infections, leg ulcers, folic acid deficiency and acute haemolytic episodes. They rarely require hospitalisation and usually have a normal lifespan. Iron overload is uncommon, but has been recorded in older patients.

The blood film and indices show typical thalassaemic features. Incubation of the red cells with brilliant cresyl blue results in inclusion bodies in all the cells. In addition, while HbA always constitutes the major haemoglobin, HbH (β_4) varies between 5% and 30% and traces of Hb Bart's (γ_4) may be found.

1.7.1.4 Haemoglobin Bart's hydrops fetalis

The Hb Bart's hydrops fetalis syndrome is due to the absence of α chain synthesis and nearly all cases are due to the --/-- genotype. All affected individuals are stillborn at 34-40 weeks gestation or die shortly after birth of severe hydrops fetalis. Clinical features are those of a pale oedematous infant with signs of cardiac failure and prolonged intrauterine anaemia. Hepatosplenomegaly is present and there may be associated skeletal and cardiovascular anomalies. There is marked extra-medullary haemopoiesis and placental enlargement, reflecting the intra-uterine hypoxia. The enlarged placenta is associated with maternal toxaemia.

The blood film of these infants shows marked anisopoikilocytosis, hypochromic macrocytes and nucleated RBCs in peripheral blood. The Hb is generally 3-10g/dℓ with over 80% Hb Bart's (γ_4), as first described by Lie-Injo and Jo (1960). The remainder of the haemoglobin is HbH and Hb Portland, which is responsible for *in utero* survival.

1.7.1.5 Acquired α thalassaemia

A rare form of acquired α thalassaemia HbH disease has been described in a myeloproliferative disorder, which occurs in elderly males and often progresses to leukaemia. There is a normal α gene structure, but a severe reduction in αmRNA from both chromosomes, due to an unknown *trans*-effect affecting transcription (Anagnou *et al.* 1983, Higgs *et al.* 1983, Higgs *et al.* 1989).

1.7.1.6 α thalassaemia/mental retardation

Individuals have been described with α thalassaemia, associated with mental retardation and other developmental abnormalities. In these cases neither of the parents appears to carry a severe α thalassaemia determinant (Weatherall *et al.* 1981). Two syndromes have now been delineated.

In the first syndrome, the probands have haematological abnormalities, together with mild to moderate handicap and mild dysmorphic features. They have a *de novo* germ-line

deletion of one chromosome 16p, variable in extent but limited to 16p13.3. The deletions remove more than 1Mb of DNA including the α cluster and probably another locus, resulting in mental retardation. The individuals have generally inherited a mild α thalassaemia determinant from one parent on the second chromosome. This thus constitutes a microdeletion syndrome (Wilkie *et al.* 1990a).

In the second syndrome, individuals have milder forms of α thalassaemia, profound mental handicap and dysmorphic features (Wilkie *et al.* 1990b). It was initially shown to be X-linked from family studies (Harvey *et al.* 1990, Cole *et al.* 1991, Donnai *et al.* 1991, Gibbons *et al.* 1991, Wilkie *et al.* 1991b,c), and has now been localised to Xq12-21.31 (Gibbons *et al.* 1992). The proposed locus may encode a *trans*-factor involved in α globin expression (Higgs 1993a), though the mechanism by which the locus causes mental retardation, dysmorphism and down-regulation of the α globin genes is still uncertain (Gibbons *et al.* 1992).

1.7.2 ß thalassaemia

As there are only two ß genes per genome, an individual can be heterozygous or homozygous for a ß thalassaemia determinant, resulting in thalassaemia minor or thalassaemia major, respectively. As the δ gene only produces about one-fiftieth of the amount of polypeptide produced by the ß gene, deficiencies of the δ gene product are not important clinically.

The clinical features are reviewed in Weatherall and Clegg 1981, Kazazian 1990, Weatherall 1991a, and the management in Fosburg and Nathan 1990, Davies and Wonke 1991.

1.7.2.1 Thalassaemia minor

ß thalassaemia heterozygotes are clinically normal, though they do have mild ineffective erythropoiesis which may result in moderate anaemia during periods of stress, such as pregnancy or severe infection. Excess α chains are removed by the red blood cell proteolytic system, and thus do not accumulate. Individuals typically have an Hb greater

than 11g/dℓ, together with an elevated red blood cell count, an MCV of 55-75fℓ and an MCH of 20-25pg. The characteristic finding is an elevated HbA_2 level of 3.5-7%. About half the individuals have an increased HbF of 1-3%.

Atypical ß thalassaemia heterozygotes with normal HbA_2 levels and/or normal haematological indices have been reported, and are called 'silent carriers' (reviewed in Pirastu *et al.* 1990).

1.7.2.2 Thalassaemia major

Individuals with two defective ß globin genes typically have thalassaemia major, though there are marked variations in severity. Because so many different ß thalassaemia alleles occur in each geographic region, many individuals with thalassaemia major carry two different alleles and are genetic compound heterozygotes. True homozygotes are in the minority, except perhaps in groups with high rates of consanguineous marriages. The genotype is important in determining the phenotype, though the extent of this is still to be fully defined. In addition, modifiers such as coinherited α thalassaemia or determinants for increased HbF may alter the phenotype.

The disease is characterised by anaemia, ineffective erythropoiesis, haemolysis and reduced haemoglobinisation of cells. The full clinical syndrome described below is rarely seen nowadays in children on an adequate transfusion programme, the mainstay of treatment in most countries.

Infants are well at birth and only present within the first year of life with progressive anaemia, as the neonatal γ to ß switch occurs and the lack of ß chain synthesis becomes significant. They may also show failure to thrive, poor weight gain, irritability as well as fever, diarrhoea and vomiting.

The excess α chains form tetramers which are insoluble and precipitate, damaging the cell membranes of the developing red blood cells in the bone marrow, and resulting in ineffective erythropoiesis. The excess haemoglobin subunits generate free oxygen radicals and form Heinz bodies which bind to membrane proteins causing early red cell death

(Schrier *et al.* 1989, Shinar and Rachmilewitz 1990). A large proportion of developing erythroblasts are destroyed in the marrow, those that are released being destroyed by the spleen. This is in contrast to α thalassaemia where $ß_4$ and γ_4 homotetramers are soluble and only precipitate as the red cells age.

Ineffective erythropoiesis leads to anaemia and a compensatory expansion of bone marrow to 15-30 times normal, resulting in deformity of the skull and face bones, porosity and fragility of the long bones and stunted growth. Apart from pathological fractures, poor development and wasting result because of the diversion of calories to the marrow. The massive cell turnover can also cause secondary hyperuricaemia and folate deficiency. In addition, the reticuloendothelial system becomes overloaded, hepatosplenomegaly and hypersplenism result, causing increasing anaemia, thrombocytopenia and neutropenia. Splenectomy may be required to control these problems. The overactive bone marrow also enhances gastrointestinal iron absorption resulting in haemosiderosis.

At presentation, Hb levels are in the 2-3g/dℓ range, with an increased level of HbF ranging from 10-100%, and variable HbA_2 levels. In $ß^0$ thalassaemia no HbA is produced. The cells show marked anisopoikilocytosis, hypochromia, target cell formation and basophilic stippling. Nucleated red cell precursors are seen in the peripheral blood.

Without treatment, children with thalassaemia major generally die before the age of five years. Life expectancy has increased considerably with treatment (despite its problems).

1.7.2.2.1 Treatment of thalassaemia major

The treatment consists of blood transfusions every three to four weeks, sufficient to maintain a pretransfusion Hb of 10-12g/dℓ, with which children grow and develop normally until the end of the first decade when the effects of iron overload, mainly due to hypertransfusion, appear. Hypertransfusion not only prevents anaemia, but also suppresses the endogenous ineffective erythropoiesis, which leads to many of the thalassaemia complications. It is associated, however, with the problems of chronic transfusion therapy, including transmission of viral infections and antigen sensitisation.

Iron causes dose-related damage to the liver, endocrine organs and heart. Skin hyperpigmentation and 'late' growth failure at age 7-8 years result, followed by pubertal failure, insulin dependent diabetes, hypothyroidism, hypoparathyroidism and finally death from cardiac failure or conduction abnormalities in the teens or twenties. Iron overload is reduced by daily chelation with desferrioxamine, the only drug available for iron chelation. It is only minimally absorbed and therefore has to be given parenterally, generally subcutaneously, delivered by continuous infusion over 8-12 hours, using a portable pump. (Figure 1.3 shows one of the oldest ß thalassaemia patients in South Africa with his new infusion pump). This is responsible for the major clinical problem with desferrioxamine - non-compliance. There have been some promising reports on oral iron chelators, though none has yet been proven safe and effective (Porter 1989, Olivieri *et al.* 1990, Töndury *et al.* 1990). Iron chelation provides sufficient iron binding capacity to bind excess intracellular mobile and extracelluar free iron and also aims to create a negative iron balance. Though relatively non-toxic, desferrioxamine can cause toxic effects if the drug is started when the iron burden is very low. Its impact on long-term survival, particularly when started at an early age, is as yet uncertain, but it is thought that daily iron chelation protects thalassaemics from the lethal complications of iron overload for at least two decades. It is still uncertain whether it prevents growth retardation and delayed sexual development, but the results are promising (Bronspiegel-Weintrob *et al.* 1990, Ehlers *et al.* 1991).

The life-long treatment of thalassaemia with medical therapy provides a good quality of life and high survival for 20-30 years, but requires compliance and is enormously expensive. Bone marrow transplantation, if successful, can offer the chance of cure and perhaps a normal life expectancy and is cheaper overall than medical therapy. Bone marrow transplantation is a difficult choice for family and physician alike. Many patients are ineligible because of a lack of HLA-matched donors. Recent studies have shown a disease-free survival of 94% for the procedure in younger patients on optimal treatment, with low iron load and fewer transfusions. In older patients, however, with liver fibrosis or enlargement and inadequate chelation, the figure drops, though it is improving with newer preparative regimens pretransplantation (Lucarelli *et al.* 1990, 1992, 1993). Morbidity and mortality are due to infections, cardiac problems related to iron deposition, graft versus host disease, which is increased because of prior sensitisation, and failure of

FIGURE 1.3 The oldest South African Asian Indian thalassaemia major patient (aged 28 years) with his new pump for desferrioxamine infusion.

engraftment with persistence of the thalassaemia. In addition, late complications including malignancy and growth retardation due to the cytotoxic drugs or radiation preparation for transplant may occur.

1.7.2.3 Thalassaemia intermedia

Thalassaemia intermedia is ill-defined, referring to patients with a clinical syndrome more severe than thalassaemia minor but milder than transfusion-dependent thalassaemia major. Thus the spectrum of disease is wide, ranging from severe types requiring transfusion to patients who are asymptomatic and have a normal haemoglobin, except under stressful conditions. The molecular basis is varied and includes homozygous ß thalassaemia, due to mild ß thalassaemia mutations (such as those in the promoter) or with coinherited α thalassaemia (reducing the chain imbalance) or co-inherited HPFH. Homozygous δß thalassaemia may also be responsible, as may heterozygous ß thalassaemia, either where the mutation is dominant or where there is a coinherited $\alpha\alpha\alpha$ chromosome (reviewed in Wainscoat *et al.* 1987).

In the dominant thalassaemias, the elongated chains aggregate and precipitate, resulting in imbalanced synthesis and ineffective erythropoiesis. In the milder forms, tetramers which fall apart in peripheral RBCs result in the picture of haemolytic anaemia. Unstable variants are rapidly degraded (Thein 1992).

1.7.3 Sickle cell anaemia

The sickle cell mutation can be present in an individual in heterozygous or homozygous form, giving rise to sickle cell trait or sickle cell anaemia, respectively. Since the mutation is in the ß globin gene, compound heterozygosity for $ß^s$ and ß thalassaemia or other structurally abnormal ß globin alleles may occur.

Many authors have reviewed these disorders (Weatherall and Clegg 1981, Nagel and Ranney 1990, Powars *et al.* 1990, Beutler 1991, Davies and Wonke 1991) and only the main features will be discussed.

1.7.3.1 Sickle cell trait

Individuals with sickle cell trait are clinically normal, although under conditions of extreme hypoxia sickling of the cells may occur. Individuals have a normal full blood count and red cell morphology. On haemoglobin electrophoresis, the cells contain 55-65% HbA and 35-45% HbS, with normal amounts of HbA_2 and HbF. α globin chains have an increased affinity for normal ß chains, thus relatively more of the abnormal ß chains are destroyed, resulting in less than 50% abnormal haemoglobin.

The number of α genes present in individuals with sickle cell trait determines about 90% of the variance in the percentage of HbS. The latter is trimodally distributed, depending on whether there are four, three, or two α genes present in heterozygotes (Brittenham 1977, Brittenham *et al*. 1979, 1980).

1.7.3.2 Sickle cell homozygotes

Sickle cell anaemia is a major risk to health with high morbidity and mortality at all ages. Like ß thalassaemia it typically presents between six months and two years of age, as the production of HbF decreases. At presentation, children have an Hb of 6-9g/dℓ, 90-95% HbS, no HbA and a small amount of HbF on electrophoresis. Sickled cells are seen on the blood film.

Infection, hypoxia, dehydration, cold and exhaustion can all precipitate sickling, which results in vaso-occlusive crises. The effects can range from mild pain, to infarction of major organs, to death. Aplastic crises also occur, often as a result of infection, but are usually self-limiting. Sequestration crises can result in up to one-third of the blood volume being trapped in a rapidly enlarging spleen.

In childhood the most dangerous complication is infection, particularly fulminant pneumococcal septicaemia. Oral penicillin prophylaxis is thus important, as is malaria prophylaxis when visiting malaria areas. Painful vaso-occlusive crises and acute splenic sequestration occur due to obstruction of the microvasculature. In adolescents, cerebrovascular accidents are an important cause of morbidity and mortality. Chronic

organ damage (including avascular necrosis of the femoral head) occurs from repeated sickling crises. In adults, deaths are frequently related to respiratory complications.

Although all sickle cell anaemia is due to homozygosity for the β^S mutation, marked differences in clinical, haematological and genetic features have been noted, linked largely to the different β^S haplotypes. The primary effect of the mutation is modified by epistatic effects, which can come from linked or unlinked genes. Environmental effects may also be important (Nagel 1993).

The Arab-Indian and Senegal haplotypes (with the presence of an *Xmn*I site 5' to $^G\gamma$) are associated with higher fetal haemoglobin levels and the preservation of the newborn preponderance of $^G\gamma$ HbF. Adult patients who are homozygous for these two haplotypes have HbF levels well above 20%. The Bantu haplotype is associated with adult type low levels of $^G\gamma$, and HbF is often less than 5%. The HbF with the Benin haplotype is intermediate, at 5-15%. HbF levels above 10% appear to protect against major organ failure including stroke and avascular necrosis, whereas a HbF above 20% protects against painful crises and pulmonary complications (Powars *et al.* 1984).

Sequence changes near the γ genes in linkage disequilibrium with the different haplotypes may be important in HbF expression (Dimovski *et al.* 1991, Lanclos *et al.* 1991). In addition, sequence changes in the ßLCR region may affect the binding of *trans*-acting factors produced under conditions of erythropoietic stress (Öner *et al.* 1992).

In general, the Arab-Indian haplotype is associated with the mildest disease, followed by Senegal, Benin and Bantu with the most severe (Nagel *et al.* 1985, 1987). Mild disease is reflected in lower reticulocyte counts, lower LDH, lower percentage dense cells and higher Hb levels, as well as the high HbF levels. Coinherited α thalassaemia has also been shown to reduce haemolysis, raise the Hb and lower the reticulocyte count and the bilirubin level (Embury *et al.* 1982, Higgs *et al.* 1982)

1.7.4 Other haemoglobinopathies

Homozygotes for HbE and HbC have mild haemolytic disease with abnormal red cell

morphology, while significant illness occurs in compound heterozygotes for $ß^S/ß^C$, $ß^S/ß$ thalassaemia and $ß^0/ß$ thalassaemia. Many other combinations of alleles exist, with enormous variation in severity. These conditions are reviewed in Weatherall and Clegg 1981, Beutler 1991.

1.7.5. Future treatment prospects for the haemoglobinopathies

Bone marrow transplantation offers the possibility of cure for the severe haemoglobinopathies and is likely to become an important mode of therapy as the risks associated with the procedure decrease.

Prospects for improvements in the medical treatment of thalassaemia include the development of oral iron chelators. Pharmacological agents that may increase γ gene expression by hypomethylating cytosine residues near the γ genes are potentially useful in ß thalassaemia and sickle cell anaemia.

The insertion of normal globin genes into bone marrow stem cells offers another possibility of cure. This requires isolation of the gene and its regulatory sequences as well as detailed knowledge of their interactions, so that expression at normal level can be achieved. At present, these techniques are associated with problems of low transformation frequency, regulation of tissue-specificity, developmental stage-specificity and transcription rate regulation. These problems are particularly important for the globin genes which require tight regulation.

Many methods of somatic gene therapy have been proposed, including the direct insertion of genes using calcium uptake, direct microinjection into the nucleus, electroporation, genome delivery using retroviruses, targeted modification of genes using site-directed mutagenesis or homologous recombination. Many of the problems of somatic gene therapy are related to efficiency. Thus, if stem cells could be expanded in culture while forestalling differentiation, major hurdles would be overcome. Growth factors offer this possibility.

The transgenic approach involves the introduction of DNA by microinjection or retroviral

transfection integrated into chromosomal DNA so it is carried in germ cells and transmitted to subsequent generations. Although this is a useful model for studying genes and regulation, it is not applicable to human gene therapy because of ethical problems associated with germ line manipulation.

The future treatment prospects for the haemoglobinopathies have been reviewed in Bank *et al.* 1988, Fosburg and Nathan 1990.

1.8 Prevention of the haemoglobinopathies

Although treatment of the severe haemoglobinopathies has resulted in a marked increase in the life span of affected individuals, it requires considerable financial resources and continuous patient co-operation. Prospects for cure of the disease are still limited. In addition, the conditions are generally inherited in an autosomal recessive manner and thus parents, who are obligate carriers, have a 25% risk of an affected child with each pregnancy. A few new mutations have been reported (Chehab *et al.* 1986, Kazazian *et al.* 1986a, Chehab *et al.* 1989).

Thus counselling for 'at-risk' couples should include the natural history of disease, problems of living with the disease, improved therapies now available and improved prognosis. As there is no cure for this disease, the options of prenatal diagnosis and selective abortion in order to 'prevent' the disease should be discussed.

For a comprehensive prenatal diagnosis programme, the first prerequisite is public education. Preventive programmes rely heavily on education, counselling and screening and detection of couples before the birth of an affected child (Cao *et al.* 1984, Angastiniotis *et al.* 1986, Kazazian 1990). A reliable screening programme is also essential. Screening includes the obtaining of an adequate family history and the performance of haematological investigations in individuals from 'high-risk' ethnic groups. In countries with high frequencies of haemoglobin disorders routine screening should be available.

1.8.1 Techniques used in prenatal diagnosis of the haemoglobinopathies

Prenatal diagnosis has evolved from globin chain biosynthesis through DNA analysis by Southern blotting to polymerase chain reaction (PCR) based techniques. Prenatal diagnosis was initially set up for ß thalassaemia using fetal blood sampling and globin chain synthesis studies, the first procedure being done by Kan *et al.* (1975b). This procedure alone reduced the incidence of the disease in the Mediterranean considerably (WHO Working Group 1982, Alter 1990). Fetal blood sampling was initially done by placental aspiration with direct visualisation via fetoscopy, with an overall fetal loss of 6-7%. A mixture of fetal and maternal blood was obtained and had to be separated. Though fetal blood sampling has largely been replaced by DNA techniques, some indications for its use still remain. Currently, cordocentesis (performed at 18-20 weeks of gestation) is used, where a pure fetal blood sample is taken from the umbilical cord under ultrasound guidance, with a fetal loss rate of no more than 1% (Weatherall 1991b).

There have also been advances in the methods of analysis of fetal blood. Initially, globin chain synthesis was done with labelled reticulocytes, followed by CMC (carboxy methyl cellulose) chromatography in which the radioactivity incorporated into each chain was measured, and the ratio of ß to γ chains was measured. The error rate was small, mainly due to maternal cell contamination or poor chromatographic technique (Weatherall 1991b). HPLC allowed separation of smaller amounts of chains in hours rather than days (Congote *et al.* 1979). Nowadays, with pure fetal blood samples, isoelectric focusing can be done and no biosynthesis is required.

Fetal blood analysis has been largely replaced by DNA analysis, in which any fetal cells can be used, not those specifically expressing the protein. DNA analysis was initially done after amniocentesis performed at 16-20 weeks of pregnancy. This procedure has the advantage of a low fetal loss rate (0.5-1%), but the disadvantage of a late second trimester prenatal diagnosis result and possible termination of pregnancy. Amniocytes are generally cultured to obtain sufficient DNA for analysis, although with PCR analysis this is no longer necessary. DNA analysis was first performed on amniocytes for α thalassaemia by DNA-DNA solution hybridisation and required 100-200μg of DNA (Kan *et al.* 1976). Restriction endonuclease and Southern blot analysis used for sickle cell anaemia in 1978

(Kan and Dozy 1978b) and ß thalassaemia in 1980 (Kazazian *et al.* 1980, Little *et al.* 1980) only required 10-20μg of DNA. Chorionic villus biopsy (CVS) material was shown to be a reliable and better source of fetal DNA for molecular analysis and many couples soon opted for this technique for prenatal diagnosis of haemoglobinopathies because it was performed at 9-11 weeks gestation, when termination is safer and involves less psychological trauma, though the fetal loss rate is slightly higher than that for amniocentesis (Williamson *et al.* 1981, Old *et al.* 1982, 1986). Precautions must be taken to dissect the maternal decidua adequately from the fetal tissue, particularly where PCR is to be used for the prenatal diagnosis (Old and Ludlam 1991).

Initially DNA analysis was carried out using Southern blotting and linkage analysis using RFLPs. These require the establishment of a linkage phase in the heterozygous parents to determine which allele is associated with ß thalassaemia, so that the ß thalassaemia status of the fetus can be determined by observing its RFLP types (reviewed in Old and Ludlam 1991).

Some mutations were detectable with Southern blotting and restriction enzyme digests or allele-specific oligonucleotides. Recently the molecular analysis of fetal DNA has been revolutionised by the discovery of PCR and particularly thermostable *Taq* polymerase, which has made analysis of fetal DNA quicker, simpler and easier, with much smaller amounts of DNA being required (Saiki *et al.* 1985, 1986, Mullis and Faloona 1987, Saiki *et al.* 1988a,b). It is possible to obtain a diagnosis within 24 hours, often without the use of radioactivity, as sufficient DNA results from the PCR amplification to visualise the products on a stained gel. The technique is so sensitive that it is possible to analyse DNA sequences in single cells (Li *et al.* 1988). PCR has allowed for the development of new techniques for DNA and particularly mutation analysis including the chemical cleavage method (Dianzani *et al.* 1991), ARMS (amplification refractory mutation system) (Old *et al.* 1990) and multiplex ARMS (Fortina *et al.* 1992), competitive priming using fluorescent labelled primers (Chehab and Kan 1989, 1990), digestion with restriction enzymes which recognise natural or artificially created restriction sites (Chang *et al.* 1992), denaturing gradient gel electrophoresis (Cai and Kan 1990, Losekoot *et al.* 1990), heteroduplex analysis (Cai *et al.* 1991) and direct sequencing of amplified DNA (Wong *et al.* 1987). In addition, RFLP analysis (Kulozik *et al.* 1988b, Semenza *et al.* 1989, Old

et al. 1990) and mutation analysis using oligonucleotides or enzyme digestion (Saiki *et al.* 1985, 1986) are also possible and more rapid with PCR.

Newer techniques have also increased the number of families to whom prenatal diagnosis can be offered. In 1978 prenatal diagnosis using the *Hpa*I RFLP for sickle cell anaemia provided a diagnosis in 60% of 'at-risk' pregnancies (Kan and Dozy 1978a). Similarly, for ß thalassaemia, prenatal diagnosis was possible in 75% of families in 1980 (Kazazian *et al.* 1980) and progressed to 92% in some ethnic groups (Wainscoat *et al.* 1986b), but required detailed family studies and the availability of critical family members. All linkage analysis is subject to error, albeit small, due to non-paternity or recombination between the marker and the allele of interest and non-paternity. Newer techniques using PCR for haplotyping single individuals may obviate some of the problems of obtaining multiple family members for linked marker studies (Ruano and Kidd 1989, Ruano *et al.* 1990, Sarkar and Sommer 1991).

Once it was recognised that the substitution in codon 6 which causes sickle cell disease destroys a restriction endonuclease site, the $ß^S$ allele could be detected directly, making prenatal diagnosis theoretically possible in 100% of 'at-risk' pregnancies, without the requirement for family studies (Chang and Kan 1981, Geever *et al.* 1981, Orkin *et al.* 1982c, Wilson *et al.* 1982). The $ß^S$ mutation could also be detected directly by oligonucleotide hybridisation (Conner *et al.* 1983). The majority of ß thalassaemia mutations have also been characterised. In any ethnic group screening for a small number of mutations by various techniques should reveal the parents' mutations in a large proportion of the cases. This allows for increased accuracy of prenatal diagnosis and decreased benchwork obviating the need for extended family studies.

1.8.2 Practical approach to prenatal diagnosis

Up to 1990, over 20 000 fetuses in 35 centres had been tested for haemoglobinopathies, 13 000 by blood sampling, the rest by DNA analysis (Alter 1990). An increasing number of prenatal diagnoses for haemoglobinopathies are now being done by DNA analysis; over 100 per month (Alter 1990). The numbers may be even higher as some prenatal diagnosis centres are not included in the analysis. Such prenatal diagnostic programmes in the

Mediterranean and for Cypriots in London have reduced thalassaemia births profoundly (Modell *et al.* 1984, Cao *et al.* 1984, Kuliev 1986, Loukopoulos *et al.* 1990).

For ß thalassaemia it is important to determine the common mutations in a particular ethnic group. Once carrier parents are ascertained, an attempt is made to identify the mutations they carry. If the mutation(s) remain(s) unknown, haplotype analysis can be used. If, in rare cases, this is not possible due to late referral or unavailability of family members or uninformative RFLP systems, fetal blood sampling can be used for prenatal diagnosis. If possible, it is recommended that both direct mutation analysis and indirect RFLP analysis be performed on all prenatal samples to minimise the risks of laboratory errors (Old and Ludlam 1991). With newer techniques ß thalassaemia results can take a few hours and it is generally less than one week from the time of a CVS to a result. The errors in analysis with such techniques are estimated to be 0.5% or less (Alter 1990).

Similar principles can be applied to sickle cell anaemia, though it is obviously unnecessary to identify the parents mutations once they are known heterozygotes. Currently prenatal diagnosis of sickle cell anaemia is achieved using PCR together with hybridisation to oligonucleotides (Saiki *et al.* 1988a), digestion (Saiki *et al.* 1985, Chehab *et al.* 1987a), ARMS (Old and Ludlam 1991) or colour complementation (Chehab and Kan 1990). The other structurally abnormal haemoglobins can also be detected using PCR techniques (Old and Ludlam 1991).

Prenatal diagnosis of α thalassaemia is mostly requested where both parents carry the -- chromosome and thus have a 25% risk of a fetus with Hb Bart's hydrops fetalis, but may be requested for HbH disease. In general, this is still achieved using Southern blotting and α and ζ probes, the latter being particularly useful to detect deletions. The α probe does not hybridise if the α genes are deleted and thus there is no signal. This procedure may take 10 days to three weeks to obtain a result. More recently PCR has been used to identify α thalassaemia. Initially, a region within the α gene was amplified and a second gene used as a control. A diagnosis of α^0 thalassaemia was made if no product was seen from the intragenic α primers (Chehab *et al.* 1987a, Lebo *et al.* 1990). This system is somewhat problematic as heterozygotes cannot be detected as the technique is non-quantitative and, in addition, results based on failure of amplification are

notoriously unreliable. A newer PCR method using amplification across the breakpoint with visualisation of a fragment specific for the deletion has been reported for detection of the $--^{SEA}$ and $--^{MED}$ chromosomes (Bowden *et al.* 1992, Ko *et al.* 1992), $-\alpha^{3.7}$ and $\alpha\alpha\alpha^{anti3.7}$ (Dodé *et al.* 1992). The technique has been used for prenatal diagnosis of a fetus 'at-risk' for compound heterozygosity for $--^{SEA}$ and Hb Constant Spring, the latter identified by an altered *Mse*I cleavage site (Ko *et al.* 1993).

New approaches to prenatal diagnosis, based on the ability of PCR to detect mutations in a single DNA molecule, have recently been reported (Li *et al.* 1988). The embryo can be tested prior to implantation, either by testing the first polar body and selection of oocytes without the defect (Monk and Holding 1990) or, after *in vitro* fertilisation, testing biopsied cells from the three-day-old embryo (eight-cell or blastocyst stage) (Adinolfi and Polani 1989, Holding and Monk 1989, Handyside *et al.* 1989, Varawalla *et al.* 1991a). Although most still consider these techniques experimental and contamination appears to be a major problem (Varawalla *et al.* 1991a), others feel that clinical preimplantation diagnosis by amplification of single copy sequences from a single cell of an eight-cell embryo could be carried out safely and reliably (Monk *et al.* 1993). One can also test for mutations in nucleated fetal cells in the maternal blood. This has been demonstrated with PCR amplification of the Y chromosome in male fetuses (Lo *et al.* 1989, Lo *et al.* 1993), and Hb Lepore at 8-10 weeks (Camaschella *et al.* 1990b). For recessive diseases this approach is problematic as only the paternal allele can be detected, and only if it differs from the maternal one. DNA ligase that has the potential to amplify DNA and discriminate single base substitution simultaneously is also potentially useful (Barany 1991).

The techniques and approaches to prenatal diagnosis have been reviewed recently by Boehm and Kazazian 1989, Old and Ludlam 1991, Weatherall 1991a,b.

1.9 Population genetics

The haemoglobinopathies are all common single gene disorders in humans, their distributions being largely limited to tropical and subtropical regions. In addition, multiple mutations have arisen independently in different populations and achieved high frequencies

by selection. These observations led to the hypothesis that such variants confer an increased fitness on heterozygous individuals living in areas where the parasite *P. falciparum* is endemic. It was Haldane in 1949 who first suggested that thalassaemia might have reached its high frequency in tropical areas because the heterozygotes are protected against malaria. The so-called 'malaria hypothesis' is supported by microepidemiological data and by *in vitro* experiments.

1.9.1 α thalassaemia and α globin cluster rearrangements

The α thalassaemias are frequent on a worldwide scale. Most individuals originate from tropical or subtropical regions where the high frequency is thought to have resulted from a selective advantage afforded to carriers of α thalassaemia in the presence of endemic *P. falciparum* malaria (Flint *et al.* 1986).

The -α thalassaemia chromosome has a worldwide distribution, with frequencies ranging from 20-90% in different African, Mediterranean, Middle East, Indian, south-east Asian and Melanesian populations (reviewed in Weatherall *et al.* 1988, Higgs *et al.* 1989, Kazazian 1990). The $-\alpha^{3.7}$ chromosome seems to be the most widely distributed, with the $-\alpha^{3.7I}$ type occurring at varying frequencies in most populations examined. Its wide distribution and association with multiple haplotypes suggest that it has arisen on more than one occasion. The $-\alpha^{3.7II}$ type has been reported in individuals from Jamaica, south-east Asia, India, Nepal and the Mediterranean (Higgs *et al.* 1984, Fodde *et al.* 1988, Modiano *et al.* 1991). The $-\alpha^{3.7III}$ chromosome, on the other hand, may have had a single origin, probably in or near Melanesia, as it is restricted mainly to Polynesia and Melanesia and is associated with a single haplotype (Higgs *et al.* 1984, Hill *et al.* 1985b, Hill *et al.* 1989). The $-\alpha^{4.2}$ variety occurs sporadically in many groups. It is relatively rare in Africa and the Mediterranean, but more common in south-east Asia (Embury *et al.* 1980a) and the Middle East (Trent *et al.* 1981c, El-Hazmi 1986). It almost always occurs at a lower frequency than the $-\alpha^{3.7}$ chromosome, except in Papua New Guinea where the $-\alpha^{4.2}$ occurs at a higher frequency and is thought to be going to fixation in the northern coastal regions where it is found in over 80% of the population (Oppenheimer *et al.* 1984, Yenchitsomanus *et al.* 1985).

In the south-west Pacific micro-epidemiological data suggest that α thalassaemia is likely to have been selected through its interaction with malaria. The frequency of α thalassaemia and malaria correlates with latitude and altitude (Oppenheimer *et al.* 1984, Flint *et al.* 1986, Hill 1986). It has not been possible to define, however, the mechanism whereby thalassaemic red cells protect against malarial parasite growth, but the protective effect may be related to enhanced immune recognition and clearance of parasitised cells [Luzzi *et al.* (1991) *J Exp Med* 173, 785-91]. Although parasite growth is inhibited by HbH cells, *in vitro* studies in cells lacking one or two α genes do not show abnormalities of invasion or growth unless the cells are under unusual oxidant stress (Ifediba *et al.* 1985), suggesting that the increase in fitness may be relatively small (Higgs *et al.* 1989).

The -- chromosome is generally rarer than the -α (reviewed in Weatherall *et al.* 1988, Higgs *et al.* 1989, Kazazian 1990). The most prevalent is the south-east Asian type, which has a frequency of 0.1-0.2 in the region. A second type occurs in the Mediterranean, and a third makes up 30% of the α thalassaemia genes in Filipinos. It is extremely rare in Africa and the Middle East (Dozy *et al.* 1979, Higgs *et al.* 1979, Orkin *et al.* 1979b, Embury *et al.* 1980b). In view of their limited geographical distributions, it is thought that each of these mutations only arose once. This is supported by haplotype analysis.

As HbH disease and Haemoglobin Bart's hydrops fetalis generally require the presence of the -- chromosome, they are found predominantly in south-east Asians and Filipinos, and only occasionally in Mediterraneans. Similarly, they are extremely rare diseases in Africa and the Middle East. The α chain termination mutant, Constant Spring, found in about 5% of individuals in south-east Asia and Thailand, contributes to the prevalence of HbH disease.

The ααα chromosome, the reciprocal of the -α chromosome, occurs at frequencies of 0.5-2% in most populations studied and, in general, has not reached high frequencies probably because there has been no selection for it (reviewed in Higgs and Weatherall 1983, Weatherall *et al.* 1988). However, in a number of groups, including Greek Cypriots (Goossens *et al.* 1980), Polynesians (Lie-Injo *et al.* 1985, Trent *et al.* 1985, 1986) and Saudi Arabians (Trent *et al.* 1981b), the chromosome has reached relatively higher frequencies, probably due to local founder effect and genetic drift.

Chromosomes with a single ζ gene are common in west Africa and in Polynesian Niueans and occur sporadically in other populations (Winichagoon *et al.* 1982, Rappaport *et al.* 1984, Felice *et al.* 1986, Higgs *et al.* 1989, Trent *et al.* 1991). They are found in association with different haplotypes and different numbers of α genes, suggesting they may have arisen more than once (reviewed in Weatherall *et al.* 1988, Higgs *et al.* 1989).

The triplicated ζ chromosome, which has the arrangement ζ2-ψζ1-ψζ1, occurs at relatively high frequencies in south-east Asia (Winichagoon *et al.* 1982), China (Chan *et al.* 1986), Micronesia (O'Shaughnessy *et al.* 1990) and Polynesia (Trent *et al.* 1986), though isolated cases occur in most populations (reviewed in Higgs *et al.* 1989). It seems to have arisen by an unusual interchromosomal recombination event and all the chromosomes studied from south-east Asia and the Pacific appear to have a single common origin (Hill *et al.* 1985a, Hill *et al.* 1987a).

1.9.2 ß thalassaemia and ß globin cluster rearrangements

ß thalassaemia genes attain high frequencies in many parts of the world including the Mediterranean Basin, the Middle East, parts of India and south-east Asia. The disease is less common in Africa, occurring only in isolated pockets in west Africa (notably in Liberia) and in parts of north Africa. It occurs sporadically, however, in all racial groups (reviewed in Kazazian 1990, Weatherall 1991a). Its major distribution again follows that of endemic *P. falciparum*, as carriers of ß thalassaemia have less morbidity when infected with malaria and thus positive selection has acted on the genes.

ß globin haplotypes have been used in an attempt to clarify the origins of ß thalassaemia mutations. The large number of ß thalassaemia mutations suggests that multiple mutations together with selection have resulted in high frequencies of ß thalassaemia. In general, in a population, the alleles with higher gene frequency are thought to have originated earlier than the rarer alleles. Even the high frequency ß thalassaemia alleles seem to postdate the divergence of human races as each ethnic group and/or racial group has its own battery of mutations, each found in strong association with only one or two haplotypes (Orkin *et al.* 1982a, Kazazian *et al.* 1984b, Hill and Wainscoat 1986). When the same mutation is

found in two different groups on strikingly different chromosomal backgrounds, two or more independent origins are likely (Huang *et al.* 1986, Wong *et al.* 1986) Not uncommonly the same mutation may be seen on different haplotypes in the same population. In general, they are consistent with crossover 5′ to the gene, within the region of increased recombination (Kazazian *et al.* 1984b, Antonarakis *et al.* 1984, 1985). However, the same mutation may be found in a population on two or more different frameworks. This is best explained by recurrent mutation, although gene conversion as a means of transmitting mutations cannot be discounted (Antonarakis *et al.* 1985). Those mutations which occur on multiple frameworks in one ethnic group all appear to be in coding regions. This may suggest a role for spliced RNA transcripts in certain rare gene conversions (Antonarakis *et al.* 1985). Thus, mutation/haplotype associations appear to be due to a combination of gene migration, interallelic gene conversion and recurrent mutation (Wong *et al.* 1986).

Micro-epidemiological and *in vitro* studies have verified the 'malaria hypothesis' in the case of ß thalassaemia. In Sardinia (Siniscalco *et al.* 1966) and Melanesia (Hill *et al.* 1988), ß thalassaemia was shown to be less common in mountainous areas where malaria transmission is low, suggesting that the incidence of ß thalassaemia was increased in the lowlands due to protection against malaria. *In vitro*, thalassaemic red cells have been shown to inhibit parasite growth (Brockelman *et al.* 1987).

Dominant ß thalassaemias are rare in all populations as they cause severe disease in heterozygotes and thus offer no selective advantage. δß thalassaemia occurs sporadically in many racial groups, as does Hb Lepore, though the latter occurs at increased frequency in central Italy and Yugoslavia (reviewed in Kazazian 1990).

γ Gene triplications have been detected in numerous populations at low frequency, though the few homozygotes detected were of Japanese descent (reviewed in Huisman 1987). The -γ arrangement has been observed in Mediterranean, Indian, Japanese and Chinese, but not in Black populations (Huisman *et al.* 1983, Hattori *et al.* 1986a). Chinese, Japanese, Mediterranean and Blacks have $^{A}\gamma$-$^{A}\gamma$ and $^{G}\gamma$-$^{G}\gamma$ rearrangements (Powers *et al.* 1984, Harano *et al.* 1985b, Hattori *et al.* 1986a, Shimizu *et al.* 1986).

1.9.3 The β^S gene

The β^S gene is found at high frequencies in Africa, especially in the sub-Saharan regions. It also occurs at lower frequencies in populations on the north coast of the Mediterranean, in the Arabian Peninsula, Iraq, Iran, India, Afghanistan and parts of the former USSR (reviewed in Nagel and Fleming 1992).

The existence of five different haplotypes in distinct geographical locations associated with the β^S gene has been seen as evidence for at least five independent origins or centres of expansion of the mutation. This is reinforced by recent sequencing data as well as the different clinical presentations of sickle cell anaemia in the different geographical locations, suggesting different epistatic influences associated with the different haplotypes (Nagel and Ranney 1990, Nagel and Fleming 1992, Flint *et al.* 1993a).

Rarer haplotypes do exist, probably as a result of crossover around the hotspot 5′ to the ß globin gene (Srinivas *et al.* 1988). The major haplotypes differ from each other both 5′ and 3′ to the ß gene and each exists in a distinct geographical area with little overlap. They are thus unlikely to have arisen by recombination (Pagnier *et al.* 1984, Nagel and Fleming 1992), but may have arisen by gene conversion in genetically isolated populations carrying a single β^S mutation on to the different haplotypes in Africa. The Asian Indian mutation is likely to have had a separate mutation, however (Livingstone 1989, Flint *et al.* 1993a).

Different mutations in the 5′ promoter and IVS2 sequences of the $^G\gamma$ and $^A\gamma$ genes have been shown to distinguish the β^S haplotypes. Further the Cameroon chromosome has an $^A\gamma^T$ gene. Though the variation does not appear to have a primary role in determining HbF levels, the high and low HbF chromosomes have different configurations. It is proposed that the sequences may be important in regulating expression *in vivo*, particularly under conditions of anaemic stress (Dimovski *et al.* 1991, Lanclos *et al.* 1991).

The configuration of the $(AT)_xT_y$ motif 0.5kb 5′ and the ATTTT variable repeat 1.5kb 5′ to the ß globin gene distinguish each of the haplotypes. There is additional nucleotide

variation in the 5' and intragenic regions of the ß globin gene, so that there are five different patterns corresponding to the five RFLP-defined haplotypes, supporting the theory of at least four $ß^S$ origins in Africa, and a fifth unicentric origin in Arab-India (Chebloune *et al.* 1988, Trabuchet *et al.* 1991a).

It has been suggested that the mutations arose or expanded coincident with malaria becoming endemic, when man progressed from a food-gathering to a food-producing way of life. Agriculture favours endemicity of malaria by creating water sources, which provide breeding areas for mosquitoes. As there are large numbers of humans, they displace other primates so that man becomes the preferred host for the parasite (Nagel and Ranney 1990, Nagel and Fleming 1992, Flint *et al.* 1993a).

The $ß^S$ gene appears to have three main centres in Africa: (i) between the Niger and Benue rivers, (ii) the Congo and (iii) Senegal, corresponding to the three main African haplotypes, Benin, Bantu or Central African Republic and Senegal, respectively. The haplotypes are virtually exclusive to each of the geographic areas, though each has expanded through diverse ethnic groups. The Benin and Senegal haplotypes are estimated to be 2 000-3 000 years old. The Benin haplotype has spread by gene flow to the north, south and east of the Mediterranean and to western Saudi Arabia. The Bantu haplotype has spread through the Bantu-speaking areas of equatorial and southern Africa and thus must have occurred before the Bantu expansion, earlier than 2 000 years ago (reviewed in Nagel and Labie 1989, Nagel and Ranney 1990, Nagel and Fleming 1992). The fourth, recently defined African haplotype, designated 'Cameroon', is limited to the Eton ethnic group from the Sanago river valley of Cameroon (Lapoumëroulie *et al.* 1992).

The Arab-Indian $ß^S$ haplotype is found in the eastern oases of Saudi Arabia and in India (Bakioglu *et al.* 1985, Wainscoat *et al.* 1985, Kulozik *et al.* 1986, Miller *et al.* 1986, 1987). In India it is found both in tribal populations living in malaria infested areas of northern India (Gujerat, Orissa and northern Andhra Pradesh), and in southern India (Tamil Nadu). This suggests that the tribal populations of India, now isolated and surrounded by mainstream Indians that have a lower frequency of $ß^S$, were previously in direct contact and subject to gene flow (Kar *et al.* 1986, 1987, Labie *et al.* 1989, Nagel and Labie 1989, Trabuchet *et al.* 1991a, Nagel and Fleming 1992).

The high frequency of β^S has been shown to be due to the higher relative fitness and survival rates of β^S heterozygotes compared to normal homozygotes, as this state affords partial protection against severe *P. falciparum* parasitaemia and its often fatal complications. The advantage is balanced against the decreased fitness of the sickle cell homozygote, reaching equilibrium in the β^S frequency. This is evidenced by the geographical distribution of the allele, population studies, hospital studies and *in vitro* experiments (reviewed in Nagel and Ranney 1990, Nagel and Fleming 1992). The β^S gene is found at polymorphic frequencies only in populations living in regions where *P. falciparum* is or was endemic. *In vitro* studies show that parasitised AS cells sickle more frequently than non-parasitised cells, possibly because the parasite reduces the cell pH, thus inducing sickling. The parasitised sickled cells are removed by the spleen and macrophages, resulting in a suicidal infection (Luzzatto *et al.* 1970, Roth *et al.* 1978, Luzzatto and Pinching (1990) *Blood Cells* 16, 340-7). Secondly, parasites that escape the first line of defence may find their intra-erythrocyte growth inhibited in AS cells during deep vascular schizogony. Recent *in vivo* experiments in β^S transgenic mice confirm that HbS is protective against malaria and implicate the spleen in this protection (Shear *et al.* 1993).

1.9.4 The β^E gene

The β^E allele is found with high frequency in Kampuchea, Laos, Thailand, Burma and the Kachari of India (reviewed in Nagel and Ranney 1990). It is associated with three haplotypes on two frameworks in south-east Asia, suggesting at least two origins (Antonarakis *et al.* 1982b, Hundrieser *et al.* 1988a). A third European origin has been proposed (Kazazian *et al.* 1984c).

The distribution is again related to areas of malaria endemicity. A moderate decrease in growth of *P. falciparum* occurs in EE cells and AE cells with oxidative stress, possibly related to the instability of HbE and its ability to generate free radicals (reviewed in Nagel and Ranney 1990).

1.9.5 The β^C gene

The β^C allele appears to have originated in a single locale on the west coast of Africa,

where it is observed. To date it is found associated with one common haplotype suggesting a single origin of the mutation. Two other rare haplotypes, probably result from 5′ crossover (Kan and Dozy 1980, Boehm *et al.* 1985, Trabuchet *et al.* 1991b).

1.10 The South African Asian Indians

1.10.1 The history of the Indian subcontinent

The history of the Indian subcontinent is complex. It has been reviewed in Encyclopaedia Britannica (1977). The peoples of India are largely the product of successive invasions that have swept into the country from prehistoric times and are broadly divided into two linguistic groups, Indo-Aryan and Dravidian, situated geographically in the north and south of the subcontinent, respectively. More than four-fifths of the population is Hindu, one-tenth Moslem, followed by Christians, Sikhs, Buddhists and Jains.

The Dravidian languages are the oldest, but it is uncertain whether the Dravidian-language-speaking Indians were autochthonous in southern India, though they are not thought to have been associated with the Harappan or Indus civilisation which thrived in north-west India from 2 300-1 750BC.

The Aryan languages were introduced by the Aryan or Indo-European peoples who invaded India across the Iranian plateau between 1 500 and 1 200BC and began the urbanisation of the Ganges valley. They introduced the Sanskrit language from which the Indo-Aryan languages are derived. Hindi developed from one of the spoken Sanskrit dialects or Prakrits and incorporated many of the words of the local inhabitants. Gujerati was also derived from Sanskrit, incorporating a language spoken in northern and western India, as well as Persian, Arabic, Turkish, Portuguese and English. Hinduism, the religion of four-fifths of India's population, evolved from Vedism, the religion of these Indo-Aryan people.

Moslem invasions of India began in the 8th century AD. The invaders overthrew the earlier Indo-Aryan rulers and controlled northern India until the 14th century when they in turn were overthrown by the Mughals (Moslems of Mongol origin from central Asia).

During this period many Hindus were converted to Islam. In the west and north-west of the subcontinent, the Arabic and Persian languages of the Moslems merged with Hindi and developed into Urdu, a language still spoken predominantly by Moslem Indians (Meer 1969).

1.10.2 The origins of the South African Indians

The origins of the South African Asian Indians have been traced to a number of discrete geographical regions in India, and they thus represent a subset of the total Indian population. Most descend from a relatively small founder population of approximately 150 000 individuals who entered South Africa between 1860 and 1911. The population has expanded rapidly, mainly due to natural increase, with a very small percentage of immigration, to the 1 000 000 Indians living in South Africa today, 90-95% of whom are urbanised (Department of Central Statistics 1990, Van Rensburg *et al.* 1992).

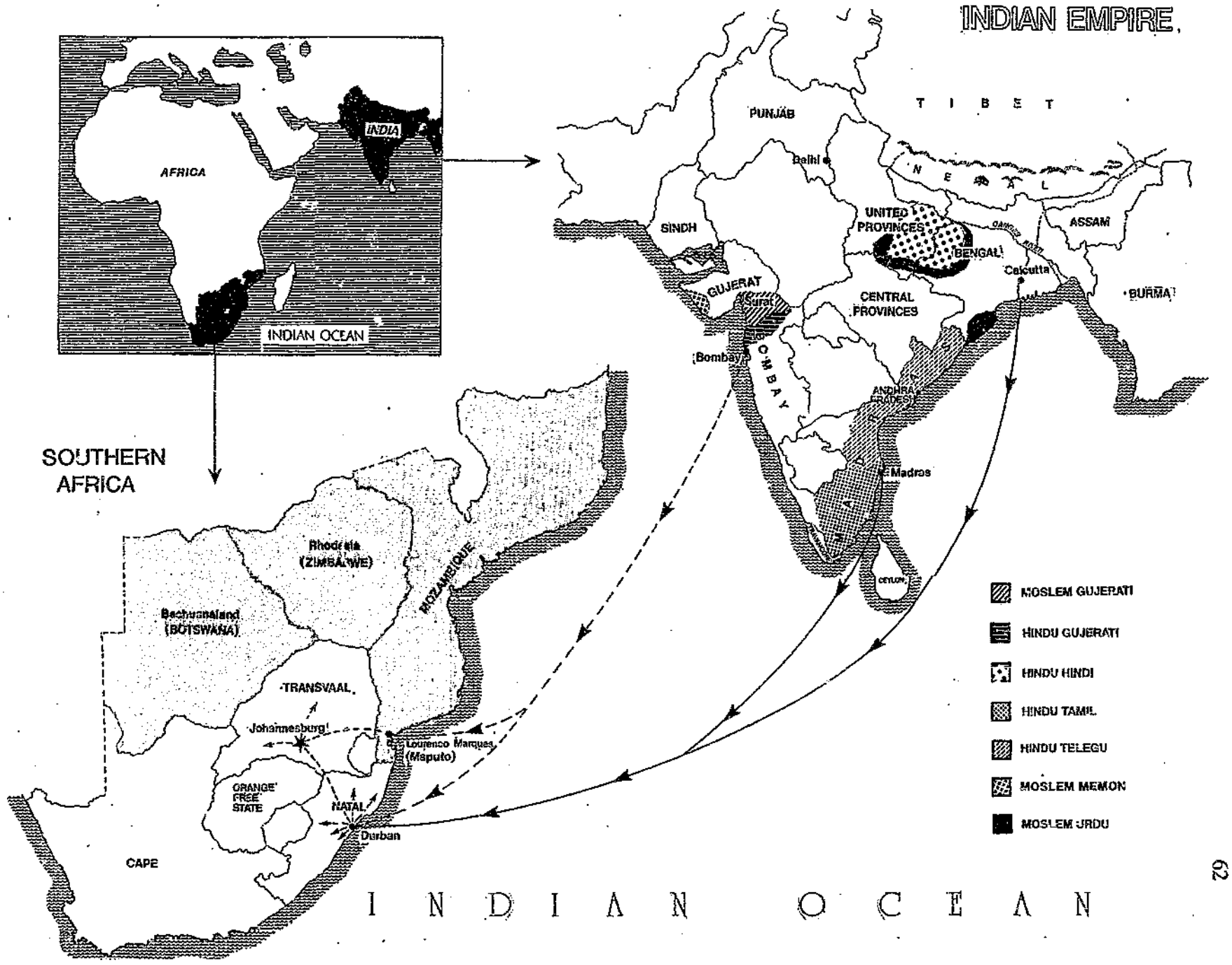
Two main groups of Indians arrived in South Africa in the late nineteenth and early twentieth century, the indentured immigrants and the passenger Indians. Detailed studies of their origins have been undertaken by Bhana and Brain (1990) and Bhana (1991) and some of their findings are detailed below. Figure 1.4a shows the geographical origins of the different religious and language subgroups of the South African Indians, together with their routes of migration to South Africa. Figure 1.4b shows the Indian subcontinent enlarged so that the areas of origin can be seen more clearly.

1.10.2.1 The indentured immigrants

The 152 184 indentured immigrants made up the largest group who arrived in Natal between November 1860 and July 1911. They were imported to man new and labour-intensive plantation crops, mainly sugar, along the coastal belt, although very few of the employers of the first indentured labourers who arrived between 1860-66 were actually engaged in sugar production. Indentured labourers were recruited in towns and in hundreds of small villages in India by professional recruiters in the employ of the emigration agents situated in Madras and Calcutta.

FIGURE 1.4a The geographical origins of the different religious and language subgroups of the South African Indians and their routes of migration to South Africa.

The indentured Indians came along the routes marked with a solid line, while the passenger Indians came along the routes marked with a dashed line.


INDIAN EMPIRE.
INDIA
AFRICA
INDIAN OCEAN
SOUTHERN AFRICA
PUNJAB
TIBET
Delhi
NEPAL
SINDH
UNITED PROVINCES
ASSAM
BENGAL
Calcutta
GUJERAT
Surat
CENTRAL PROVINCES
BURMA
Bombay
BOMBAY
ANDHRA PRADESH
MADRAS
Madras
Rhodesia (ZIMBABWE)
MOZAMBIQUE
Bechuanaland (BOTSWANA)
TRANSVAAL
Johannesburg
Lourenco Marques (Maputo)
ORANGE FREE STATE
NATAL
Durban
CAPE
CEYLON
MOSLEM GUJERATI
HINDU GUJERATI
HINDU HINDI
HINDU TAMIL
HINDU TELEGU
MOSLEM MEMON
MOSLEM URDU
INDIAN OCEAN

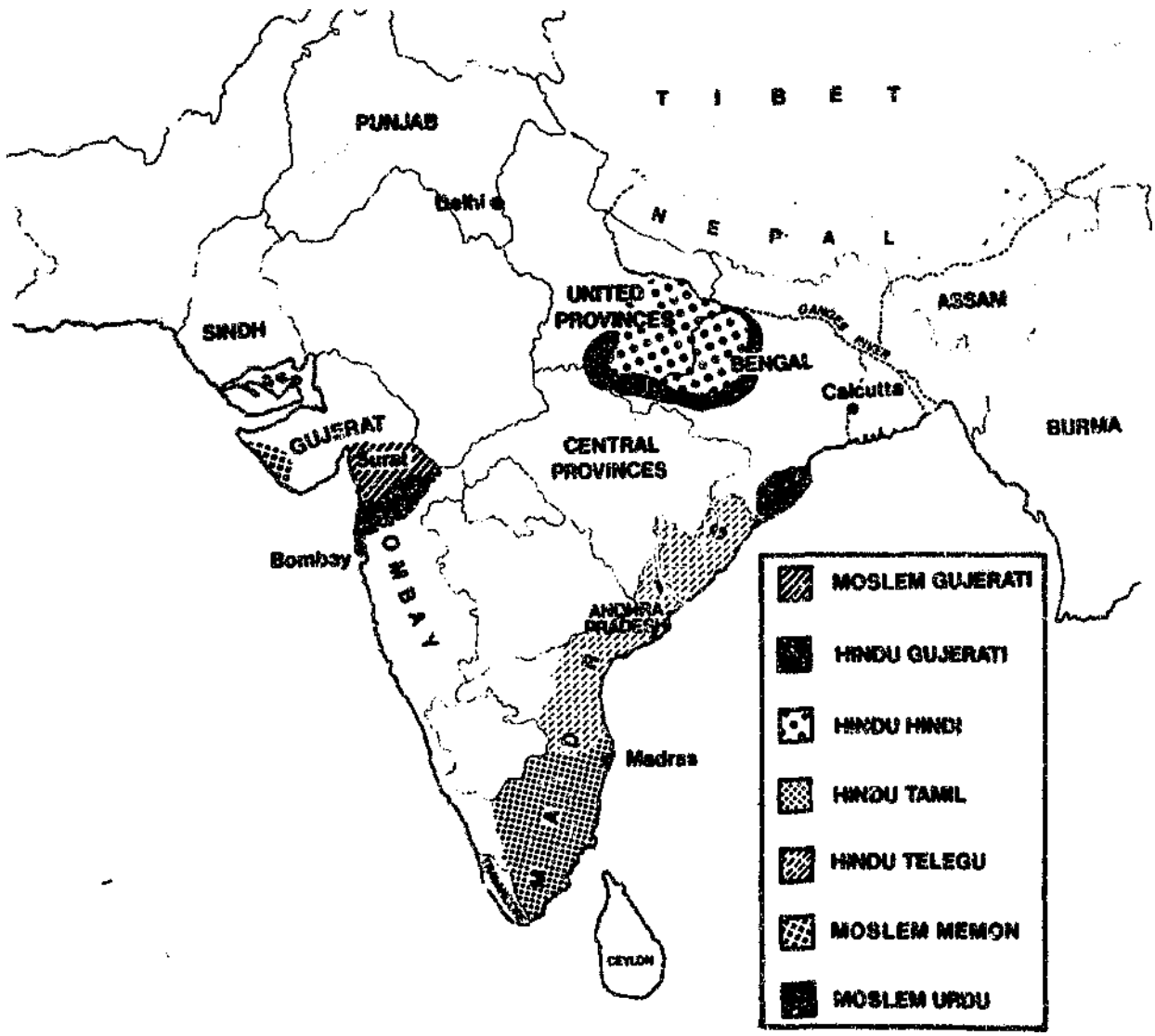


FIGURE 1.4b Map of the Indian subcontinent showing the areas of origin of the different religious and language subgroups of the South African Indians.

Nineteen ships arrived in South Africa between November 1860 and July 1866, mainly from Madras, with indentured immigrants predominantly from the Madras Presidency and the Ganges Valley, though individual immigrants came from further afield. Although immigrants spoke a variety of languages as well as being of many different castes, they were predominantly Hindu, speaking Tamil or Telegu. The period 1866-1870 was one of economic depression and many labourers were indentured to insolvent estates. Many were shipped back to India and lodged complaints of their poor treatment in Natal. While this was investigated no further immigration was permitted.

A genuine labour shortage in Natal led to the re-establishment of the indenture system with a protector of Indian immigrants appointed and new regulations, after which importation of Indian labourers started again in 1874. In the 1870's the majority of indentured labourers were recruited in the north-eastern states of India and sailed from Calcutta. These immigrants were allocated to employers along the coast, corporations of Durban and Pietermaritzburg and the Natal Government railways. A large proportion were Hindi-speaking Hindus.

The indentured system continued until 1911 with ships coming from both Madras and Calcutta. While for the first decade most indentured labourers were employed in the coastal belt, they gradually spread inland with the demand for labour. When the indentured system came to an end indentured Indians were distributed over most parts of Natal. Upon termination of their contracts, the indentured Indians sometimes returned to India at government expense and at other times elected to stay, swelling the numbers of the so-called 'free Indians'. They were free to settle in Natal or to leave Natal and seek their fortunes elsewhere in the country. Many of the labourers settled where they had been stationed, bringing their families and seeking employment. They became market gardeners, business men and hawkers of fresh fruit, vegetables and fish. Increasing numbers were employed inland as waiters, domestic servants, stockmen and labourers. Others became wage earners in factories, municipal employees or shop assistants. Some took to independent occupations including jewellery-making, tinsmithing, blacksmithing, basket-making or carpentry. The largest numbers became artisans and labourers. Some of the Indians who terminated their indentures went into farming, where they played an

important part in the agricultural sector in the 1890's. They produced most of the fruit and vegetables for the urban markets. They were mainly Hindi- and Tamil- speakers and were largely confined to Natal.

There were also some small groups on the immigrant ships of 'special servants', skilled men brought in under contract and paid far more than indentured labourers to occupy positions as waiters, carriage drivers, dhobies (laundrymen) and grooms in hotels and clubs.

1.10.2.2 The passenger Indians

The second major group consisted of the passenger Indians, individuals who arrived at their own expense as British subjects and under the ordinary laws of the colony. They came predominantly from western India, though some came via Mauritius. Most came between 1396 and 1900. They included teachers and interpreters, though the majority were traders and hawkers. The presence of the indentured immigrants offered them opportunities for trade in goods that could be supplied from India. Though they came initially to serve the needs of the indentured and free Indians, they quickly saw opportunities for trading with Blacks engaged in subsistence farming or wage labour, and later also sold to white clients. They first settled in Durban and Natal in the 1870's, but later moved into the interior. Gold and diamond mining offered further opportunities for trade and the Indians were soon found in the Transvaal and Cape. They were prohibited from mining but capitalised on opportunities in petty trade and services created by the mining activity.

After 1897, when Natal passed a law requiring a literacy test for immigrants, as the Europeans feared the control exercised by the Indian traders, many Indians chose to disembark at the Portuguese East African (Mozambique) port of Delagoa Bay or at ports in the Cape. By 1911 passenger Indians were to be found in nearly all the main towns and cities in South Africa, with a fairly even distribution of settlement between the Transvaal and Natal, although less in the Cape. A 1911 census showed 149 791 Indians in South Africa of which about 30 000 were passenger Indians.

About half the passenger Indians came from the districts of Surat and Valsad, the major towns being Surat, Rander, Kholvad, Kathor, Bardoli, Baroda and Navsari. Kholvad and Kathor supplied mainly Moslem immigrants while Navsari and Bardoli sent Hindus. Kathiawar was the third major source, from towns and villages between Jamnagar near the Gulf of Kutch and Porbandar on the coast of the Arabian sea, including Rajkot, Kalavad, Lalpur, Gondal, Bhanvad and Ranavav. The majority of passenger Indians were Gujerati-speaking Moslems and Hindus. There were a few Tamil-speakers and among the Moslems were Memons and Vohras. The Moslems, particularly, engaged in large business ventures and expanded Indian trade considerably, making extensive use of family connections. The Hindus engaged in small trade or service, becoming shop assistants, accountants, bookkeepers and teachers, though some became prominent businessmen. Kinship networks were important in areas of settlement, especially in outlying areas. For example, many of the immigrants from Bhavnad and Ranavav in Kathiawar, who were mainly Memons, settled in Potchefstroom and Pietersburg, in the western and northern Transvaal respectively, where they engaged in trade with the Blacks and the Boers.

The offspring of both the indentured and passenger Indians became a new class as they were free to go where their work took them. Those with indentured parents sought work in the civil service or entered professions; those of passengers often went into their fathers' businesses.

1.10.3 The present South African Indian population

The present South African Asian Indian population can be considered heterogeneous, but can be divided along religious and linguistic lines into smaller subgroups, corresponding to their geographical origins in India, as shown in Figure 1.4b (Grierson 1927). About 70% of the population are Hindu, speaking either Gujerati, Hindi, Tamil or Telegu, while 20% are Moslem, speaking mainly Gujerati, but also Urdu or Memon, a dialect of Gujerati. The remainder of the population are Christians, Buddhists and Jains. The Tamil and Telegu are people of Dravidian culture, whereas the other groups are of Indo-Aryan culture (Mistry 1965).

The groups differ from one another in dietary habits, marriage and social customs and

cultural activities. Religion, home language, caste or community continue to play a significant role in the life of Indians who retain their strong traditions despite living amongst other ethnic groups. Marriages still tend to take place among members of the same religion, language and caste, so the subgroups are maintained and can be used as a guide to the geographical origins of individuals (Mistry 1965).

Today at least 80% of South African Asian Indians still live in Natal, approximately 15% live in the Transvaal, and 4% in the Cape (Department of Central Statistics 1990). They are concentrated around the large cities. While the majority of Natal Indians are Hindu, Moslems form over 50% of the Transvaal Indians, reflecting their historical origins.

1.11 Aims of the study

In South Africa the majority of cases of ß thalassaemia, sickle cell anaemia and other haemoglobinopathies are known to occur in Indians, but no population surveys have been carried out to determine the exact frequency of these disorders. In addition, the molecular basis of these disorders in this population had not been characterised. Thus, the aims of this study were as follows:

1. To carry out a population-based survey to determine the allele frequencies of ß thalassaemia, $ß^s$, α thalassaemia and other haemoglobinopathies in the major subgroups of the South African Asian Indians.

2. To determine the frequencies of non-pathological variants, including numerical variation in gene number, RFLPs and VNTRs, in both the α and ß globin clusters.

3. To establish the α and ß globin haplotypes associated with normal α and ß globin genes, with ß thalassaemia, $ß^s$, and other abnormal ß globin alleles, as well as with α thalassaemia alleles.

4. To delineate the different types of deletions or mutations which cause α thalassaemia.

5. To characterise the ß thalassaemic mutations in the South African Indian population.

6. To use the mutation/haplotype associations of both the abnormal α and ß globin alleles to indicate the possible geographical origins of the mutations and the number of times they have arisen.

7. To use the population frequency data to assess the need for establishing a screening programme for the major haemoglobinopathies.

8. To evaluate the local prenatal diagnostic service for the major haemoglobinopathies by determining the most useful linked markers and the most common mutations in the families who have used the service, and by assessing the service itself.

9. To assess the differences between the major religious and linguistic subgroups and to compare them with other Asian Indians, using the normal and pathological variation in both globin clusters.

CHAPTER 2 - SUBJECTS AND METHODS

2.1 Subjects

The individuals studied were all South African Asian Indians, who were classified into a religious and linguistic subgroup after an interview. Seven religious and linguistic subgroups are included in the study, namely: Moslem Gujerati, Hindu Gujerati, Hindu Hindi, Hindu Tamil, Moslem Memon, Moslem Urdu and Hindu Telegu. As the first four constitute the majority of the South African Indian population, these were classified as major subgroups for the purposes of this study, while the latter three were classified as minor subgroups.

An attempt was made to determine the town or village of origin of the family in India, though this was often difficult as most of the individuals interviewed were third or fourth generation South Africans and were not aware of their origins. This information has not been retained through the generations, whereas religion, language and caste, important for marriage, have been passed down.

The purposes of the study were explained to all individuals and informed consent, either verbal or written was obtained prior to phlebotomy. In the case of minors, parental consent was obtained.

The study was carried out in a number of stages, with samples from different individuals used in separate parts of the study. This was necessary to ensure random selection, where required, but also sufficient numbers of individuals with rare genotypes.

2.1.1 Lenasia schools survey

Blood samples were collected from 638 unrelated teachers and senior pupils (Standard eight upwards) at five Lenasia high schools. Lenasia is a municipality south of Johannesburg where the majority of Johannesburg Indians reside. The average age of the individuals was 22.2 $\pm$ 10.3 years, with a range from 15-60. Written parental consent

was obtained in the case of pupils who were minors. The consent form used can be found in Appendix I.

The group consisted of:

305 Moslem Gujerati
175 Hindu Gujerati
115 Hindu Tamil
21 Hindu Hindi
13 Moslem Urdu
2 Hindu Telegu
1 Moslem Memon
2 Other
4 Unknown

As the sampling was done as part of a collaborative study, with blood being collected for HLA and other genetic marker studies simultaneously, only a 5mℓ EDTA specimen was obtained from each individual and a small aliquot of serum, sufficient for iron studies, was obtained in most cases. The aim of this part of the survey was to determine the frequencies of ß thalassaemia and other haemoglobin variants. In addition, the causes of microcytic hypochromic anaemia could be assessed. Thus, on each sample a full blood count, qualitative haemoglobin electrophoresis and HbF determination were done, as well as a serum iron and transferrin determination, where possible. If the MCV was $<80fℓ$, the MCH $<27pg$ or the Hb less than $13g/dℓ$ in males or $12g/dℓ$ in females, HbA_2 was quantitated. It is estimated that in excess of 95% of ß thalassaemia heterozygotes have microcytic and/or hypochromic indices (Weatherall and Clegg 1981). In addition, if an abnormal haemoglobin was detected on qualitative electrophoresis, this was also quantitated.

As a service to the individuals studied and to confirm our results, we attempted to get repeat samples from the 177 individuals with microcytosis, hypochromia and/or anaemia. Blood was collected from 137 of these individuals and DNA studies to identify the presence of α thalassaemia, iron studies to exclude iron deficiency anaemia, and further haematological investigations were carried out to confirm the presence of ß thalassaemia

or other haemoglobin variants.

Individuals who were iron deficient were sent a letter advising them of the results and urging them to contact their doctor. Where ß thalassaemia or $ß^S$ was identified in an individual, the family was offered genetic counselling and testing of 'at-risk' individuals and their spouses. Where α thalassaemia carriers were identified, families were offered counselling to explain the benign cause of their microcytosis or hypochromia. All families with a haemoglobinopathy were sent a written report detailing the diagnosis and implications.

When a family member with a haemoglobinopathy was identified, blood was collected from the nuclear family, where possible. Eleven such families were obtained. These families were used to identify ß thalassaemia mutations and their associated haplotypes, as well as the haplotypes associated with the $ß^S$ mutation and normal ß globin haplotypes. α haplotypes associated with $\alpha\alpha$ and $-\alpha$ chromosomes were also studied as well as other non-pathological variation in both the α and ß globin clusters.

2.1.2 Random families

Although the Lenasia school survey provided estimates of the frequency of ß thalassaemia and other ß globin variants in the Indian population, it only provided a minimum estimate of the α thalassaemia frequency. Firstly, a significant percentage of individuals who have the $-\alpha/\alpha\alpha$ genotype may have normal haematology (Bowden *et al.* 1985) and would thus not have been detected in the initial survey. Secondly, although individuals with the $-\alpha/-\alpha$ or $--/\alpha\alpha$ genotype would have microcytosis and hypochromia and would thus have been detected haematologically, it was only possible to obtain DNA samples from about 80% of these individuals to determine their α thalassaemia genotypes. Determination of the α thalassaemia frequency thus requires DNA studies on all individuals.

Blood was collected from a second subject group of 60 nuclear families, with at least two children each, 15 from each of the major religious and linguistic groups, namely Moslem Gujerati-speaking, Hindu Gujerati-speaking, Hindu Tamil-speaking and Hindu Hindi-speaking. These families were collected randomly mainly in Lenasia, though about one-

fifth were collected in Durban, Natal. Individuals in the community were used as contacts to ascertain families. As far as possible, families in which there were consanguineous marriages were excluded. No families had any history of thalassaemia or any other genetic condition. In addition to a verbal explanation, each family was provided with an information sheet explaining the study (see Appendix II) and verbal informed consent was obtained from all individuals, with parental consent in the case of minors.

The 120 parents were considered random individuals and were used to estimate the frequencies of α thalassaemia, as well as of other α and ß globin cluster variants, in each of the major religious and language groups. In addition, the families were used to study normal α and ß globin haplotypes and those associated with α thalassaemia.

2.1.3 Families with major haemoglobinopathies

Blood was collected from nuclear families in which at least one child had a major haemoglobinopathy. This included 17 families where the affected child(ren) was(were) being treated at Coronation Hospital in Johannesburg, one family where the affected children were being treated at Johannesburg General Hospital, and 14 families under the care of the Natal Genetics Services' Sisters. All these families had requested a 'workup' in preparation for prenatal diagnosis. They included 26 thalassaemia major families, three $ß^S/ß^T$ families, two $ß^E/ß^T$ families and one family with HbH disease. These families were used to identify ß thalassaemia mutations and the haplotypes associated with the different abnormal ß globin genes. In addition, normal ß globin haplotypes, α globin variation and α globin haplotypes were also studied. In four consanguineous families, the chromosomes of one parent were excluded from the analysis, so that the same chromosome was not counted twice.

In four families, the parents were identified as ß thalassaemia carriers, prior to the birth of an affected child. The parents were used to identify ß thalassaemia mutations, but no haplotype studies were possible.

During the course of the study, from 1984-1993, 22 prenatal and two preclinical (at birth) diagnoses were carried out for these families.

2.1.4 Additional families

Through the Genetic Counselling Clinic, an additional seven families were ascertained, in which an abnormal haemoglobin had been identified. These families were used to characterise both normal and abnormal α and ß variation and haplotypes. Consanguinity was again taken into consideration for the analysis, as detailed above.

2.1.5 Venous blood sampling and processing

As described in Section 2.1.1, the Lenasia school survey was a larger study and only 5mℓ of the blood collected from all subjects by venepuncture was used in this part of the study. For the remaining subjects in this study, approximately 30-50mℓ of blood was collected. From each individual one 5mℓ tube, with EDTA as anticoagulant, was used for haematological studies. A 5mℓ sample was collected into a plain glass tube, allowed to clot and an aliquot of the serum used for iron studies. The remainder of the serum was frozen for possible future use in genetic marker studies. The rest of the blood was collected into either ACD or EDTA tubes and was used for DNA studies.

The blood intended for DNA studies was centrifuged at 3 000rpm in a bench-top centrifuge for 15 minutes to separate it into its major components. The plasma was removed and stored for possible later use. The leucocytes, or buffy coat, were removed with a Pasteur pipette and frozen at -20°C until required for DNA extraction. The remaining red cells were mixed in a 1:1 ratio with preserving fluid (see Appendix III) and stored at -20°C for future use in genetic marker studies.

2.2 Haematological methods

2.2.1 Full blood counts (FBCs)

FBCs were performed on a Sysmex K1000 or equivalent machine, which automatically determined the red blood cell count (RCC), haemoglobin (Hb), haematocrit (Hct), mean cell volume (MCV) and mean cell haemoglobin (MCH). The machines were located in

2.1.4 Additional families

Through the Genetic Counselling Clinic, an additional seven families were ascertained, in which an abnormal haemoglobin had been identified. These families were used to characterise both normal and abnormal α and ß variation and haplotypes. Consanguinity was again taken into consideration for the analysis, as detailed above.

2.1.5 Venous blood sampling and processing

As described in Section 2.1.1, the Lenasia school survey was a larger study and only 5mℓ of the blood collected from all subjects by venepuncture was used in this part of the study. For the remaining subjects in this study, approximately 30-50mℓ of blood was collected. From each individual one 5mℓ tube, with EDTA as anticoagulant, was used for haematological studies. A 5mℓ sample was collected into a plain glass tube, allowed to clot and an aliquot of the serum used for iron studies. The remainder of the serum was frozen for possible future use in genetic marker studies. The rest of the blood was collected into either ACD or EDTA tubes and was used for DNA studies.

The blood intended for DNA studies was centrifuged at 3 000rpm in a bench-top centrifuge for 15 minutes to separate it into its major components. The plasma was removed and stored for possible later use. The leucocytes, or buffy coat, were removed with a Pasteur pipette and frozen at -20°C until required for DNA extraction. The remaining red cells were mixed in a 1:1 ratio with preserving fluid (see Appendix III) and stored at -20°C for future use in genetic marker studies.

2.2 Haematological methods

2.2.1 Full blood counts (FBCs)

FBCs were performed on a Sysmex K1000 or equivalent machine, which automatically determined the red blood cell count (RCC), haemoglobin (Hb), haematocrit (Hct), mean cell volume (MCV) and mean cell haemoglobin (MCH). The machines were located in

the routine haematology laboratory and standards and controls were run daily, to ensure that they were accurately calibrated.

2.2.2 Preparation of haemolysates

Cells from 2-3mℓ of whole blood from an EDTA tube were washed three times with normal (0.9%) saline and then once with 1.2% saline. Drabkin's solution (see Appendix III) was added, about half as much as the total red cell volume, and the mixture was left to stand for about 10 minutes. This procedure converted the haemoglobin to cyanmethaemoglobin. An equal volume of carbon tetrachloride was added, the solution vigorously mixed to lyse the cells and then spun for 20 minutes at 3 000rpm in a bench top centrifuge. The haemolysate was removed, the haemoglobin concentration read on the Sysmex K1000, and adjusted to a final concentration of 10g/dℓ by adding additional Drabkin's solution.

2.2.3 Haemoglobin electrophoresis

2.2.3.1 Cellulose acetate electrophoresis

Both qualitative and quantitative electrophoresis were carried out using the Gelman Hemoglobin Electrophoresis System (Gelman Sciences Incorporated), designed to separate by electrophoresis, detect and quantitate haemoglobins on cellulose acetate membranes in an alkaline medium.

Electrophoresis was carried out in a Gelman semi-micro electrophoretic chamber, using Gelman Hemoglobin Buffer (No. 51126), reconstituted according to the manufacturer's specifications, and Gelman Sepraphore III cellulose acetate membrane (No. 51010). Haemolysate at a concentration of 3g/dℓ was loaded onto the strips and allowed to run at 400V for 45 minutes.

For qualitative electrophoresis, eight samples were loaded on a cellulose acetate strip. After electrophoresis, the strips were stained in Gelman Ponceau S solution (No. 51284) for one minute, destained in acetic acid for 30 minutes and analysed. With this technique,

HbA_2, HbC and HbE run together, as do HbS and HbD, while HbF and HbA are clearly separated. Thus, this technique was used as a screen to detect qualitative electrophoretic variation in haemoglobin, though the exact identity of variants had still to be determined (see Section 2.2.3.2).

For quantitative electrophoresis, a single sample was loaded on a cellulose acetate strip. Two strips were run for each sample. After electrophoresis, the bands were cut out of the strips, eluted in distilled water for three hours and the optical densities measured at λ=413nm. The contribution of each band to the total haemoglobin could thus be calculated.

2.2.3.2 Citrate agar electrophoresis

Citrate agar electrophoresis, which separates haemoglobins based on their electric charge and adsorption in an agar medium, was use to confirm the identity of haemoglobin variants detected on the Gelman system, particularly to distinguish HbS from HbD and HbC from HbE.

The Helena Laboratories System was used, with electrophoresis being carried out in a Helena electrophoretic chamber, using Helena citrate buffer (No. 5121), reconstituted according to the manufacturer's specifications, and a Helena Titan IV citrate agar plate (No. 2400). Haemolysate at a concentration of 3g/dℓ was loaded onto the agar plate and run at 40mA for 45 minutes. The plate was stained with an O-toluidine stain (see Appendix III), then rinsed with 5% acetic acid, after which the bands could be clearly visualised.

2.2.4 HbF quantitation - alkali denaturation

This technique is based on the fact that HbF is more resistant to alkali denaturation than the other normal haemoglobins, HbA and HbA_2 (Betke *et al.* 1959). 3.2mℓ of 0.08N NaOH was added to 0.2mℓ of haemolysate, mixed and allowed to stand for two minutes, after which 3.2mℓ of 0.08N HCl, 3.2M ammonium sulphate was added. The mixture was left to stand for one minute and then filtered through Whatman 1MM paper. The OD of

the filtrate at λ=540nm was measured and compared to that of a non-denatured haemolysate, thus enabling the percentage of HbF to be calculated.

2.2.5 Iron studies

Serum iron and transferrin levels were determined in the routine chemical pathology laboratory of the South African Institute for Medical Research (SAIMR). The transferrin saturation was calculated from these two parameters.

2.3 DNA methods

In these studies human genomic DNA, isolated from white blood cells, fetal fibroblasts or trophoblastic tissue, was analysed by Southern blot transfer techniques or the polymerase chain reaction (PCR).

During the course of the study, newer, generally more efficient techniques, gradually replaced the older ones. Only the most recent methods will be included in this chapter. Details of media and solutions used are listed in Appendix III. All centrifugations were done in a Beckman benchtop centrifuge model TJ-6, unless otherwise stated.

2.3.1 Preparation of DNA probes

In order to study the α and ß globin gene clusters a number of probes were used which detect variation in these regions. The probes used were all contained in plasmid vectors and the details of the probes, their vectors, insert sizes and characteristics are shown in Tables 2.1 and 2.2 for the α and ß clusters respectively. The α globin gene cluster probes were all supplied by Dr DR Higgs, John Radcliffe Institute, Oxford and the ß globin gene cluster probes by Dr JS Wainscoat and Dr J Old, John Radcliffe Institute, Oxford.

2.3.1.1 Transformation of plasmid DNA

Probes were generally received as DNA preparations and it was thus necessary to transform the recombinant plasmid DNA into a bacterial host.

TABLE 2.1 DNA probes used for analysis of the α globin gene cluster

PROBE	VECTOR	ANTIBIOTIC RESISTANCE	INSERT	DESCRIPTION OF PROBE	REFERENCE
p5'HVR.14	pSP64	Ampicillin	770bp *Eco*RI/*Hin*dIII	Hypervariable region 100kb 5' to the α cluster, consisting of a 57bp tandem repeat	Jarman and Higgs 1988
pBRζ1	pBR322	Ampicillin	4.5kb *Bam*HI/*Eco*RI	Human $\psi\zeta$1 gene	Lauer *et al.* 1980
pSG21	pBR322	Tetracycline	1.1kb *Alu*I	Interzeta hypervariable region, consisting of a 36bp tandem repeat	Goodbourn *et al.* 1983, 1984
pDHα	pBR322	Tetracycline	3.7kb *Bgl*II/*Eco*RI	Human α1 genomic DNA	Lauer *et al.* 1980
pDH12	pSP64	Ampicillin	1.9kb *Hin*dIII/*Sac*I	Genomic single-copy sequence 3kb 3' to α1	Higgs *et al.* 1986
pα3'HVR.64	pSP64	Ampicillin	4kb *Hin*fI	Hypervariable region 8kb 3' to α1, consisting of 17bp tandem repeat	Higgs *et al.* 1981 Jarman *et al.* 1986

TABLE 2.2 DNA probes used for analysis of the β globin gene cluster

PROBE	VECTOR	ANTIBIOTIC RESISTANCE	INSERT	DESCRIPTION OF PROBE	REFERENCE
pε1.3	pBR322	Ampicillin	1.3kb *Bam*HI/*Eco*RI	Human ε globin genomic DNA (from 3′ end of exon 2)	Baralle *et al.* 1980
pHd3.2 ($p^{A}\gamma$)	pBR322	Ampicillin	3.2kb *Hin*dIII	Human $^{A}\gamma$ globin gene and flanking DNA	Dr J Wainscoat - personal communication to Dr M Ramsay (1985)
pP3.9	pBR322	Tetracycline	3.9kb *Pst*I	Human $\psi\beta$1 genomic DNA	Fritsch *et al.* 1980
βIVS2	pBR322	Ampicillin	920bp *Bam*HI/*Eco*RI	Human β globin IVS2	Burns *et al.* 1981
pRK29	pBR322	Ampicillin	1.2kb *Eco*RI	Genomic DNA situated 18kb 3′ of β	Tuan *et al.* 1983

Competent *E. coli* cells (strain HB101), were prepared using a $CaCl_2$ technique. *E. coli* was grown in Luria broth (LB) at 37°C until the culture had an OD_{550} of 0.5. The culture was chilled on ice for 10 minutes and centrifuged at 3 000rpm for 15 minutes to harvest the cells. The pellet was resuspended in 10mℓ 50mM $CaCl_2$/10mM Tris HCl (pH=8.0) and centrifuged once again. The pellet was then resuspended in 1mℓ of the $CaCl_2$/Tris HCl solution, aliquotted into 0.1mℓ amounts and kept on ice until required.

One µl of plasmid DNA was added to the competent cells and placed on ice for 30 minutes. The solution was heat-shocked in a 45°C waterbath for two minutes and returned to ice for 1 minutes. The transformation mix was added to 3mℓ LB and incubated at 37°C for 1-3 hours prior to being plated out and grown overnight on LA plates, with the appropriate antibiotic (to which the plasmid was resistant), after which single transformed colonies were identified.

2.3.1.2 Long term storage of bacterial strains

Once a plasmid had been transformed, a single colony was grown overnight in 10mℓ of LB, with the appropriate antibiotic. 0.85mℓ amounts were added to 0.15mℓ of sterile glycerol, mixed, frozen rapidly in an ethanol/NaCl bath and stored at -70°C. These transformed strains could be used at a later date for further culture.

2.3.1.3 Large scale preparation of plasmid DNA

The quick plasmid preparation from Promega (Promega Protocols and Applications Guide 1989) was used. After overnight growth of a transformed bacterial colony in 250mℓ LB, with the appropriate antibiotic, bacterial cells were pelleted and lysed using Lysozyme (2mg/mℓ) in 6mℓ ice-cold lysis buffer. After 20 minutes on ice, 12mℓ 0.2M NAOH:1% SDS was added and lysis allowed to proceed for another 10 minutes before 7.5mℓ 3M Na Acetate (pH=4.6) was added. The bacterial debris was pelleted by spinning the solution at 15 000rpm for 20 minutes in a Sorvall RC5C centrifuge. The supernatant was placed in a clean tube and the RNA removed by adding 5µl RNase A (10mg/mℓ) and incubating the solution at 37°C for 20 minutes. Contaminating proteins were then removed by two phenol:chloroform (1:1) and one chloroform extraction (all chloroform extractions in

these techniques were done with a 24:1 chloroform:isoamyl alcohol solution). The DNA was precipitated by the addition of 2.5 volumes of absolute ethanol and storage at -70°C for 20 minutes. The precipitated DNA was pelleted by centrifugation at 3 000rpm for 30 minutes, washed with 70% ethanol to remove excess salt and resuspended in 1mℓ TE (pH=8.0).

Though not specifically recommended by the protocol, the plasmid DNA was purified on a caesium chloride ethidium bromide density equilibrium gradient to remove any bacterial or RNA contaminants. The dissolved plasmid DNA pellet was added to 4g CsCl in 3mℓ TE and 320μl of ethidium bromide (10mg/mℓ). The solution was transferred into a Beckman Quickseal polyallomer tube (13x51mm) and spun in a vertical rotor (Beckman Vti65.2) for 16 hours at 45 000rpm in a Beckman L8-55 Ultracentrifuge. Using this method the bands were visualised under UV light, a thin upper band of bacterial chromosomal DNA and nicked linearised plasmid DNA (not always visible), a large central band of closed circular plasmid DNA, which was removed by inserting a wide bore needle attached to a syringe into the side of the centrifuge tube and aspirating the band, and a lower band of RNA (not always visible).

The ethidium bromide was removed from the plasmid sample by repeated isoamyl alcohol extraction, until the pink colour in the lower phase had disappeared. The sample was then dialysed against TE on a Millipore VS filter (pore size 0.025μm) to remove the caesium chloride.

2.3.1.4 Verification of the plasmid

A small sample of the purified plasmid DNA was run on an agarose gel alongside a marker of known size, usually the BRL 1kb ladder, to check that the plasmid was of the correct size. In addition, the size of the insert was verified by cutting it out of the plasmid with the appropriate restriction enzyme(s) and running the digested sample on an agarose gel with a sized marker.

2.3.1.5 Isolation and purification of DNA fragment for radiolabelling

Although in some cases it was possible to radioactively label the whole plasmid, certain of the plasmids gave poor hybridisation signals and some of the inserts contained repetitive DNA. In these cases it was necessary to isolate and purify a DNA fragment from the plasmid for oligolabelling.

A digestion was set up with about 50-100μg of plasmid and the appropriate restriction enzyme(s) and buffer(s). In the case of double digests, some could be done by adding the two enzymes sequentially in the same buffer. In other cases it was necessary to adjust the salt concentrations for the two enzymes, starting with the enzyme requiring a lower salt buffer, and then adding salt prior to adding the second enzyme. Digestions were checked at each step, by agarose gel electrophoresis of an aliquot, to ensure the reaction had gone to completion.

Once the digestion was complete, the sample was loaded onto a 1% low melting point Sea Plaque agarose gel and run in 1xTBE buffer at 4°C, until the fragment of interest and the rest of the DNA were clearly separated. The fragment was then cut out of the gel with a blade and the fragment-containing agarose finely chopped up. Aliquots of 0.3mℓ were placed in an Eppendorf tube with an equal volume of phenol, vortexed vigorously, then frozen rapidly in an ethanol/NaCl bath for five minutes and then spun in a microcentrifuge for five minutes. The vortexing, freezing and centrifuging cycles, which released the DNA from the agarose, were repeated twice more. The aqueous phase was then removed and extracted with chloroform to remove any remaining phenol, and the DNA was precipitated using 1/10 volume 3M NaAcetate and 2.5 volumes of absolute ethanol. It was placed at -70°C for 20 minutes and then spun in a microcentrifuge for at least 30 min to pellet the DNA. The pellet was rinsed in 70% ethanol and resuspended in 50-250μl of TE. This technique provided extremely pure DNA which could be oligolabelled to high specific activity and gave strong hybridisation signals with minimal background.

2.3.2 Genomic DNA extraction

2.3.2.1 DNA extraction from peripheral blood

DNA was extracted from the buffy coat using the method of Sykes (1983), in which cells were washed with a non-ionic detergent solution, 0.2% Triton-X/0.9% NaCl and then lysed with a urea-containing buffer and SDS. Protein was removed using phenol:chloroform and chloroform extractions and the DNA precipitated by adding absolute ethanol to the solution. The DNA was collected by spooling, then dissolved and stored in TE, at 4°C for short term storage or -70°C for long term storage.

2.3.2.2 DNA extraction from fetal fibroblasts

After amniocentesis, fetal fibroblasts in amniotic fluid were cultured by the Cytogenetics Unit of the Department of Human Genetics, SAIMR, until a confluent layer of adherent cells was obtained. One to three flasks of cells were generally obtained.

The medium was poured off and the cells in each flask were rinsed with phosphate-buffered saline (PBS). The cells in the first flasks were lysed by incubating them in 5mℓ Tris/EDTA lysing buffer with 25μl Proteinase K (10mg/mℓ) at 45°C for 10 minutes. The lysate was transferred to the second flask, incubated again, transferred to the third flask and reincubated. The lysate was placed at 68°C for 30 minutes and then at 45°C for 90 minutes. Protein was removed by one phenol:chloroform and two chloroform extractions and the DNA was precipitated by adding 5M NaCl (to a final concentration of 0.3M) and 2.5 volumes of absolute ethanol, followed by storage at -70°C for 20 minutes. The solution was centrifuged to pellet the DNA, the pellet was rinsed with 70% ethanol, and dissolved in TE.

2.3.2.3 DNA extraction from trophoblastic tissue

After chorionic villus sampling (CVS), sufficient material was generally obtained to extract DNA directly without culture. If the cells were cultured, the same method as that used for fetal fibroblasts was used to extract DNA (Section 2.3.2.2).

After the Cytogenetics Unit staff had dissected the villi from the maternal tissue, to avoid maternal contamination, the villi were spun down at 3 000rpm for 10 minutes. They were resuspended in 500μl of amniocyte buffer, with 10μl 10% SDS and 10μl Proteinase K (10mg/mℓ) and incubated at 55°C for 2-3 hours, until the villi were completely lysed. Protein was removed using one phenol:chloroform and two chloroform extractions. The DNA was precipitated using 1/10 volume 3M NaAcetate and 2.5 volumes absolute ethanol and, after spooling or centrifugation, was resuspended in TE.

2.3.2.4 Quantitation of DNA

A small volume of DNA was run on a 0.8% ethidium bromide stained agarose gel, with DNA of known concentration alongside. This crude quantification provided sufficient information for calculating the amount of DNA required for restriction endonuclease digestion. In addition, visualisation of the DNA as a single, slow-travelling, high molecular weight band indicated that the DNA was of high quality, non-sheared and non-degraded.

2.3.2.5 Dialysis of DNA

As the DNA from fibroblasts and trophoblastic tissue was precipitated with high salt solutions and dissolved in relatively small volumes, it was necessary to dialyse the samples prior to either restriction enzyme digestion or PCR amplification to remove the salt. Samples were dialysed against TE using Millipore VS filters (pore size 0.025μm).

2.3.3 Analysis of genomic DNA by 'Southern blotting'

2.3.3.1 Digestion of genomic DNA

DNA was digested with Type II restriction endonucleases which had been shown to reveal polymorphic variation in and around the α and ß globin gene clusters. The restriction endonucleases were purchased from Amersham, Anglian, Boehringer Mannheim, Bethesda Research Laboratories or Promega, depending on availability and cost.

Digestions were set up according to the manufacturer's specifications using 5-10μg of genomic DNA, the recommended buffer, 1μl of 0.1M spermidine trihydrochloride, which uncoils DNA and aids digestion, 25-30 units of enzyme and distilled water to a final volume of 40μl. Digestions were allowed to proceed for 6-18 hours at the specified temperature, after which 1/10 of the volume was run on a trial gel at 80-100V for 1.5-2 hours to check that the digestion was complete. Complete digestion was indicated by an even smear on the gel, with no high molecular weight DNA present. In the event of incomplete digestion, additional enzyme was added and the sample incubated for an additional 3-4 hours, which generally completed the reaction. Once the samples were completely digested, 5μl of ficoll loading dye was added. This solution stops the reaction, raises the density which facilitates loading, and acts as a tracking dye.

2.3.3.2 Agarose gel electrophoresis

After digestion the samples were loaded onto agarose gels. Agarose acts as a filter, slowing down larger fragments and allowing smaller fragments to migrate further, such that fragments are separated on a gel inversely proportional to the logarithm of their size or molecular weight.

Horizontal slab gels were prepared for gel electrophoresis with agarose concentrations between 0.6 and 1.0%, the higher concentration gels being used for separation of smaller fragments. In general, fragments from about 1-30kb can be effectively separated on agarose gels. Gels were made using the appropriate concentration of Seakem HGT agarose in TBE buffer. The agarose was dissolved by boiling. After the solution had cooled, ethidium bromide was added to a final concentration of 0.5μg/mℓ so that the DNA could be visualised after electrophoresis. The gel mixture was poured into a perspex gel mould (18x20cm), with a gel comb in place to form loading wells, and allowed to set at room temperature.

Prior to electrophoresis, the gel was submerged in TBE buffer, and samples were loaded into individual wells. In addition, two molecular weight markers were loaded on each gel so that the sizes of all fragments could be determined. Boehringer Mannheim markers II and III, consisting of bacteriophage λ digested with *Hin*dIII and *Hin*dIII/*Eco*RI

respectively, were used. The markers could be detected by including radiolabelled bacteriophage λ in the hybridisation mix.

Electrophoresis was carried out at a constant voltage of 35-50V for 15-18 hours at room temperature. The distance the Ficoll dye had moved from the origin was used to standardise the length of each run. Conditions varied slightly, depending on the sizes of the fragments to be separated.

2.3.3.3 Southern blotting

Southern gel transfer allows DNA fragments that have been separated according to size by agarose gel electrophoresis to be transferred *in situ* to a solid support (Southern 1975).

For most of this study the nylon membrane, Hybond-N (Amersham), was used as a solid support and the manufacturer's protocol for Southern blotting and capillary transfer was followed, except that the first depurination step was found to be unnecessary. After electrophoresis, the gel was photographed on a transilluminator (Spectroline TC312A, λ=312nm) using Polaroid 667 film, then denatured, neutralised and placed in 20xSSC. A standard capillary blot was set up (Sambrook *et al.* 1989), using 20xSSC as transfer buffer. Gels were inverted and transfer was allowed to proceed for 24-48 hours to improve the efficiency of transfer and subsequent hybridisation signals, particularly of larger fragments. After transfer, the membranes were rinsed in 2xSSC to remove the residual agarose, baked at 80°C for two hours to fix the DNA and sealed in plastic until required.

2.3.3.4 Radioactive labelling of probe DNA

DNA probes were radioactively labelled by the incorporation of ^{32}P-dCTP using the Amersham Multiprime DNA labelling kit. This is based on the random priming method of Feinberg and Vogelstein (1983) in which random sequence hexanucleotides prime DNA synthesis on a denatured DNA template at numerous sites along its length. This technique, using the Klenow fragment of DNA polymerase I, results in higher specific activities and higher levels of radionucleotide incorporation than nick translation.

Both whole plasmids and purified DNA fragments were labelled using 25ng of DNA and the Amersham RPN 1601 protocol, supplied with the kit. The procedure was modified by adding 1μl of 0.1M spermidine after the reaction had proceeded for five minutes at 37°C. The reaction was allowed to continue for another 15 minutes, prior to terminating it by adding an equal volume of TE. Incorporation was highly efficient and thus not routinely checked.

All reactions were put through TE-saturated Sephadex G50 spin columns to separate the labelled DNA from the unincorporated nucleotides. This reduced the non-specific background on the autoradiographs after hybridisation (Sambrook *et al.* 1989). The volume of the labelled probe was made up to 1mℓ and the specific activity of the probe was estimated by measuring the activity of 10μl in a scintillation counter. Prior to hybridisation, the labelled DNA was denatured by heating to 95-100°C for 10 minutes and then placed on ice.

2.3.3.5 Hybridisation and autoradiography

The Hybond-N filters, prewet in 2xSSC, were placed in plastic bags with 7-10mℓ prehybridisation solution per filter for at least one hour. The solution was modified from the manufacturer's specification, firstly by adding 50% deionised formamide so that the hybridisation temperature could be reduced from 65°C to 42°C, and secondly by incorporating the denatured sonicated herring sperm DNA directly into the solution.

The denatured labelled probe DNA was added to the packet and hybridisation allowed to proceed for 36-42 hours. An increased time of hybridisation appeared to improve the signals considerably.

The hybridisation solution was removed and the filters washed. The washes for pSG21 were: 10 minutes in 3xSSC at room temperature; two washes in 3xSSC at 65°C for 30 minutes each; one wash in 0.5xSSC/0.1% SDS at 65°C for 30 minutes; one wash in 0.2xSSC/0.1% SDS at 65°C for 30 minutes; and two washes in 0.1xSSC/0.1% SDS at 65°C for 30 minutes each.

For all probes, except pSG21, the filters were washed twice in 2xSSPE/0.1% SDS for 15 minutes at room temperature, followed by two 20 minute washes in 1xSSPE/0.1% SDS at 65°C. These washes were sufficient for probes pDHα, pDH12, pε1.3 and pᴬγ. The other probes (p5'HVR.14, pBRζ1, pα3'HVR.64, pP3.9, ßIVS2, pRK29) required two additional higher stringency washes, 0.5xSSPE/0.1% SDS at 65°C for 30 minutes each. In addition, pP3.9 required two additional 0.1xSSPE/0.1% SDS washes for 30 minutes at 65°C followed. All filters were finally rinsed three times in 0.1xSSPE/0.1% SDS and then individually sealed between plastic sheets for autoradiography.

The probe could be reused immediately by adding it to another set of blots which had been prehybridised in plastic bags. The prehybridisation solution was removed and the probe in solution simply added to the blots. If there was a delay before reusing the probe, it was heated to 70°C to ensure that it was denatured.

The filters, sealed individually between plastic, were placed in Okamoto cassettes with Du-Pont Cronex tungstate intensifying screens. An Agfa Curix X-ray film was placed over the filters and the cassette was kept at -70°C for 2-14 days, depending on the intensity of the signal. The film was removed, developed in a Kodak RP X-omat Developer and analysed.

2.3.3.6 Reuse of nylon filters

One of the major advantages of nylon filters is that repeat hybridisation can be done to the same filter without significant loss of signal intensity or membrane disintegration. Prior to rehybridisation, it is necessary to strip the membrane to remove any bound probe from the previous hybridisation. This was achieved by soaking the membrane for 10 minutes in denaturing solution and then soaking it twice for 10 minutes in a neutralising solution, all at room temperature. The filters were stored in 2xSSC at 4°C until required.

2.3.4 Approach to characterisation of individuals using 'Southern blotting'

Both the α and ß globin gene clusters of all individuals were investigated in order to determine the presence of gene rearrangements, the allele frequencies of different RFLPs

and to characterise the α and ß globin haplotypes present. Prior to RFLP and haplotype analysis, it is important to assess the number of genes on a chromosome, as this may influence the interpretation of results.

2.3.4.1 Variation in the α globin gene cluster

It was found that the pBRζ probe hybridised to DNA digested with *Eco*RI was the most efficient way to screen for α cluster rearrangements, as both variations in the number of ζ and α genes could be detected. Rearrangements of the ζ and α genes were confirmed on both *Bam*HI and *Bgl*II digests with the pBRζ1 and pDHα probes respectively. The DNA fragments produced by the different rearrangements are shown in Table 2.3.

The $-\alpha^{3.7}$ chromosomes were subtyped using *Apa*I digests and the pDHα probe. While the αα chromosome is recognised by fragments of 2.7kb, 1.7kb and 0.89kb, the $-\alpha^{3.7I}$ is associated with a 2.5kb fragment, the $-\alpha^{3.7II}$ with a 3.5kb and the $-\alpha^{3.7III}$ with 2.7kb and 0.7kb fragments (Higgs *et al.* 1984).

The polymorphic sites studied in the α cluster are shown in Figure 2.1. Details of the associated probes, enzymes and fragment sizes for ζζαα chromosomes are shown in Table 2.4. Alterations produced by the different rearrangements are not shown.

The α globin haplotypes were constructed using RFLP sites 2-9 and typed using the nomenclature of Higgs *et al.* (1986). The enzyme *Acc*I was excluded from the analysis as it was too costly. Thus types II and VI haplotypes and haplotypes IIIe and VIIa could not be distinguished, as they differ only at the *Acc*I site. Family studies were used to determine the linkage phase of the markers, though it was not always possible to assign haplotypes unequivocally. The linkage phase of sites 1 and 10 with the haplotypes was also determined.

2.3.4.2 Variation in the ß globin gene cluster

γ Gene rearrangements were detected using *Bgl*II and the probe, pγ^. The γγγ chromosome is associated with an 18kb fragment, the normal γγ with a 13kb and the -γ

TABLE 2.3 DNA fragments produced by ζ and α gene rearrangements[1]

PROBE	ENZYME	CHROMOSOME			
		ζζαα	ζζζαα	-ζαα	ζζ--$^{SEA\,8}$
pBRζ1	*Eco*RI	23.0 5.0	23.0 11.5 5.0	18.0	17.2 5.0
	*Bam*HI	9.1-14.0[2] 5.9	9.1-14.0[2](2)[3] 5.9	6.5 or 6.2	20.0 5.9
	*Bgl*II	12.6 10.0-12.0[2]	12.6 10.0-12.0[2](2)[3]	12.6	10.5
		ζζαα	ζζαααα[4]	ζζ-$\alpha^{3.7}$	ζζ-$\alpha^{4.2}$
pDHα	*Eco*RI[5]	23.0	27.0	19.0	19.0
	*Lam*HI	14.5	18.0	10.3[6]	9.8[6]
	*Bgl*II	12.6[7] 7.4	12.6[7] 7.4 3.7	15.8	8.7

[1] Embury *et al.* 1980a, Goossens *et al.* 1980, Pressley *et al.* 1980, Higgs *et al.* 1981, Trent *et al.* 1981b, Winichagoon *et al.* 1982, Lie Injo *et al.* 1985, Felice *et al.* 1986.
All fragment sizes are approximate and are in kb.

[2] Fragment size may vary as it spans the inter-ζ hypervariable region.

[3] Two fragments in the size range are present.

[4] Fragments are shown for the $\alpha\alpha\alpha^{anti3.7}$. They are similar for $\alpha\alpha\alpha^{anti4.2}$ with *Eco*RI and *Bam*HI, but with *Bgl*II, fragments of 16.0 and 7.4 are seen.

[5] These fragments can be detected with pBRζ1 as well.

[6] Fragments are difficult to distinguish, unless in the same lane.

[7] The presence of an additional *Bgl*II polymorphism may result in a 5.2kb fragment with pBRζ1 or a 7.4kb fragment with pDHα in place of the 12.6kb fragment.

[8] The --^{SA} variant hybridised with the pBRζ1 probe produces fragments of 7.0kb with *Bgl*II, 5.9 with *Bam*HI and 5.0kb with *Eco*RI (Vandenplas *et al.* 1987).

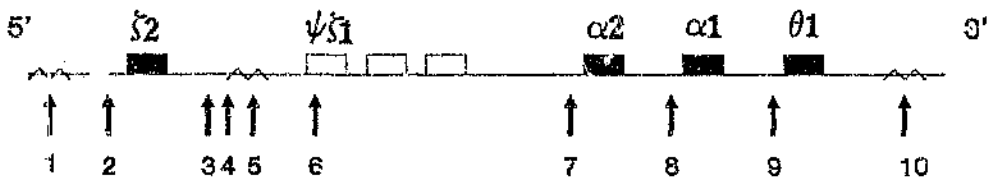


FIGURE 2.1 Polymorphic sites studied in the α globin cluster

1: 5′ hypervariable region, 2: *Xba*I, 3: *Sst*I, 4: *Bgl*I, 5: Interζ hypervariable region, 6: *Sst*I site, distinguishing $\psi\zeta$1 (PZ) from ζ1 (Z) form of gene, 7: *Rsa*I, 8: *Pst*I, 9: *Pst*I, 10: 3′ hypervariable region.

TABLE 2.4 α globin gene cluster RFLPs

PROBE	RFLP SITE[1]	FRAGMENT FOR OLIGOLABELLING	ENZYME	FRAGMENT SIZES[2]	REFERENCE
p5'HVR.14	1	770bp *Eco*RI/*Hin*dIII	*Rsa*I	1.9-10.2kb	Jarman and Higgs 1988
pBRζ1	2	1.9kb *Hin*fI	*Xba*I	20kb (-) 16kb (+)	Felice *et al.* 1986
pSG21	3,5	1.1kb *Alu*I	*Sst*I	5.7-7.2kb(-)[3] 3.0-5.0kb(+)	Goodbourn *et al.* 1983, 1984
pSG21	4,5	1.1kb *Alu*I	*Bgl*I	7.5-9.5kb(-)[3] 4.2-4.5kb(+)	Higgs *et al.* 1986
pBRζ1	6	1.16kb *Bgl*II/*Pvu*II	*Sst*I	1.85-2.0kb(PZ)[4] 1.55kb(Z)	Hill *et al.* 1985a
pDHα	7	1.5kb *Pst*I	*Rsa*I	3.5kb (-) 2.4kb (+)	Higgs *et al.* 1984
pDHα	8	1.5kb *Pst*I	*Pst*I	1.5kb (-) 1.3kb (+)	Higgs *et al.* 1986
pDH12	9	0.8kb *Bam*HI	*Pst*I	2.9kb (-) 1.9kb (+)	Higgs *et al.* 1986
α3'HVR	10	4kb *Hin*fI	*Pvu*II	1.6-9.4kb	Jarman *et al.* 1986

[1] The positions of RFLP sites are as shown in Figure 2.1.

[2] Fragment sizes are shown for chromosomes with the normal complement of genes ($\zeta\zeta\alpha\alpha$). Rearrangements in the cluster may alter these. - and + denote the absence and presence of a polymorphic restriction enzyme site respectively. A range of fragment sizes are seen where a VNTR region is included in the analysis.

[3] Fragments span the interζ hypervariable region, which has alleles broadly classified into S, M and L.

[4] Fragments span the hypervariable region of the first intron of the ζ1 gene. ψζ1 forms of the gene are designated PZ and ζ1 forms Z.

with an 8kb (Trent *et al.* 1981a, Sukumaran *et al.* 1983). The $^{G}\gamma^{G}\gamma$ and $^{A}\gamma^{A}\gamma$ rearrangements were sought using *Pst*I and the probe, $p\gamma^{A}$ (Powers *et al.* 1984).

The polymorphic sites studied in the ß cluster are shown in Figure 2.2, with details of the probes, enzymes and fragment sizes shown in Table 2.5. Sites 1, 3-7 and 9 can now be studied using PCR (as described in Section 2.3.5.1).

ß globin haplotypes were constructed using sites 1 and 3-9 and family studies were again used to establish the linkage phase. The linkage phase of site 2 with the haplotypes was also determined.

2.3.5 Analysis of genomic DNA using the polymerase chain reaction (PCR)

The PCR reaction, first described by Saiki *et al.* (1985) and Mullis and Faloona (1987), enables DNA to be amplified rapidly *in vitro* and, compared with Southern blotting, reduces the time for analysis considerably. As the technique only became available relatively late in the study, most of the work was done by Southern blotting. However, it was used for mutation analysis, DNA sequencing and limited RFLP analysis.

All PCR reactions were done in a Perkin-Elmer Cetus or a Hybaid thermocycler, in a 25-50μl volume, using 2 units of Promega Taq Polymerase, 1/10 volume of the recommended buffer (final concentration 50mmol/ℓ KCl, 10mmol/ℓ Tris HCl, 1.5mM $MgCl_2$), 0.5-1xdNTPs (62.5-125μM of each nucleotide) and 5-12.5pm of each primer.

2.3.5.1 RFLP analysis

Towards the end of this study, primer sequences became available that allowed detection of all but one of the RFLPs in the ß globin haplotype. No primers are available for the *Hin*dIII/3'ß RFLP as the sequence of the region has not been published (Dr J Old, personal communication 1990). The primer sequences are shown in Table 2.6, together with the enzyme/probe combinations for the corresponding RFLP sites, and the sizes of the fragments produced after amplification and digestion respectively.

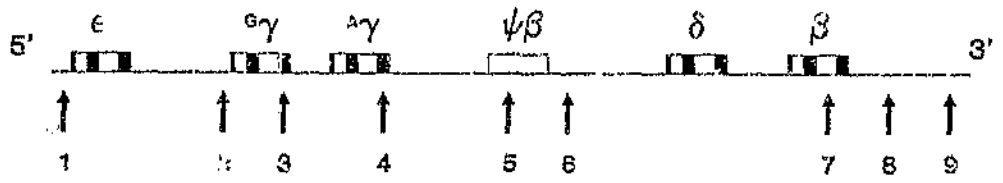


FIGURE 2.2 Polymorphic sites studied in the β globin cluster

1: *Hin*cII, 2: *Xmn*I, 3: *Hin*dIII, 4: *Hin*dIII, 5: *Hin*cII, 6: *Hin*cII, 7: *Ava*II, 8: *Hin*dIII, 9: *Bam*HI.

TABLE 2.5 β globin gene cluster RFLPs

PROBE	RFLP SITE[1]	FRAGMENT FOR OLIGOLABELLING	ENZYME	FRAGMENT SIZES[2]	REFERENCE
pa1.3	1	1.3kb *Bam*HI/*Eco*RI	*Hinc*II	8.0kb (-) 3.7kb (+)	Antonarakis *et al.* 1982a
$p^A\gamma$	2	3.2kb *Hind*III	*Xmn*I	8.0kb (-) 7.0kb ( )	Gilman and Huisman 1985
$^A\gamma$	3,4	3.2kb *Hind*III	*Hind*III	$^G\gamma$: 7.8kb (-) 7.1kb (+) $^A\gamma$: 3.5kb (-) 2.8kb (+)	Jeffreys 1979 Tuan *et al.* 1979
pP3.9	5,6	1.7kb *Bgl*II/*Xba*I	*Hinc*II	7.6kb (--) 6.0kb (-+) 4.6kb/3.0kb (+-) 3.0kb/3.0kb (+ +)	Antonarakis *et al.* 1982a
βIVS2	7	0.92kb *Bam*HI/*Eco*RI	*Ava*II	2.2kb (-) 2.0kb (+)	Antonarakis *et al.* 1982a
pRK29	8	0.9kb *Bgl*II/*Eco*RI	*Hind*III	15.5kb (-) 13.5kb (+)	Tuan *et al.* 1983
βIVS2	9	0.92kb *Bam*HI/*Eco*RI	*Bam*HI	22.0kb (-) 9.3kb (+)	Kan *et al.* 1980

[1] The positions of the RFLP sites are as shown in Figure 2.2.
[2] - and + denote the absence or presence of the polymorphic restriction enzyme site respectively.

TABLE 2.6 PCR primers for β globin gene cluster RFLP analysis

RFLP SITE[1]	PRIMER SEQUENCES[2,3]	PRODUCT SIZE (bp)	FRAGMENTS AFTER DIGESTION[4] (bp)		
			-	+	C
*Hinc*II/ε (1)	TTAAGAGAGCTAGAACTGGTGAG AAGCCTCATATAAAGGAGCAAATC	158	158	86/72	-
*Hin*dIII/Gγ (3)	AGTGCTGCAAGAAGAACAACTACC CTCTGCATCATGGGCACTGAGCTC	323	323	235/98	-
*Hin*dIII/Aγ (4)	ATGCTGCTAATGCTTCATTAC TCATGTGTGATCTCTCTCAGCAG	635	635	327/308	-
*Hinc*II/5'ψβ (5)	TCCTATCCATTACTGTTCCTTGAA ATTGTCTTATTCTAGAGACGATTT	794	794	687/107	-
*Hinc*II/3'ψβ (6)	GTACTCATACTTTAAGTCCTAACT TAAGCAAGATTATTTCTGGTCTCT	914	914	480/434	-
*Ava*II/β (7)	ACTCCCAGGAGCAGGGAGGGCAGG TTCGTCTGTTTCCCATTCTAAACT	676	315	214/101	361
*Hin*fI/β[5] (9)	TGGATTCTGCCTATTAAAA GGGCCTATGATAGGGTAAT	741	341	213/128	244/154

[1] The numbers correspond to the positions of the RFLP sites shown in Figure 2.2.

[2] Primer sequences were kindly supplied by Dr J Old in 1990. The *Hin*dIII/^{G}γ and *Hin*fI/β sequences were previously published (Kulozik *et al.* 1988b, Semenza *et al.* 1989). The *Hin*dIII/^{A}γ primer sequences are those published by Varawalla *et al.* 1992).

[3] Primer sequences are shown from 5' to 3' for the forward and reverse primers respectively.

[4] - and + denote absence and presence of the polymorphic site respectively. C denotes constant bands.

[5] This site has been shown to be in complete linkage disequilibrium with the *Bam*HI site and can thus be used in place of it (Semenza *et al.* 1989). If the Asian Indian 619bp deletion is present, the binding site for the forward primer is deleted. There is therefore no amplification from the chromosome carrying the deletion.

The PCR conditions and methods used followed those described previously (Old *et al.* 1990). A final concentration of 62.5μM of each of the deoxynucleotides and 5pm of each primer was used, with the exception of the *Hin*fI/ß system where 12.5pm of each primer was required for efficient amplification. Samples were denatured at 93°C for seven minutes, after which annealing was allowed to occur for seven minutes at 53°C. During the annealing period the enzyme was added, a so-called 'hot-start'. Thirty step cycles of denaturing at 93°C for 60 seconds, annealing at 65°C for 60 seconds and extension at 72°C for 90 seconds, respectively, were followed by a final extension step at 72°C for three minutes. Conditions were identical for all sets of primers, with the exception of primers *Hin*fI/ß and *Hae*I 3'ψß which gave better results with annealing temperatures of 59°C and 61°C respectively.

Products were digested after amplification by adding 5-10 units of the appropriate enzyme, together with the recommended buffer, directly into the PCR tube and incubating for 1-2 hours at the recommended temperature for the enzyme. Digested products were run on 1.5% Nusieve/1.5% HGT composite agarose gels for all the systems except *Hin*cII/ε and *Hin*fI/ß, where a 3% Nusieve/1% HGT composite agarose gel was required to separate the fragments.

Although the bulk of the population screening work was completed using Southern blotting, the primers have been used in a number of prenatal diagnoses, where the time for analysis has been reduced dramatically (see Section 7.2.3.3.2)

2.3.5.2 Mutation analysis

2.3.5.2.1 The ARMS technique for detection of ß thalassaemia mutations

The ARMS (Amplification Refractory Mutation System) technique, originally described by Newton *et al.* (1989), was applied by Old *et al.* (1990) to the detection of ß thalassaemia mutations. It relies on the principle that an oligonucleotide with a 3' mismatch will not function as a primer for PCR amplification under controlled conditions. Thus, an oligonucleotide complementary to the normal sequence will not amplify the

DNA if a mutation is present at the 3′ site. Similarly, an oligonucleotide complementary to the mutant sequence at the 3′ site will not amplify the normal DNA. Absence or presence of a particular amplification product can therefore indicate the absence or presence of a mutation. Specificity of primers is increased by introducing a deliberate mismatch four base pairs away from the end of the primer. A second set of primers is introduced into the reaction as a control, so that non-amplification due to mismatch can be distinguished from non-amplification for other technical reasons (Old *et al.* 1990).

In order to identify the ß thalassaemia mutations present in the South African cohort, we first aimed to identify most of the 17 previously described Asian Indians mutations (Kazazian *et al.* 1984a, Thein *et al.* 1988, Old *et al.* 1990, Varawalla *et al.* 1991b,c). The two mutations described most recently (Garewal *et al.* 1994) were not specifically sought, as the data were published after completion of the study.

The ARMS technique was used to screen for 13 mutations, including the five thought to be most common (Thein *et al.* 1988, Varawalla *et al.* 1991b), namely IVS1nt5, IVS1nt1, codon 41/42, codon 8/9 and the 619bp deletion, the latter detectable with the control primers. The primers used are shown in Tables 2.7 and 2.8, together with the mutations they detect and the size of the amplified product. Primers for the normal DNA sequence at the positions of the four common mutations were also synthesised so that heterozygotes and homozygotes could be distinguished (Old *et al.* 1990). These are also shown in Table 2.7.

Primers for the remaining four rare mutations, namely codon 5 (-CT), codon 30 (G→A), codon 88 (+T) and IVS1nt110 (G→A), were not synthesised as the mutations were considered rare and DNA sequencing was considered a more cost-effective way to screen for them. DNA sequencing also covered the regions where two new mutations have recently been described (Garewal *et al.* 1994).

In all reactions two sets of primers were used, those designed to detect the sequence at the site of the mutation and a control set of primers to ensure that the DNA had amplified. 5pM of each primer were used, as well as 62.5 μM of each nucleotide and 0.25mM spermidine trihydrochloride, found to increase amplification specificity (Straub

TABLE 2.7 PCR primers for detection of the five common Asian Indian mutations

ALLELE-SPECIFIC PRIMER[1]	PRIMER SEQUENCE[2]	PAIRED PRIMER[3]	PRODUCT SIZE (bp)	CONTROL PRIMERS[4]
IVS1nt5 (G→C) m	CTCCTTAAACCTGTCTTGTAACCTTGTTAG	C	285	A/B
IVS1nt5 n	CTCCTTAAACCTGTCTTGTAACCTTGTTAC	C	285	A/B
IVS1nt1 (G→T) m	TTAAACCTGTCTTGTAACCTTGATACGAAA	C	281	A/B
IVS1nt1 n	GATGAAGTTGGTGGTGAGGCCCTGGGTAGG	D	450	A/B
Codon 41-42 (-CTTT) m	GAGTGGACAGATCCCCAAAGGACTCAACCT	C	439	A/B
Codon 41-42 n	GAGTGGACAGATCCCCAAAGGACTCAAAGA	C	443	A/B
Codon 8-9 (+G) m	CCTTGCCCCACAGGGCAGTAACGGCACACC	C	215	A/B
Codon 8-9 n	CCTTGCCCCACAGGGCAGTAACGGCACACT	C	214	A/B

[1] m = mutation-sequence specific, n = normal-sequence specific.
[2] Primer sequences were kindly supplied by Dr J Old, Oxford in 1990 and all sequences are shown from 5′ to 3′.
[3] Sequence of primer C: ACCTCACCCTGTGGAGCCAC
Sequence of primer D: CCCCTTCCTATGACATGAACTTAA
[4] Sequence of primer A: CAATGTATCATGCCTCTTTGCACC
Sequence of primer B: GAGTCAAGGCTGAGAGATGCAGGA
Primers A and B normally amplify a fragment of 861bp. They span the common Asian Indian 619bp deletion which, if present, results in a fragment of 242bp (Old *et al.* 1990).

TABLE 2.8 PCR primers for detection of eight of the rare Asian Indian mutations

ALLELE-SPECIFIC PRIMER[1]	PRIMER SEQUENCE[2]	PAIRED PRIMER[3]	PRODUCT SIZE (bp)	CONTROL PRIMERS[3]	REFERENCE
Codon 15 (G→A) m	TGAGGAGAAGTCTGCCGTTACTGGCCAGTA	D	500	A/B	Dr J Old, personal communication 1990
Codon 16 (-C) m	TCACCACCAACTTCATCCACGTTCACGTTC	C	238	A/B	Dr J Old, personal communication 1990
-88 (C→T) m	TCACTTAGACCTCACCCTGTGGAGCCTCAT	D	655	A/B	Dr J Old, personal communication 1990
Cap +1 (A→C) m	ATAAGTCAGGGCAGAGCCACTTATTGGTTC	D	567	A/B	Dr J Old, personal communication 1990
IVS1 (-25bp 3') m	CTCTGGGTCCAAGGGTAGACCACCAGCATA	C	354	A/B	Dr J Old, personal communication 1990
Codon 30 (G→C) m	TAAACCTGTCTTGTAACCTTGATACCTACG		280	A/B	Varawalla *et al.* 1991b
IVS2nt1 (G→A) m	AAGAAAACATCAAGGGTCCCATAGACTGAT	C	634	A/B	Varawalla *et al.* 1991b
IVS2nt837 (T→G) m	CCTTTTGCTAATCATGTTCATACCTCGTAG	B	646	$^{G}\gamma$ RFLP	Varawalla *et al.* 1991c

[1] m = mutation-sequence specific.
[2] All primer sequences are shown from 5' to 3'.
[3] Sequences of primers A, B, C and D are shown below Table 2.6.
The sequences of the $^{G}\gamma$ RFLP primers are shown in Table 2.5.

and Bale 1990). All amplifications were done using the same PCR programme as that used for the RFLPs, except that the number of cycles was reduced to 25 to prevent the generation of misprimed products in significant quantities. The annealing temperature for IVS2nt837 m (Table 2.8) was reduced to 64°C.

In each set of reactions, negative controls were included and a positive control, where possible. When a mutation was identified in a proband, its presence in one or both parents was always confirmed. Apart from using the ARMS technique to screen for ß thalassaemia mutations in the population survey, it has also been used for prenatal diagnosis (see Section 7.2.3).

Products were loaded on ethidium bromide stained 1.5% Nusieve/1.5% HGT composite agarose gels and run for 1-2 hours at 100V. Gels were photographed on an ultraviolet transilluminator and analysed.

2.3.5.2.2 Detection of the $ß^S$ mutation

The sickle cell mutation was detected by PCR amplification of the 5′ region of the ß globin gene, followed by *Dde*I digestion. This was used for a number of prenatal diagnoses.

The primers, synthesised based on information supplied by Dr J Old (personal communication 1990), were:

5′: GGCCAATCTACTCCCAGGAG

3′: ACATCAAGGGTCCCATAGAC

5pM of each primer and 125μM of each deoxynucleotide were used in the reaction, which consisted of 30 cycles of denaturation at 94°C for 48 seconds, annealing at 59°C for 48 seconds and extension at 72°C for 90 seconds, after an initial denaturation and annealing step as in the RFLP reaction (see Section 2.3.5.1). A final extension period of 10 minutes at 72°C completed the reaction.

The product of 605bp was digested with *Dde*I and run on a 3% Nusieve/1% HGT

composite agarose gel. The presence of a $ß^S$ mutation results in a fragment of 351bp, while the $ß^A$ chromosome results in fragments of 201bp and 150bp. Constant fragments of 89bp, 88bp, 40bp and 37bp are also produced. The results of a typical reaction are shown in Figure 2.3.

2.3.5.2.3 Detection of the $ß^E$ and $ß^{D\ Punjab}$ mutations

The $ß^E$ mutation can be detected by amplification of part of the ß globin gene with primers codon 41/42n and C (described in and below Table 2.7 respectively), followed by *Mnl*I digestion (Old and Ludlam 1991), as the $ß^E$ mutation abolishes an *Mnl*I site (Thein *et al.* 1987b).

The presence of the $ß^E$ mutation is indicated by a 230bp fragment, while the $ß^A$ gene results in fragments of 170bp and 60bp. These can be detected on a 3% Nusieve/1% HGT composite agarose gel.

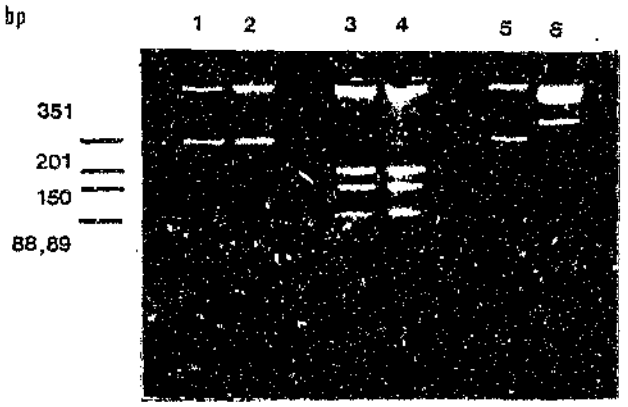
The $ß^{D\ Punjab}$ mutation can be detected by amplification of part of the ß globin gene with primers A and B (described below Table 2.7), followed by *Eco*RI digestion (Old and Ludlam 1991), as the $ß^{D\ Punjab}$ mutation abolishes an *Eco*RI site (Trent *et al.* 1984).

The presence of the $ß^{D\ Punjab}$ mutation is indicated by a 867bp fragment, whereas the $ß^A$ gene results in fragments of 558bp and 309bp.

Neither of these detection methods has been used yet for a prenatal diagnosis, although families with these haemoglobin variants have requested workups.

2.3.5.3 DNA sequencing

Direct DNA sequencing of the ß globin gene from PCR products was carried out in an attempt to identify the unknown mutations on the remaining ß thalassaemia chromosomes, where no mutations had been detected with the ARMS technique. Primers that had already been synthesised as part of the ARMS screening technique were used to limit costs. The primer pairs C and D and A and B were used to amplify the two fragments of the ß



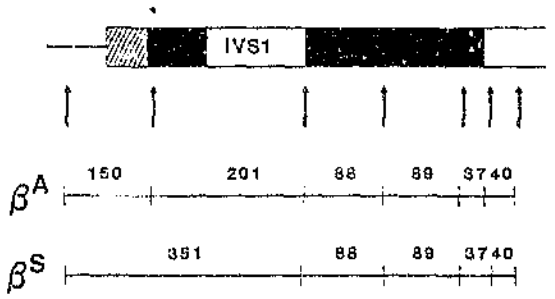


FIGURE 2.3 Detection of the β^S mutation by PCR.

Lanes 1,2 β^S homozygotes, lanes 3,4 β^A homozygotes, lane 5 β^S/β^A heterozygote. lane 6 1kb molecular size marker.
The fragments produced after *Dde*I digestion of the 695bp product are indicated alongside the photograph and shown schematically below the photograph for the β^A and β^S genes, respectively.
The * indicates codon 6 of the β globin gene, the site of the β^S mutation.

globin gene shown in Figure 2.4. They span the areas in which the majority of ß thalassaemia mutations has been found (Kazazian 1990). Due to the cost of primer synthesis, it was not possible to synthesise additional primers to span the remainder of the gene.

The PCR conditions were the same as those used for the RFLP analysis (see Section 2.3.5.1). The product was run into a 1% low-melting point agarose gel, until the product could be seen as a distinct band. A DNA plug containing the product was removed with a pipette tip and placed in 500μl of distilled water. This was heated to 65°C for 5-10 minutes and vortexed vigorously to dissolve the agarose, following which 10μl of the solution was used to set up a second round of PCR in which conditions were again identical to those used in Section 2.3.5.1, except the number of cycles was increased to 35 and one of the primers in each reaction was diluted 25-100 fold to obtain single-stranded product for sequencing. An aliquot of the reaction product was again run on a gel to check that asymmetric product was obtained. In both sets of reactions it was only possible to obtain single-stranded product in one direction, namely when primers D and A respectively were in excess, despite numerous attempts at altering reaction conditions.

The PCR product was purified on a Millipore Ultrafree-MC 30 000 NMWL Filter Unit. TE was passed through the filter three times, by spinning the unit at 5 000rpm for five minutes. The DNA, which was used as a sequencing template, was recovered from the membrane, resuspended in 30-50μl of TE and stored at 4°C.

DNA sequencing was performed by the dideoxy chain termination method of Sanger *et al.* (1977) following the Sequenase version 2 protocol (United States Biochemical), supplied with the Sequenase kit. 7μl of template DNA was used, together with 4pm of the appropriate sequencing primer, either that which was limiting in the PCR (C or B) or the IVS1nt1 n primer as an additional internal primer with the C/D product (see Figure 2.4). Samples were radioactively labelled using ^{35}S-dATP and both 1/5 and 1/20 dilutions of the labelling mix were used, so that a greater area of the sequence could be read.

After the sequencing reaction was complete, samples were run on a 0.2mm 6% denaturing polyacrylamide gel at 50-55°C and 1200-1400V for 1,5-4 hours, depending on the region

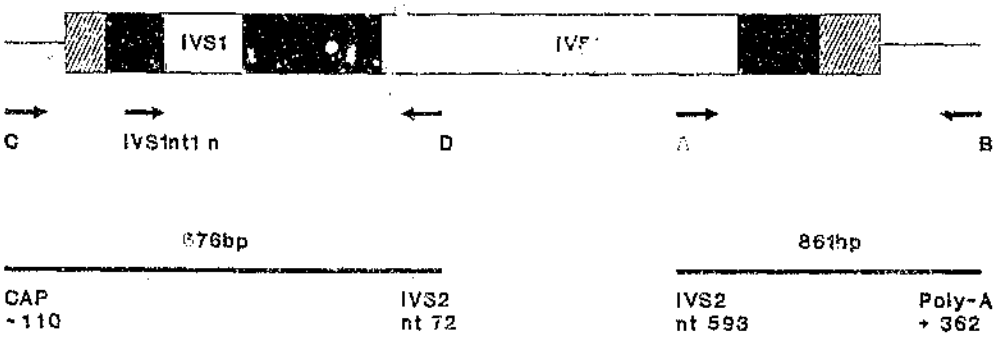


FIGURE 2.4 Regions of the β globin gene amplified for sequencing

The sequences of primers A, B, C and D are listed below Table 2.7, while that of IVS1nt1 n is shown in Table 2.7. The DNA sequences of the regions from -70 to the cap site to IVS2nt33 and from +34 to the termination codon to +300 to the Poly-A site could be clearly read on sequencing gels. As sequencing was only done in one direction and intermediate primers for the A/B product were unavailable, the complete region amplified could not be sequenced.

of the sequence which was to be read. Gels were fixed with a 10% methanol:10% acetic acid solution and transferred on to filter paper, prior to drying and autoradiography.

All the samples sequenced came from individuals who were compound heterozygotes for two different ß thalassaemia mutations. Thus, a point mutation appeared as two bands at the same position on a sequencing gel and a frameshift mutation appeared as a double ladder from the site of the mutation. Once a mutation had been identified in a proband its presence was confirmed by sequencing the same region in the parent whose mutation was unknown and also sequencing DNA from a normal individual to exclude artefacts in the region.

2.4 Statistical analysis of results

The student's t-test was used to compare normally distributed data, while a chi-squared test was used for comparison of allele frequencies between populations. Genetic diversity was calculated according to Nei (1987).

CHAPTER 3 - A HAEMATOLOGICAL PROFILE OF THE SOUTH AFRICAN ASIAN INDIAN POPULATION

In this chapter, a haematological profile of the South African Asian Indian population is presented, based mainly on the results of the survey of Lenasia schools. The survey also provided data from which the frequencies of the ß thalassaemia and $ß^S$ alleles in the major religious and linguistic subgroups could be estimated and the major causes of microcytosis, hypochromia and/or anaemia could be determined. Minimum estimates of α thalassaemia frequencies in the different subgroups, based on the survey, were refined using molecular studies on the parents of the randomly collected families. These data were used to assess the value of a screening programme for carriers of the ß thalassaemia and $ß^S$ alleles since these disorders, in homozygous form, cause serious and disabling diseases, and thus present a major drain on health resources (see Section 7.1).

3.1 Results

3.1.1 Haematological parameters

Haematological parameters were determined on an electronic cell counter. Table 3.1 shows the means and standard deviations of the red cell count (RCC), haemoglobin level (Hb), mean cell volume (MCV) and mean cell haemoglobin (MCH) for the total Lenasia sample studied (including the minor religious and language groups) and, individually, for the major religious and linguistic subgroups. The mean values in the different groups were compared using the Student's t-test.

Since gender influences the haematological parameters, each group, except the Hindu Hindi, was divided into males and females for comparison, and also compared as a complete group, with the sexes combined. The Hindu Hindi group was considered too small to divide by sex and the whole group was not used in statistical comparisons with the other religious and language groups, as there were uneven numbers of males and females.

TABLE 3.1 Results of red blood cell parameters determined in the Lenasia schools survey

GROUP	n[1]	RCC ($x10^{12}/\ell$)		Hb (g/dℓ)		MCV (fℓ)		MCH (pg)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Lenasia[2]	638	5.33	0.56	15.0	1.8	85.3	7.0	28.3	2.8
Male	329	5.64	0.46	15.2	1.4	86.0	6.9	28.8	2.5
Female	309	4.99	0.45	13.8	1.4	84.6	7.0	27.8	3.0
Moslem Gujerati	305	5.36	0.56	15.1	1.8	85.3	6.1	28.2	2.7
Male	155	5.66	0.47	16.2	1.3	86.2	5.0	28.7	2.2
Female	150	5.06	0.47	13.9	1.5	84.4	7.0	27.7	3.0
Hindu Gujerati	175	5.29	0.59	14.7	2.1	84.4	9.4	28.0	3.3
Male	89	5.64	0.51	16.0	1.7	84.9	10.4	28.6	3.0
Female	86	4.94	0.43	13.4	1.6	83.6	8.2	27.3	3.4
Hindu Tamil	115	5.21	0.47	15.1	1.5	87.0	4.6	29.1	2.0
Male	57	5.51	0.32	16.3	0.9	87.5	3.9	29.6	1.6
Female	58	4.92	0.40	14.0	1.1	86.4	5.2	28.6	2.2
Hindu Hindi	21	5.59	0.56	15.5	1.5	84.8	5.8	27.8	2.5

[1] n = number of individuals studied.
[2] Lenasia refers to the total sample studied in the Lenasia schools survey, including the minor religious and language groups.

The RCCs, Hbs and MCHs of the males were significantly higher than those of the females ($p<0.001$) in all groups studied (namely the total Lenasia sample, the Moslem Gujerati, Hindu Gujerati and Hindu Tamil). While the males' MCVs were higher than those of the females in all the groups, the differences between the sexes in the total Lenasia sample and the Moslem Gujerati were the only ones to reach significance ($p<0.05$)

When the subgroups, consisting of approximately equal numbers of males and females, were compared, the Moslem Gujerati had the highest mean RCC, the Hindu Tamil the lowest and the Hindu Gujerati were intermediate. The RCCs of the Moslem Gujerati and Hindu Tamil were significantly different in the combined and male samples ($p<0.05$). Hb levels were similar in the Moslem Gujerati and Hindu Tamil, but lower in the Hindu Gujerati. Hb levels in Hindu Tamil females were significantly higher than in Hindu Gujerati females ($p<0.01$). The Hindu Tamil had the highest mean MCV and MCH, followed by the Moslem Gujerati and the Hindu Gujerati. The mean MCV in the Hindu Tamil was significantly higher than in both the Moslem Gujerati and the Hindu Gujerati in the combined ($p<0.001$) and female ($p<0.05$) samples. Similarly the MCH was significantly higher in the Hindu Tamil than in both the Hindu Gujerati and Moslem Gujerati combined ($p<0.001$ for both these groups) and male ($p<0.05$) samples and than Hindu Gujerati females ($p<0.05$).

Iron studies and HbF determinations were done on as many samples as were available. Table 3.2 shows the mean and standard deviations for the serum iron, transferrin, percentage saturation of transferrin and HbF for the total Lenasia group studied as well as for the major subgroups. In some cases, the samples were insufficient for all the assays and thus the numbers are reduced from those presented in Table 3.1. The Hindu Hindi were again excluded from the comparative analysis because of the small sample, with uneven numbers of males and females.

The mean values in the subgroups were compared using the Student's t-test. The mean serum iron levels and percentage transferrin saturation were higher ($p<0.001$) and the transferrin level was lower ($p<0.001$) in males compared to females for the total Lenasia

TABLE 3.2 Results of iron studies and HbF determination in the Lenasia schools survey

GROUP	n[1]	SERUM IRON (μmol/ℓ)		TRANSFERRIN (g/ℓ)		TRANSFERRIN SATURATION (%)		n[2]	HbF (%)	
		Mean	SD	Mean	SD	Mean	SD		Mean	SD
Lenasia[3]	560	17.4	7.8	3.46	0.71	20.3	10.6	629	0.5	0.3
Male	289	20.0	7.9	3.30	0.67	24.2	10.9	325	0.5	0.3
Female	271	14.6	6.6	3.62	0.72	16.2	8.4	304	0.5	0.3
Moslem Gujerati	257	17.4	7.4	3.41	0.71	20.6	10.3	299	0.5	0.3
Male	135	19.9	7.5	3.27	0.70	24.3	10.6	153	0.5	0.3
Female	122	14.5	6.2	3.56	0.70	16.5	8.1	146	0.5	0.3
Hindu Gujerati	155	16.7	8.7	3.66	0.69	18.6	11.3	172	0.4	0.2
Male	75	19.8	9.1	3.45	0.68	23.2	12.2	87	0.4	0.2
Female	80	13.8	7.1	3.86	0.66	14.2	8.4	85	0.4	0.2
Hindu Tamil	108	18.5	7.6	3.30	0.66	22.2	10.4	116	0.4	0.4
Male	53	21.5	15.6	3.08	0.48	26.9	10.2	58	0.4	0.2
Female	55	15.6	6.5	3.50	0.74	17.7	8.5	58	0.5	0.5
Hindu Hindi	18	17.6	7.5	3.53	0.60	19.3	8.8	21	0.4	0.2

[1] n = number of individuals for iron studies.
[2] n = number of individuals for HbF determination.
[3] Lenasia refers to the total sample studied in the Lenasia school survey, including the minor religious and language groups.

group and the subgroups studied. The Hindu Gujerati had the highest transferrin levels and the lowest serum iron and percentage transferrin saturation. The Hindu Tamil had the lowest transferrin levels, with the highest serum iron and percentage transferrin saturation, while the Moslem Gujerati were intermediate. Transferrin levels were significantly higher in the Hindu Gujerati than in the Hindu Tamil and Moslem Gujerati in the male, female and combined samples ($p<0.001$). Similarly the transferrin saturation was significantly lower in the Hindu Gujerati than in the Hindu Tamil ($p<0.001$). Average HbF levels were higher in the Moslem Gujerati than in the Hindu Gujerati ($p<0.001$) or Hindu Tamil ($p<0.0^{5}$ In general, HbF levels were <1% in individuals with normal haematology. A few individuals with levels between 1% and 2% were found, the majority being carriers of either ß thalassaemia or $ß^S$. One Hindu Tamil individual with normal haematology and a normal HbA_2 had a HbF of 3.6%. Unfortunately it was not possible to obtain additional samples from her or her family to determine the cause.

3.1.2 ß thalassaemia, $ß^S$ and other ß globin variants

HbA_2 levels, which are raised in heterozygous ß-thalassaemia, were determined on all 171 individuals who had microcytosis (MCV<80fℓ), hypochromia (MCH<27pg) and/or anaemia (Hb<13g/dℓ in males and <12g/dℓ in females). The MCV/RCC ratio was not useful, since it is not diagnostic in cases with concurrent iron deficiency (see section 7.1.3). ß thalassaemia heterozygotes were identified in the Moslem Gujerati and Hindu Gujerati groups. These results were confirmed on a repeat blood sample. The estimated frequencies are shown in Table 3.3. ß thalassaemia alleles do occur in the other major groups, Hindu Hindi and Hindu Tamil, as well as in the minor groups, Moslem Memon, Moslem Urdu and Hindu Telegu, as evidenced by the families with major haemoglobinopathies encountered in the study. The families all had individuals with thalassaemia major, except where stated otherwise. They were distributed as follows:

Moslem Gujerati	13 (2 compound heterozygotes with $ß^S$, 1 with $ß^E$)
Hindu Gujerati	6
Hindu Hindi	5 (1 compound heterozygote with $ß^S$, 1 with $ß^E$)
Hindu Tamil	3
Moslem Urdu	6 (2 compound heterozygote with $ß^S$)
Moslem Memon	2
Hindu Telegu	1

TABLE 3.3 Frequencies of the β thalassaemia and β^S alleles in the South African Indian groups studied

GROUP[1]	n[2]	β THALASSAEMIA			β^S		
		HETEROZYGOTE RATE (%)	ALLELE FREQUENCY	SE[3]	HETEROZYGOTE RATE (%)	ALLELE FRFQUENCY	SE[3]
Moslem Gujerati	305	2.0	0.010	0.004	1.3	0.007	0.003
Hindu Gujerati	175	1.1	0.006	0.004	-	-	-
Hindu Tamil	115	-	-	-	0.9	0.004	0.004

[1] One Hindu Telegu β^S heterozygote was also observed in the random survey. Only two Hindu Telegu individuals were studied and thus * frequency was not calculated.
[2] n = number of individuals studied.
[3] SE = standard error, calculated as $\sqrt{(pq/2n)}$.

Insufficient population data are available on the numbers of individuals in the Indian subgroups in South Africa and it is thus rather difficult to estimate allele frequencies from the homozygote frequencies. In addition, the cases studied do not represent all major haemoglobinopathy families in South Africa, but only those encountered in our research and diagnostic work on the haemoglobinopathies, and thus ascertainment was incomplete.

Qualitative cellulose acetate electrophoresis was carried out on the 638 samples to identify any haemoglobin variants present. HbS, confirmed on citrate agar electrophoresis, was the only variant detected in the survey. Estimates of the allele frequency were only possible in the Moslem Gujerati and Hindu Tamil. These are shown in Table 3.3. In the Hindu Gujerati, no β^S heterozygotes were found among the 177 individuals tested and although the sample was small the frequency can thus be estimated to be less than 0.003. In the Hindu Hindi too few individuals (20) were studied to estimate the frequency.

The β^S allele also occurs in at least some of the other subgroups, namely Hindu Gujerati, Hindu Hindi, Moslem Urdu and Hindu Telegu, as evidenced by families encountered in the course of the study who carry the allele. Similarly, Moslem Gujerati and Hindu Hindi families with a β^E allele and Moslem Gujerati and Moslem Memon families with a $\beta^{D\,Punjab}$ (henceforth referred to as β^D) allele were encountered.

The frequencies of the β thalassaemia and β^S alleles did not deviate from Hardy-Weinberg equilibrium and were not significantly different in the groups in which there were sufficient individuals for comparison.

3.1.3 Causes of microcytosis and hypochromia in South African Indians

Of the 638 individuals surveyed, 171 individuals or 27.5% had microcytosis (MCV < 80fℓ), hypochromia (MCH < 27pg) and/or anaemia (Hb < 13g/dℓ in males and < 12g/dℓ in females). Although there were individuals who had microcytosis and/or hypochromia without anaemia, none of the anaemic individuals had normocytic, normochromic or macrocytic indices.

The causes of microcytosis and/or hypochromia were:

6 ß thalassaemia heterozygotes and $\alpha\alpha/\alpha\alpha$

1 ß thalassaemia heterozygote and $-\alpha/-\alpha$

1 ß thalassaemia heterozygote and $\alpha\alpha\alpha/\alpha\alpha$

54 $-\alpha/\alpha\alpha$

5 $-\alpha/-\alpha$

54 Iron deficiency and $\alpha\alpha/\alpha\alpha$

13 Iron deficiency and $-\alpha/\alpha\alpha$

1 Iron deficiency and $-\alpha/-\alpha$

1 $\alpha\alpha\alpha/\alpha\alpha$

19 Normal α globin genes and haematological indices on repeat sampling

13 Normal iron levels, no repeat sample available for determination of α thalassaemia status

3 Unknown - no serum available for iron studies or repeat sample for DNA studies

The α thalassaemia status of individuals was determined by DNA analysis, as discussed in Section 2.3.4.1.

Thus, of the 171 individuals, eight (4.7%) were ß thalassaemia heterozygotes, 68 (39.8%) were iron deficient and at least 67 (39.2%) were $-\alpha/\alpha\alpha$ and seven (4.1%) $-\alpha/-\alpha$. Fourteen individuals had α thalassaemia and iron deficiency and two individuals were found to be $\alpha\alpha\alpha/\alpha\alpha$. The number of individuals with α thalassaemia is an underestimate as it was only possible to get repeat samples for DNA analysis from 137 of the individuals and thus 35 of the individuals, including 13 of the non iron deficient individuals did not have their α thalassaemia status determined.

The rate of microcytosis and hypochromia in the subgroups were:

Moslem Gujerati	27.9%
Hindu Gujerati	31.0%
Hindu Hindi	23.8%
Hindu Tamil	16.5%

The relative contributions of thalassaemia, both α and ß, and iron deficiency differ in the

groups. In the Moslem Gujerati, Hindu Tamil and Hindu Hindi the ratio of thalassaemia carriers to iron deficient individuals was 1.6, 2.2 and 1.5 respectively, while in the Hindu Gujerati it was 0.7.

3.1.3.1 Iron deficiency

The average rates for iron deficiency in the major groups were estimated and are shown in Table 3.4. All iron deficient individuals were microcytic (MCV < 80fℓ) or hypochromic (MCH < 27pg), and all had low serum iron (< 14μmol/ℓ in males, < 11μmol/ℓ in females) and transferrin saturation (< 20% in males, < 15% in females) and high transferrin (> 4.3g/ℓ). Twenty-seven of them were anaemic (Hb < 13g/dℓ in males, < 12g/dℓ in females). However, these average rates do not really reflect the true profile, as the rates differ dramatically between males and females, as shown in Table 3.4.

3.1.3.2 α thalassaemia

The frequency of the -α haplotype in the subgroups was estimated using the Lenasia data. In those groups where the -α/-α genotype was found, the homozygote frequency could be used to estimate the haplotype frequency, in conjunction with the Hardy-Weinberg equation. This method is relatively imprecise because of the small number of homozygotes and, in addition, ascertainment may have been incomplete. The standard error is thus large, particularly in the smaller sample. The frequency of the -α haplotype could also be determined by gene counting from the frequency of the -α/αα genotype. However, this is likely to be an underestimate, as again ascertainment was incomplete, not only because all microcytic hypochromic individuals could not be tested, but also because those individuals with the -α/αα genotype and normal haematology would not have been detected (Walford and Deacon 1976, Bowden *et al.* 1985). The estimated frequencies using both methods are shown in Table 3.5.

In view of the problems with the estimations of the -α haplotype frequency, an attempt was made to refine the data by estimating the -α frequency in the major groups using the parents of the randomly collected families. Table 3.6 shows the frequency of the -α chromosome in the different religious and language groups together with the expected

TABLE 3.4 Rates of iron deficiency in South African Indians

GROUP	n[1]	IRON DEFICIENT INDIVIDUALS[2]	RATE OF IRON DEFICIENCY (%)
Moslem Gujerati	305	29	9.5
Male	155	4	2.6
Female	150	25	16.7
Hindu Gujerati	175	29	16.6
Male	89	7	7.9
Female	86	22	25.6
Hindu Tamil	115	5	4.3
Male	57	0	0
Female	58	5	8.6
Hindu Hindi	21	2	9.5
Male	14	1	7.1
Female	7	1	14.3

[1] n = number of individuals studied.
[2] Iron deficient individuals were defined as those with low serum iron levels, high transferrin levels and a low percentage transferrin saturation.
Normal serum iron levels: 14-31μmol/ℓ (males), 11-29μmol/ℓ (females).
Normal transferrin levels: 1.9-4.3g/ℓ.
Normal transferrin saturation 20-50% (males), 15-50% females.

TABLE 3.5 Estimates of -α haplotype frequencies using the Lenasia schools survey

GROUP	n[1]	-α/αα	-α/-α	USING GENE COUNTING		USING HARDY-WEINBERG LAW	
				FREQUENCY	SE[2]	FREQUENCY	SE[3]
Moslem Gujerati	305	34	5	0.075	0.011	0.140	0.040
Hindu Gujerati	176	18	0	0.051	0.012	-	-
Hindu Hindi	21	2	1	0.095	0.045	0.218	0.151
Hindu Tamil	115	11	0	0.048	0.014	-	-
Moslem Urdu	13	2	0	0.077	0.052	-	-

[1] n = number of individuals studied.
[2] SE = standard error calculated as $\sqrt{(pq/2n)}$.
[3] SE = standard error calculated as $\sqrt{(1-q^2/2n)}$.

TABLE 3.6 Estimates of the -α frequency using the parents in the random families

GROUP	NUMBER OF CHROMOSOMES STUDIED	-α HAPLOTYPE		HETEROZYGOTE RATE (%)
		FREQUENCY	SE	
Moslem Gujerati	60	0.217	0.053	34
Hindu Gujerati	62	0.145	0.045	25
Hindu Hindi	60	0.050	0.028	10
Hindu Tamil	63	0.032	0.022	6

carrier rates. The frequency was significantly higher in the Moslem Gujerati than in the Hindu Hindi ($p<0.02$) and Hindu Tamil ($p<0.01$), while the Hindu Gujerati had an intermediate frequency.

The red cell indices for the $-\alpha/\alpha\alpha$ individuals from the randomly collected families were:

RCC ($x10^{12}/\ell$)	5.17 ± 0.52
Hb ($g/d\ell$)	13.8 ± 1.49
MCV ($f\ell$)	80.2 ± 4.6
MCH (pg)	26.7 ± 1.4

A low MCH (<27pg) appeared to be the most consistent feature in these individuals, with a low MCV (<80fℓ), the next most constant feature. However, about 35% of these individuals had completely normal haematological indices and would thus have been missed in the Lenasia survey. All the $-\alpha/-\alpha$ individuals had microcytic hypochromic indices. The frequencies for the Moslem Gujerati and Hindu Gujerati in Table 3.5 are clearly underestimates of the true frequency. For the Hindu Hindi and Hindu Tamil, the smaller sample size and lower $-\alpha/\alpha\alpha$ frequencies probably resulted in larger errors in the original estimates.

In the entire study, in which the α globin status of at least 360 random individuals was determined, no -- chromosomes were detected. These would have been suspected if abnormal sized fragments had been seen with the enzymes and probes described in Table 2.3.

There was some evidence for a non-deletion type α thalassaemia, one consanguineous Moslem Gujerati family studied had two children with HbH disease, but an intact α cluster on the chromosomes carrying α thalassaemia. They had normal sized fragments after *Bam*HI, *Bgl*II and *Eco*RI digestion, when probed with both the pDHα and pBRζ1 probes, and also with the additional enzymes and probes described in Table 2.4. This allele, however, probably does not occur at any significant frequency as no individuals with unexplained microcytosis or hypochromia were found after adequate haematological and DNA assessment.

3.2 Discussion

3.2.1 Haematological profiles of the Indian groups in South Africa

In all the groups studied males had higher RCCs, Hbs, MCVs and MCHs as well as higher serum iron and transferrin saturation, with lower transferrin, than the corresponding females. In addition, the rates of iron deficiency were higher in the females than males (Table 3.4). These findings are consistent with the well-defined increased physiological iron requirements of females in the reproductive years. They have a greater daily requirement for iron, mainly because of increased menstrual losses. Pregnancy also results in a strenuous nutritional challenge due to the transfer of iron to the fetoplacental complex. In addition, females have a lower iron intake and are more prone to iron deficiency. Adult males generally have a favourable iron balance and nutritional iron deficiency is thus rarer (reviewed in Bothwell *et al.* 1979). In the Hindu Gujerati and Hindu Hindi groups, however, iron deficiency was found in about 7% of males.

The haematological indices in the different subgroups can be explained by combining the data on the frequencies of thalassaemia, mainly α, and the iron studies. In general, for the subgroups studied in the Lenasia survey, the mean values obtained for the haematological and iron parameters were within the normal range, albeit towards the lower end. As Lenasia is situated in Johannesburg, at 2 000m above sea level, the mean RCC, Hb, MCV and MCH levels could be considered somewhat low. This is reflected in the relatively high rates of microcytosis and hypochromia in all groups. The Hindu Tamil had the highest mean Hb, MCV, and MCH, correlating with the highest serum iron levels and percentage transferrin saturation, the lowest transferrin levels and the lowest rate of both iron deficiency and α thalassaemia. In contrast, the Hindu Gujerati had the lowest Hb, MCV and MCH, correlating with the highest rate of iron deficiency, the lowest serum iron and transferrin saturation values, the highest transferrin levels and a relatively high rate of α thalassaemia. The Moslem Gujerati had intermediate haematological values, iron parameters and rate of iron deficiency. They did, however, have the highest rate of α thalassaemia, which probably accounts for this group having the highest mean RCC. The frequencies of ß thalassaemia are too low to influence the overall haematological profile.

3.2.1.1 Iron deficiency

The carrier state for thalassaemia does not have a major effect on the Hb level and thus most anaemic individuals were iron deficient. Iron deficiency has been reported to occur commonly in South African Indians (Mayet *et al.* 1972, Adams 1973). In a study in Durban in Natal, 33% of females and 20% of males were thought to be iron deficient, and a significant percentage were thought to have latent deficiency, as they had a normal Hb with microcytic indices (Mayet 1976). A more recent study showed 14.4% of Indian women in Natal to have iron deficiency anaemia and a further 26% to have depleted iron stores (Macphail *et al.* 1981). In the Transvaal a study on 100 pregnant Indian women found a high prevalence of iron deficiency. Interestingly, no significant differences were found between the religious subgroups (Lamparelli *et al.* 1988), possibly due to small sample size, as similar trends to those observed in this study were documented. α thalassaemia was not excluded as a cause of microcytosis and hypochromia in any of these studies.

The rates of iron deficiency correlate broadly with the dietary habits of the subgroups. The Hindu Gujerati was the group with the highest rate of iron deficiency. Although many Indo-Aryan Hindus are non-vegetarian, some communities, particularly among the Gujerati-speakers, are largely lacto-vegetarian (Mistry 1965), possibly accounting for the higher rate of iron deficiency. Some Hindi-speaking Hindus are also vegetarian, whereas the Tamil-speakers and the Moslem Gujerati, with lower rates of iron deficiency, are non-vegetarian (Mistry 1965). Although iron intake studies in South African Indians have shown no direct relationship between daily iron intake on a quantitative basis and iron deficiency, poor iron availability, rather than poor dietary content, does appear to be important. Legumes and cereals are important sources of dietary iron, particularly for the vegetarians, but these contain high dietary phytate and low calcium which reduce iron availability for absorption. Those who have diets rich in protein iron have less deficiency (Mayet *et al.* 1972, Adams 1973).

In general, ß thalassaemia heterozygotes are thought to have better iron nutrition than normal individuals, though the diagnoses of iron deficiency and thalassaemia are not mutually exclusive (Economidou *et al.* 1980) and investigations should be done for both in microcytic hypochromic individuals. A study in India of ß thalassaemia carriers showed

better iron nutrition than controls, though iron deficiency still occurred in up to 6% of males and 24% of females (Mehta and Pandya 1987). ß thalassaemia carriers appear to have some advantage in maintaining iron balance, perhaps through the mild ineffective erythropoiesis increasing iron absorption. The difference in iron status may not be as pronounced in populations with good iron nutrition, but if iron nutrition is poor, it seems as if factors that increase iron absorption may contribute (Mehta and Pandya 1987). Conversely, it is thought that long-term iron therapy in ß thalassaemia heterozygotes may cause haemosiderosis (Pearson *et al.* 1974), though more recent studies have shown that the risk of iron overload does not appear to be large (Pippard and Wainscoat 1987). Data on the potential for iron overload in α thalassaemia are scant, though the risks are not considered large.

In view of the high incidence of microcytosis and hypochromia in South African Indians, the data on α and ß thalassaemia may be of clinical significance. In many cases the diagnosis of iron deficiency is made through response to a therapeutic trial of iron. In Indians, as in other tropical and subtropical populations, hypochromic anaemia should not necessarily be attributed to iron deficiency, and detailed studies of iron status should be undertaken prior to iron therapy (Bowden *et al.* 1985). In addition, the risks of iron overload, particularly in α thalassaemia, need to be assessed, prior to embarking on dietary iron supplementation in the South African Indian population, as has been proposed (Lamparelli *et al.* 1987).

3.2.1.2 α and ß thalassaemia in South African Indians

Although iron deficiency has been extensively studied in South African Indians, few studies have addressed the other causes of microcytosis and hypochromia. α thalassaemia has been studied in a number of families (Drysdale and Higgs 1988, Fei *et al.* 1992) but no population surveys have been attempted. For ß thalassaemia, reports of local management of cases provide evidence that the disorder is known to occur in Indians (Poole *et al.* 1989), but no other reports of local families or population surveys have been published.

3.2.1.2.1 α thalassaemia

The frequencies of α thalassaemia in the non-tribal Asian Indians in South Africa appear to be higher in the two groups originating on the west coast (the Moslem Gujerati and Hindu Gujerati), and lowest in the southern Indians (the Hindu Tamil), and they are probably related to past malarial exposure, as they are elsewhere in the world (see discussion in Section 1.9.1). The southern part of the Indian subcontinent has historically had less malaria than the central parts and it was in the Hindu Tamil from the southern parts that the lowest frequencies of α thalassaemia were noted. Studies of α thalassaemia have been mainly confined to forest tribes in India, where high frequencies have correlated with malarial exposure (Brittenham *et al.* 1980, Kar *et al.* 1986, Fodde *et al.* 1988, Kulozik *et al.* 1988a, Fodde *et al.* 1991). In some of the tribal populations, the -α chromosome is so common that most individuals in these populations have only two active α globin genes (Brittenham *et al.* 1979). A study of Dravidian Indians in Singapore reported a -α frequency of 0.04 (Tan *et al.* 1991). This is very close to the frequency in the Hindu Tamil (also Dravidian Indians) in this study. A frequency of 0.012 was reported in a non-tribal Indian population in Andhra Pradesh (Fodde *et al.* 1991).

In view of the extremely low prevalence of HbH disease and absence of Hb Bart's hydrops fetalis in the South African Indians, as in other Indians (Saha and Banerjee 1973), the absence of a -- chromosome in this study was not unexpected. A unique -- chromosome has been described in South Africa. It was originally identified in a Coloured individual (Vandenplas *et al.* 1987). It was subsequently found in an Asian Indian family and it was therefore postulated that the mutation originated in the Gujerat region of India (Drysdale and Higgs 1988). Another Indian family in South Africa, whose religious and linguistic affiliation is not specified, has recently been described with the same mutation (Fei *et al.* 1992).

There was also evidence for the presence of a non-deletion α thalassaemia in the South African Indians; one consanguineous Moslem Gujerati family was ascertained during the course of this study in which two of the children had HbH disease, with no deletions of the α genes detected (see Section 4.2.2).

3.2.1.2.2 ß thalassaemia

In India, ß thalassaemia is thought to represent the most common clinically significant haemoglobin disorder, with carrier frequencies between 1-17% (reviewed in Chatterjea 1966, Saha and Banerjee 1973, Mitra 1983, Sangani *et al.* 1990). The population of nearly 700 million is divided into thousands of highly endogamous groups, many of which are virtual genetic isolates. Consequently results from a survey of haemoglobin abnormalities in one group may not be generalised to others (Brittenham 1981).

In the local population ß thalassaemia carrier rates were highest in the Moslem Gujerati and Hindu Gujerati, corresponding with the highest α thalassaemia rates and thus the areas of origin with higher malarial exposure. The ß thalassaemia alleles attain polymorphic frequencies in the Moslem Gujerati.

Since the Department of Human Genetics, University of the Witwatersrand, Johannesburg, has the only laboratory in South Africa offering prenatal diagnosis for haemoglobinopathies, affected families are drawn from the entire country, though predominantly from the Transvaal and Natal where over 95% of South African Indians reside (Department of National Health and Population Development 1990). Greater numbers of Moslem than Hindu patients have been ascertained, despite the South African Indian population being 70% Hindu and 20% Moslem (Cooper *et al.* 1989/90).

Although the ascertainment of thalassaemia major families at the Department of Human Genetics is probably incomplete, one might expect the distribution of the families, according to religious and language groups, to reflect the general distribution of families in the community, unless particular subgroups have different access to resources or are more resistant to genetic counselling and prenatal diagnosis than others. The Moslem groups, for example, appear to have stronger objections to prenatal diagnosis and termination of pregnancy than the Hindu groups, and thus more Moslem families may have been missed. All groups are likely to have equal access to services as there is a wide referral system through hospital thalassaemia clinics, genetic counselling clinics, the Department of National Health Genetic Services' nurses and medical practitioners in private practice. As far as can be determined, the majority, if not all, patients are treated

in state hospitals because of the high costs of therapy.

The Transvaal Indian population is approximately 55% Moslem and 30% Hindu (Mistry 1965) both predominantly Gujerati-speaking. Using the estimated allele frequencies for ß thalassaemia obtained in the present study, we would expect between three and five Moslem and zero and one Hindu thalassaemia major homozygotes to live in the Transvaal, at present [see Appendix IV(Ai) for calculation]. Using the number of 'at-risk' couples, a maximum of 12 Moslem and two Hindu affected patients would be expected to live in the Transvaal at present [see Appendix IV(Aii) for calculations]. At present there are 22 thalassaemia patients, 18 Moslem and four Hindu, from 14 Moslem and four Hindu families, who live in the Transvaal and are known to the Department of Human Genetics. This is probably an underestimate of the tot .mber as some may have died young and some may be treated by private doctors or at peripheral hospitals. The calculations have assumed that thalassaemia major children live to a median age of 25 years, but some of the patients, particularly those who were born up to 25 years ago, have died at an earlier age. Further, prenatal diagnosis has been offered in South Africa since 1984, and a number of cases were referred overseas prior to this date. A number of affected pregnancies have thus been terminated.

Thus there appears to be a greater number of observed homozygotes than expected, though the numbers are not statistically significant and may reflect sampling error. The increased numbers may result from the high rate of consanguineous and intra-caste marriages in the Moslems, first-cousin marriages being greatly encouraged. In the Hindu Gujerati cousin marriages are prohibited, although intra-caste marriages are common (Mistry 1965). Further, it appears as if the Moslem Urdu and Moslem Memon are over-represented in the thalassaemia major families, as they constitute a very small percentage of the total population. These groups are highly endogamous and also have a high rate of consanguineous marriages. A founder effect may have resulted in a high frequency of ß thalassaemia alleles in these two groups.

As the data are rather less complete for other geographical areas, it is difficult to extrapolate information useful for Natal or the entire country from these Transvaal figures. About 40 thalassaemia major cases in 34 families are known in the Natal region

where 80% of Indians live (Department of National Health and Population Development 1990), though again ascertainment may not have been complete. If one assumes a median survival age of 25 years, an average allele frequency of 0.010 is obtained [see Appendix IV(B)], a figure compatible with those obtained in the Transvaal. However, if the number of homozygotes is similarly raised, this may suggest that the frequency is an overestimate and that the rates for ß thalassaemia are actually lower in the Hindu Hindi, Telegu and Tamil who constitute the majority of Natal Indians (Mistry 1965). Among the Hindu Hindi, cousin marriages are forbidden by custom and religion, but marriages still take place among members of the same caste (Mistry 1965) and thus inbreeding may be important. In the Hindu Tamil and Telegu, cousin marriages are favoured (Mistry 1965) and one might thus expect consanguinity to be important again in raising the proportion of families from the latter two groups.

In general, however, the number of thalassaemia major homozygotes appears to be greater than that expected from the estimated allele frequencies and constitutes a significant disease load. Thirty-five percent of the ascertained thalassaemia cases come from the Transvaal, where only 15% of the total Indian population reside. This may be related, in part, to access to the Department of Human Genetics, and to the publicity and heightened interest which inevitably accompanies a study like the present one.

3.2.1.3 The $ß^S$ allele and other ß globin variants

The sickle cell gene was first described in tribal populations in India by Lehmann and Cutbush (1952). Sickle cell anaemia has since been detected in many groups in India. It is thought to occur predominantly in the tribal populations, which have acted as a reservoir from which the gene has permeated the rest of society (Kar *et al.* 1987), such that it occurs at lower frequency in the caste communities (reviewed in Saha and Banerjee 1973, Brittenham *et al.* 1977, Nagel and Ranney 1990, Nagel and Fleming 1992). Many other haemoglobin variants have been described on the Indian subcontinent, including HbE, HbD, HbJ, HbK, HbM (reviewed in Chatterjea 1966, Saha and Banerjee 1973, Mitra 1983)

Berk and Bull (1943) described the first Asian Indian patient with sickle cell anaemia; he

was a South African. Other heterozygotes and homozygotes have since been described in South Africa (Naude and Neame 1961, Reddy and Ward 1969). The frequency of sickle cell trait was estimated to be 1% in a sample of Natal Indians, not classified by religion or language (Naude and Neame 1961). One Indian family with homozygous HbE has been described (Botha *et al.* 1967).

Sickle cell anaemia in Indians is a uniformly mild disease (Brittenham *et al.* 1977). This phenomenon has been noted both in our South African cases and those reported in the literature. The patients have generally presented as adults, mainly with bone pain (Berk & Bull 1943). One Indian sickle cell anaemia homozygote was investigated at the molecular level, in an attempt to determine a cause for her mild disease (O'Byrne 1983).

Only three sickle cell anaemia homozygotes are known in the Transvaal, all extremely mildly affected. Two are Moslem Gujerati siblings and the third a Hindu Gujerati. We would expect three to five homozygotes, based on the average β^s allele frequency, assuming that such individuals have a near normal lifespan [see Appendix IV(C)]. However, in view of the number of thalassaemia homozygotes observed, we would expect more cases due to consanguinity. It is likely that the cases are not coming to medical attention or are being misdiagnosed because of the mild nature of their disease.

CHAPTER 4 - CHARACTERISATION OF α THALASSAEMIA AND THE α GLOBIN CLUSTER IN SOUTH AFRICAN ASIAN INDIANS

In the chapter that follows, the types and subtypes of -α thalassaemia chromosomes, as well as other α globin gene cluster rearrangements that occur in South African Asian Indians, are characterised. The allele frequencies for the common α globin cluster RFLPs on αα and -α chromosomes are presented. Using family studies, α globin haplotypes have been constructed for the different types of rearranged chromosomes. Further, characteristics of two VNTR systems, which flank the α cluster, have been studied. The data obtained have been used to compare the different Indian groups and to try to define the numbers of origins of the different rearrangements. Data presented for the major groups, namely Moslem Gujerati, Hindu Gujerati, Hindu Hindi and Hindu Tamil, have been analysed statistically. For the minor groups, namely Moslem Urdu, Moslem Memon and Hindu Telegu, the data are presented, but the numbers are too small for statistical analysis.

4.1 Results

4.1.1 Types of -α chromosomes

As discussed in Chapter 3, -α appears to be the only type of α thalassaemia which occurs at significant frequency in South African Indians. The subtypes of -α chromosome found are shown in Table 4.1. The majority of -α chromosomes in all the groups were of the $-\alpha^{3.7I}$ type. Though the $-\alpha^{4.2}$ chromosome occurs in all groups, the proportion of this type was significantly increased in the Hindu Hindi compared to both the Moslem Gujerati ($p<0.001$) and the Hindu Gujerati ($p<0.05$). Only two $-\alpha^{3.7II}$ chromosomes were seen, both in Moslem Gujerati individuals. No $-\alpha^{3.7III}$ chromosomes were found.

4.1.2 Other α globin cluster rearrangements

The $\alpha\alpha\alpha^{anti3.7I}$ rearrangement was found on five out of 60 unrelated chromosomes in the

TABLE 4.1 Subtypes of -α chromosomes in South African Indian groups

GROUP	$-\alpha^{3.7I}$		$-\alpha^{3.7II}$		$-\alpha^{4.2}$		TOTAL
	N[1]	%	N[1]	%	N[1]	%	N[1]
Moslem Gujerati	68	96	2	3	1	1	71
Hindu Gujerati	30	97	0	0	1	3	31
Hindu Hindi	8	67	0	0	4	33	12
Hindu Tamil	13	93	0	0	1	7	14

[1] N = number of -α chromosomes studied.

randomly collected Hindu Hindi families. The frequency in the Hindu Hindi can be estimated to be 0.083 ± 0.036. Although the rearrangement was not observed on the 60 chromosomes studied in the randomly collected families of the other major groups, it was also detected in three Moslem Gujerati individuals from the Lenasia schools survey. They were all investigated because of microcytosis, not necessarily related to the rearrangement. Only a minimum estimate of the frequency in the Moslem Gujerati (0.005) could be calculated. The rearrangement was also found in one non-random Moslem Gujerati family. The $\alpha\alpha\alpha^{anti4.2}$ chromosome was not found in any of the groups studied.

One Moslem Gujerati, one Hindu Hindi and one Moslem Urdu family in the study had individuals with a ζζζ rearrangement. One Moslem Gujerati family had a -ζ rearrangement.

4.1.3 α globin cluster RFLPs

The allele frequencies for the eight common α globin cluster RFLPs studied on αα and -α chromosomes are shown in Table 4.2a for the major groups. The -α subtypes have not been analysed separately because the numbers of chromosomes were too small. The RFLP frequencies for αα chromosomes in the minor groups are shown in Table 4.2b. The allele frequencies were not calculated for the -α chromosomes as very few chromosomes were studied. In some cases, there was insufficient DNA to type the individuals for all the RFLPs, and thus the numbers of individuals studied vary.

Because of the presence of polymorphic repeat sequences in the α globin cluster, alleles classified as one type may actually vary in size. In the IZHVR system (system 5), at least three sizes of *L* (large) allele [4.6kb, 4.8kb and 5.0kb (-)] and two of *M* (medium) allele [4.2kb or 4.4kb (+) and 6.9kb or 7.2kb (-)] 4.could be distinguished. All, except the third *L* allele of 4.6kb correspond to those previously described by Goodbourn *et al.* (1984). Similarly, there were at least three sizes of *PZ* allele (system 6), due to variation in the size of IVS1 of ζ1, the commonest 1.85kb *PZ2* allele, a less common 2.0kb *PZ1* allele and a rarer 1.75kb *PZ3* allele (Goodbourn *et al.* 1983, Fodde *et al.* 1991). In the *Bgl*I/ζ system a new allele of 5.0kb was observed and termed '*H*' in this study.

TABLE 4.2a Allele frequencies of the common α globin cluster polymorphisms for $\alpha\alpha$ and $-\alpha$ chromosomes in the major South African Indian groups

SYSTEM[1]	ALLELE[2]	CHROMOSOME	MOSLEM GUJERATI			HINDU GUJERATI			HINDU HINDI			HINDU TAMIL[3]		
			N[4]	Freq	SE	N[4]	Freq	SE	N[4]	Freq	SE	N[4]	Freq	SE
2	+	αα	103	0.709	0.044	77	0.688	0.052	75	(733	0.051	67	0.701	0.056
		-α	21	0.667	0.103	14	0.643	0.128	8	0.125	0.117	-	-	-
3	+	αα	107	0.682	0.045	76	0.763	0.049	76	0.658	0.054	75	0.640	0.055
		-α	24	0.833	0.076	13	0.923	0.074	8	0.500	0.177	-	-	-
4	+	αα	107	0.252	0.042	76	0.184	0.044	76	0.237	0.049	75	0.333	0.054
		-α	24	0.125	0.068	14	0.071	0.069	8	0.500	0.177	-	-	-
	H	αα	107	0.009	0.009	76	0.026	0.018	76	0.053	0.026	75	0.000	-
		-α	24	0.000	-	14	0.000	-	8	0.000	-	-	-	-
5	S	αα	107	0.037	0.018	76	0.039	0.022	76	0.132	0.039	75	0.027	0.019
		-α	24	0.042	0.040	13	0.000	-	8	0.000	-	-	-	-
	M	αα	107	0.636	0.047	76	0.645	0.055	76	0.539	0.057	75	0.627	0.056
		-α	24	0.542	0.102	13	0.538	0.138	8	0.625	0.171	-	-	-
	L	αα	107	0.327	0.045	76	0.316	0.053	76	0.329	0.054	75	0.347	0.055
		-α	24	0.417	0.100	13	0.462	0.138	8	0.375	0.171	-	-	-

Continued /

TABLE 4.2a Allele frequencies of the common α globin cluster polymorphisms for αα and -α chromosomes in the major South African Indian groups (continued)

SYSTEM[1]	ALLELE[2]	CHROMOSOME	MOSLEM GUJERATI			HINDU GUJERATI			HINDU HINDI			HINDU TAMIL[3]		
			N[4]	Freq	SE	N[4]	Freq	SE	N[4]	Freq	SE	N[4]	Freq	SE
6	Z	αα	107	0.355	0.046	77	0.273	0.051	76	0.237	0.049	73	0.370	0.057
		-α	22	0.182	0.082	13	0.076	0.073	8	0.750	0.152	-	-	-
	PZ1	αα	107	0.112	0.030	77	0.117	0.037	76	0.158	0.042	73	0.205	0.047
		-α	22	0.591	0.105	13	0.692	0.128	8	0.125	0.117	-	-	-
	PZ2	αα	107	0.533	0.048	77	0.597	0.056	76	0.539	0.057	73	0.562	0.058
		-α	22	0.227	0 089	13	0.231	0.117	8	0.125	0.117	-	-	-
	PZ3	αα	107	0.000	-	77	0.013	0.013	76	0.066	0.028	73	0.000	-
		-α	22	0.000	-	13	0.000	-	8	0.000	-	-	-	-
7[5]	+	αα	107	0.271	0.043	77	0.390	0.056	76	0.329	0.054	71	0.254	0.052
		-α	22	0.043	-	13	0.308	0.128	6	0.167	0.152	-	-	-
8[6]	+	αα	108	0.037	0.018	77	0.026	0.018	76	0.066	0.028	74	0.027	0.019
9	+	αα	108	0.037	0.018	77	0.039	0.022	75	0.079	0.031	74	0.027	0.019
		-α	22	0.000	-	14	0.000	-	8	0.000	-	-	-	-

[1] The numbers of the systems correspond to the sites labelled in Figure 2.1 and the RFLPs described in Table 2.4.
[2] In the biallelic systems only the frequency of the '+' allele is shown. In System 4, the frequency of the newly described '*H*' allele is also included. For the other multi-allelic systems frequencies of all alleles are shown.
[3] Allele frequencies were not calculated for the -α chromosomes in the Hindu Tamil as only 3 chromosomes were studied.
[4] N = number of chromosomes studied.
[5] The *Pst*I site detected with the pDHα probe is deleted on $-\alpha^{4.2}$ chromosomes, thus the number of -α chromosomes is reduced in some groups.
[6] The *Pst*I site detected with the pDHα probe is deleted on $-\alpha^{3.7}$ chromosomes, thus insufficient -α chromosomes were available for analysis.

TABLE 4.2b Allele frequencies of the common α globin cluster polymorphisms for $\alpha\alpha$ chromosomes in the minor South African Indian groups

SYSTEM[1]	ALLELE[2]	CHROMOSOME[3]	MOSLEM URDU			MOSLEM MEMON			HINDU TELEGU		
			N[4]	Freq	SE	N[4]	Freq	SE	N[4]	Freq	SE
2	+	αα	16	0.750	0.108	12	0.833	0.108	8	0.875	0.117
3	+	αα	16	0.625	0.121	12	0.833	0.108	8	0.875	0.117
4	+	αα	12	0.000	-	12	0.167	0.108	8	0.000	-
	H	αα	12	0.083	0.080	12	0.000	-	8	0.125	0.117
5	S	αα	16	0.188	0.098	12	0.000	-	8	0.125	0.117
	M	αα	16	0.625	0.121	12	0.917	0.080	8	0.375	0.171
	L	αα	16	0.188	0.098	12	0.083	0.080	8	0.500	0.177
6	Z	αα	16	0.125	0.083	12	0.167	0.031	8	0.000	-
	PZ1	αα	16	0.125	0.083	12	0.000	-	8	0.250	0.153
	PZ2	αα	16	0.688	0.116	12	0.833	0.031	8	0.750	0.153
	PZ3	αα	16	0.063	0.061	12	0.000	-	8	0.000	-
7	+	αα	14	0.286	0.121	12	0.750	0.125	8	0.500	0.177
8	+	αα	16	0.125	0.083	12	0.000	-	8	0.000	-
9	+	αα	16	0.125	0.083	12	0.000	-	8	0.000	-

[1] The numbers of the systems correspond to the sites labelled in Figure 2.1 and the RFLPs described in Table 2.4.
[2] In the biallelic systems only the frequency of the '+' allele is shown. In System 4, the frequency of the newly described '*H*' allele is also included. For the other multi-allelic systems frequencies of all alleles are shown.
[3] Allele frequencies were not calculated for the -α chromosomes in any of the groups as very few chromosomes were studied.
[4] N = number of chromosomes studied.

There were no significant allele frequency differences among the major religious and language groups for the $\alpha\alpha$ chromosomes. Statistical comparisons were not done for the minor groups as the numbers of chromosomes studied were too small. For the *RsaI*/α polymorphism (system 7), the Moslem Memon group appeared to have a high frequency (0.750) of the + allele, though this may be due to a sampling error, as only 12 chromosomes were studied.

For the -α chromosomes, the Hindu Hindi had a significantly lower frequency of the *XbaI* (system 2) + allele than the Moslem Gujerati ($p<0.05$). They also had a higher frequency of the *Z* allele at the PZ/Z locus (system 6) than the Moslem Gujerati ($p<0.05$) and the Hindu Gujerati ($p<0.01$).

The allele frequencies on $\alpha\alpha$ and -α chromosomes were compared in each of the groups. In the Hindu Hindi the -α chromosomes had significantly lower *XbaI*/ζ (system 2) + ($p<0.01$), higher PZ/Z (system 6) *Z* ($p<0.01$) and lower *RsaI*/α (system 7) + ($p<0.01$) allele frequencies compared to the $\alpha\alpha$ chromosomes. In the Moslem Gujerati, the frequency of the *RsaI*/α (system 7) + allele was significantly lower on the -α chromosomes than on the $\alpha\alpha$ chromosomes. In addition, the PZ/Z (system 6) *PZ1* subtype was more common on -α chromosomes, while the *PZ2* subtype occurred more commonly on $\alpha\alpha$ chromosomes in the Moslem Gujerati ($p<0.001$) and Hindu Gujerati ($p<0.001$).

4.1.4 α globin haplotypes

4.1.4.1 αα haplotypes

α globin haplotypes were determined using the RFLP sites 2-9 (Figure 2.1) and family studies to assign the linkage phase. A total of 375 $\alpha\alpha$ chromosomes were typed from the different religious and language groups and it was possible to determine 322 haplotypes unequivocally. The $\alpha\alpha$ haplotypes are shown in Tables 4.3a and b for the major and minor groups respectively. The haplotypes are divided into groups according to their four shared 3′ sites. The groups are numbered with Roman numerals. Within each group the four 5′ sites define a subtype, which is named with a small letter following the Roman numeral.

TABLE 4.3a Frequencies of α globin haplotypes on αα chromosomes in the major South African Indian groups

HAPLOTYPE									MOSLEM GUJERATI		HINDU GUJERATI		HINDU HINDI		HINDU TAMIL	
Type[1]	Description[2]								(N=91)[3]		(N=74)[3]		(N=68)[3]		(N=61)[3]	
									Freq	SE	Freq	SE	Freq	SE	Freq	SE
Ia	+	+	–	M	P	+	–	–	0.264	0.046	0.351	0.055	0.309	0.056	0.262	0.056
Ib	–	+	–	M	P	+	–	–	0.000	-	0.000	-	0.029	0.020	0.016	0.016
Id	–	+	–	L	P	+	–	–	0.000	-	0.014	0.014	0.000	-	0.000	-
Ic*	+	+	–	L	P	+	–	–	0.011	0.011	0.041	0.023	0.015	0.015	0.000	-
IIa	–	+	–	L	P	–	–	–	0.132	0.035	0.122	0.038	0.103	0.037	0.148	0.045
IIb	+	+	–	L	P	–	–	–	0.143	0.037	0.108	0.036	0.118	0.039	0.197	0.051
IIc	+	+	–	M	P	–	–	–	0.022	0.015	0.027	0.019	0.000	-	0.016	0.016
IIe	+	–	–	S	P	–	–	–	0.011	0.011	0.000	-	0.015	0.015	0.000	-
IIf	+	–	+	S	P	–	–	–	0.000	-	0.000	-	0.015	0.015	0.000	-
IIh*	+	–	+	M	P	–	–	–	0.011	0.011	0.000	-	0.000	-	0.016	0.016
IIi*	–	–	+	M	P	–	–	–	0.011	0.011	0.000	-	0.000	-	0.000	-
IIj*	+	–	H	S	P	–	–	–	0.000	-	0.014	0.014	0.059	0.029	0.000	-
IIk*	–	–	H	M	P	–	–	–	0.000	-	0.014	0.014	0.000	-	0.000	-
IIIa	–	–	+	M	Z	–	–	–	0.055	0.024	0.149	0.041	0.118	0.039	0.098	0.038
IIIb	+	–	+	M	Z	–	–	–	0.187	0.041	0.027	0.019	0.118	0.039	0.197	0.051
IIIc	–	+	–	L	Z	–	–	–	0.033	0.019	0.000	-	0.000	-	0.000	-
IIIe	–	+	–	M	Z	–	–	–	0.011	0.011	0.014	0.014	0.000	-	0.016	0.016
IIIf	–	–	–	M	Z	–	–	–	0.011	0.011	0.000	-	0.000	-	0.000	-
IIIg*	+	+	–	M	Z	–	–	–	0.055	0.024	0.068	0.029	0.015	0.015	0.000	-
IIIi*	+	+	–	L	Z	–	–	–	0.000	-	0.014	0.014	0.000	-	0.000	-
IIIj*	–	+	+	M	Z	–	–	–	0.000	-	0.000	-	0.015	0.015	0.000	-
IVa	+	–	–	S	P	–	+	+	0.000	-	0.027	0.019	0.044	0.025	0.033	0.023
IVc	–	+	–	L	P	–	+	+	0.000	-	0.000	-	0.015	0.015	0.000	-
IVd*	+	+	–	L	P	–	+	+	0.011	0.011	0.000	-	0.000	-	0.000	-
Vc	+	–	–	S	P	–	–	+	0.011	0.011	0.000	-	0.000	-	0.000	-
Ve*	+	+	–	L	P	–	–	+	0.000	-	0.014	0.014	0.015	0.015	0.000	-
VIIIa	+	–	–	S	P	–	+	–	0.011	0.011	0.000	-	0.000	-	0.000	-

[1] * Denotes a haplotype not described previously.

[2] Sites 2-9 in Figure 2.1 were used to construct haplotypes. As the *Acc*I site (not shown) was excluded types II and VI haplotypes and types IIIc and VIIa could not be distinguished (Higgs *et al.* 1986).

[3] N = number of chromosomes on which haplotypes could be determined. In each religious/language group a number of haplotypes were indeterminate and were excluded from the analysis.

TABLE 4.3b Frequencies of α globin haplotypes on αα chromosomes in the minor South African Indian groups

HAPLOTYPE										MOSLEM MEMON		MOSLEM URDU		HINDU TELEGU	
Type[1]	Description[2]									(N=12)[3]		(N=8)[3]		(N=8)[3]	
										Freq	SE	Freq	SE	Freq	SE
Ia	+	+	–	M	P	+	–	–		0.750	0.125	0.500	0.177	0.375	0.171
Ib	–	+	–	M	P	+	–	–		-	-	-	-	-	-
Id	–	+	–	L	P	+	–	–		-	-	-	-	-	-
Ic*	+	+	–	L	P	+	–	–		-	-	-	-	0.125	0.117
IIa	–	+	–	L	P	–	–	–		0.083	0.080	0.125	0.117	0.125	0.117
IIb	+	+	–	L	P	–	–	–		-	-	0.125	0.117	0.250	0.153
IIc	+	+	–	M	P	–	–	–		-	-	-	-	-	-
IIe	+	–	–	S	P	–	–	–		-	-	-	-	-	-
IIf	+	–	+	S	P	–	–	–		-	-	-	-	-	-
IIh*	+	–	+	M	P	–	–	–		-	-	-	-	-	-
IIi*	–	–	+	M	P	–	–	–		-	-	-	-	-	-
IIj*	+	–	H	S	P	–	–	–		-	-	0.125	0.117	0.125	0.117
IIk*	–	–	H	M	P	–	–	–		-	-	-	-	-	-
IIIa	–	–	+	M	Z	–	–	–		0.083	0.080	-	-	-	-
IIIb	+	–	+	M	Z	–	–	–		0.083	0.080	-	-	-	-
IIIc	–	+	–	L	Z	–	–	–		-	-	-	-	-	-
IIIe	–	+	–	M	Z	–	–	–		-	-	-	-	-	-
IIIf	–	–	–	M	Z	–	–	–		-	-	0.125	0.117	-	-
IIIg*	+	+	–	M	Z	–	–	–		-	-	-	-	-	-
IIIi*	+	+	–	L	Z	–	–	–		-	-	-	-	-	-
IIIj*	–	+	+	M	Z	–	–	–		-	-	-	-	-	-
IVa	+	–	–	S	P	–	+	+		-	-	-	-	-	-
IVc	–	+	–	L	P	–	+	+		-	-	-	-	-	-
IVd*	+	+	–	L	P	–	+	+		-	-	-	-	-	-
Vc	+	–	–	S	P	–	–	+		-	-	-	-	-	-
Ve*	+	+	–	L	P	–	–	+		-	-	-	-	-	-
VIIa	+	–	–	S	P	–	+	–		-	-	-	-	-	-

[1] * Denotes a haplotype not described previously.

[2] Sites 2-9 in Figure 2.1 were used to construct haplotypes. As the *Acc*I site (not shown) was excluded types II and VI haplotypes and types IIIe and VIIa could not be distinguished (Higgs *et al.* 1986).

[3] N = number of chromosomes on which haplotypes could be determined. In each religious/language group a number of haplotypes were indeterminate and were excluded from the analysis.

Twenty-seven different haplotypes were observed on the $\alpha\alpha$ chromosomes, 17 of which had been previously described (Higgs *et al.* 1986). Haplotypes IId, Va and Vb previously observed in Asian Indians, at low frequency (Higgs *et al.* 1986), were not seen in the present study. Haplotype Ia was the commonest in all the groups studied.

The distribution of haplotype groups was compared among the major religious and language groups and no statistical differences were found. However, within the group III haplotypes, differences in the distribution of subtypes occur. The Moslem Gujerati had a ratio of type IIIa:IIIb of 1:3.4, the Hindu Tamil of 1:1.3, while for the Hindu Gujerati the ratio was 5.5:1. The Hindu Hindi had an intermediate ratio of 1:1. The differences between the Hindu Gujerati and both the Moslem Gujerati and Hindu Tamil were statistically significant ($p < 0.01$).

The distribution of *PZ* alleles among the haplotypes was interesting. Groups I, IV, V and VIII haplotypes only had *PZ2* alleles, whereas in group II haplotypes, except subtypes IIj and IIk, about 50% had *PZ1* and 50% *PZ2* alleles. *PZ3* only occurs in association with haplotypes IIj and IIk, the majority of chromosomes bearing this allele. However, *PZ2* alleles also occur with these haplotypes. Haplotypes IIj and IIk were also associated with the new *BglI H* allele. In addition, on the chromosome with the IIj haplotype, there was an insertion into $\zeta 2$ of $\pm$0.4kb (Dr DR Higgs, personal communication 1992), detectable as a band of 5.4kb and 6.3kb on *Eco*RI/pBRζ1 and *Bam*HI/pBRζ1 blots respectively.

Genetic diversities (Nei 1987) calculated for the haplotype distributions on $\alpha\alpha$ chromosomes in the major groups were as follows:

Moslem Gujerati	0.858
Hindu Gujerati	0.829
Hindu Hindi	0.857
Hindu Tamil	0.837

4.1.4.2 α thalassaemia haplotypes

The haplotypes associated with the -α chromosomes were also determined. The numbers

of chromosomes studied were relatively small and thus statistical analysis was not possible. Table 4.4 shows the haplotypes associated with the $-\alpha^{3.7\text{II}}$ chromosomes, together with the associated PZ alleles. The haplotypes were classified according to their most likely 5′ origin, but other possibilities for classification exist, as shown below the table. No $-\alpha^{3.7\text{II}}$ haplotypes were determined in the Moslem Memon or Hindu Telegu groups. On all $-\alpha^{3.7}$ chromosomes, the *Pst*I/pDHα site is deleted.

The $-\alpha^{3.7\text{II}}$ rearrangement occurs on 10 different haplotypes, 12 if *PZ* subtyping is included. A new haplotype, Xa, has been assigned which has a IIIa 5′ end but a 3′ end derived from a group IV or V chromosome. The haplotypes did not appear to be confined to any religious or language group, though there appeared to be some clustering, especially in the Hindu Hindi where most $-\alpha^{3.7\text{II}}$ chromosomes appeared to be associated with a IIIa haplotype. In the Moslem Gujerati most of the $-\alpha^{3.7\text{II}}$ chromosomes were associated with a IIb or IIc haplotype and in the Hindu Gujerati with a Ia or IIa haplotype.

Only two $-\alpha^{3.7\text{III}}$ chromosomes were studied and these both had the same haplotype, namely type IIc derivative:

(+ + - M P - * -), with a *PZ2* allele.

The $-\alpha^{4.2}$ chromosome occurred with two haplotypes:

(+ - + M Z * - -) in one Hindu Gujerati family and

(- + - L Z * - -) in two Hindu Hindi and one Moslem Gujerati family.

They appear to be derived from a IIIb and IIIc chromosome at their 5′ ends respectively, but their 3′ ends could originate from a type I, II, III or VI haplotype.

In one consanguineous Moslem Gujerati family, two children with HbH disease were shown to be homozygous for the haplotype IIIb.

4.1.4.3 Haplotypes associated with other α globin cluster rearrangements

4.1.4.3.1 $\alpha\alpha\alpha$ haplotypes

Three haplotypes were associated with the $\alpha\alpha\alpha$ chromosomes studied:

TABLE 4.4 α globin haplotypes on $-\alpha^{3.7}$[1] chromosomes in the South African Indian groups

HAPLOTYPE[1]	PZ TYPE	MOSLEM GUJERATI		HINDU GUJERATI		HINDU HINDI		HINDU TAMIL		MOSLEM URDU	
		N[2]	%[3]	N[2]	%[3]	N[2]	%	N[2]	%	N[2]	%
Ia	2	1	7	3	27	1	17	1	33	-	-
IIa[4]	1	2	14	4	36	1	17	-	-	-	-
	2	1	7	-	-	-	-	-	-	-	-
IIb[5]	1	3	21	2	18	-	-	1	33	-	-
IIc	1	3	21	2	18	-	-	-	-	1	33
	2	1	7	-	-	-	-	-	-	-	-
IIIa	-	-	-	-	-	4	67	-	-	1	33
IIIb	-	1	7	-	-	-	-	-	-	-	-
IIIe	-	1	7	-	-	-	-	-	-	-	-
IIIg	-	-	-	-	-	-	-	-	-	1	33
IIIh[6]	-	1	7	-	-	-	-	-	-	-	-
Xa[7]	-	-	-	-	-	-	-	1	33	-	-
Indeterminate		8	-	1	-	0	-	0	-	0	
Total		22		12		6		3		3	

[1] In the assignment of $-\alpha^{3.7}$ haplotypes, the *Rsa*I polymorphism is useful in defining the 5′ origin of the haplotypes, but the 3′ end is difficult to assign as the *Pst*I/α site is deleted.

[2] N = number of chromosomes observed.

[3] The percentages of each haplotype were calculated from the number of chromosomes with assigned haplotypes only (chromosomes with indeterminate haplotypes excluded) so that the groups could be compared.

[4] The 5′ end may also have been derived from a IVa, Va, VIa or VIIIb chromosome.

[5] The 5′ end may also have been derived from a IVd, Ve or VIb chromosome.

[6] The IIIh haplotype (+ − − S Z − * −) has a unique 5′ end not seen on the $\alpha\alpha$ chromosomes studied.

[7] The Xa haplotype (− − + M Z − * +) appears to be the result of a crossover between a IIIa haplotype (5′ end) and a type IV or V haplotype (3′ end), and thus a new haplotype results.

(- - + M Z - - - -) in three Hindu Hindi and one Moslem Gujerati family

(- + - L P - - - -) in one Hindu Hindi family

(- + - M P - - - -) in one Hindu Hindi family.

Their 5′ origins appeared to be from a Ib, IIa and IIIa chromosome respectively, while the 3′ origins were difficult to determine. Two *RsaI*/α fragments were seen because of the additional α gene.

4.1.4.3.2 $\zeta\zeta\zeta$ haplotypes

The $\zeta\zeta\zeta$ chromosome was associated with two haplotypes:

(+ +- -- MS PZPZ - - -) in a Moslem Urdu family and

(+ -- ++ MM PZPZ - - -) in a Hindu Hindi family.

The Moslem Urdu $\zeta\zeta\zeta$ is typical of the south-east Asian type and is the result of an interchromosomal crossover between a Ia chromosome on the 5′ end and a IId/e at the 3′ end (Hill *et al.* 1985a, Hertzberg *et al.* 1988, O'Shaughnessy *et al.* 1990). The Hindu Hindi chromosome appears to be the result of a inter- or intrachromosomal recombination event between two IIf chromosomes.

4.1.4.3.3 -ζ haplotypes

One -ζ chromosome in a Moslem Gujerati family was associated with the haplotype:

(± del del del PZ - + +)

The 5′ region of the haplotype was indeterminate, while the 3′ part derives from a type IV haplotype.

4.1.5 VNTR analysis

The distribution of the alleles at the 5′ and 3′ hypervariable loci, 5′HVR and 3′HVR, were characterised in the different groups. Fragments of similar sizes were binned for frequency analysis, in bin sizes of 0.1xfragment size, as suggested by Budowle *et al.* (1991). The distributions on $\alpha\alpha$ chromosomes for the major and minor groups are shown in Tables 4.5a and b for 5′HVR and Tables 4.6a and b for 3′HVR respectively. The sample sizes for the minor groups were too small for any statistical analysis.

TABLE 4.5a 'Allele bin' frequencies for 5'HVR on $\alpha\alpha$ chromosomes in the major South African Indian groups

ALLELE BINS (kb)	MOSLEM GUJERATI (N=108)[1]		HINDU GUJERATI (N=77)[1]		HINDU HINDI (N=76)[1]		HINDU TAMIL (N=68)[1]	
	Freq	SE	Freq	SE	Freq	SE	Freq	SE
<1.9	0.000	-	0.000	-	0.000	-	0.015	0.015
1.91-2.10	0.130	0.032	0.039	0.022	0.145	0.040	0.191	0.048
2.11-2.30	0.167	0.036	0.273	0.051	0.132	0.039	0.221	0.050
2.31-2.50	0.213	0.040	0.247	0.049	0.211	0.047	0.221	0.050
2.51-2.70	0.148	0.034	0.065	0.028	0.211	0.047	0.074	0.032
2.71-2.90	0.083	0.027	0.052	0.025	0.105	0.035	0.074	0.032
2.91-3.20	0.028	0.016	0.039	0.022	0.000	-	0.015	0.015
3.21-3.50	0.037	0.018	0.039	0.022	0.000	-	0.015	0.015
3.51-3.80	0.009	0.009	0.026	0.018	0.000	-	0.000	-
3.81-4.10	0.009	0.009	0.000	-	0.000	-	0.044	0.025
4.11-4.50	0.000	-	0.026	0.018	0.000	-	0.000	-
4.51-4.90	0.009	0.009	0.000	-	0.000	-	0.015	0.015
4.91-5.40	0.000	-	0.039	0.022	0.000	-	0.000	-
5.41-5.90	0.009	0.009	0.000	-	0.000	-	0.000	-
5.91-6.50	0.000	-	0.000	-	0.000	-	0.000	-
6.51-7.10	0.000	-	0.000	-	0.039	0.022	0.059	0.029
7.11-7.80	0.019	0.013	0.039	0.022	0.053	0.026	0.044	0.025
7.81-8.60	0.083	0.027	0.091	0.033	0.079	0.031	0.000	-
8.61-9.50	0.028	0.016	0.000	-	0.013	0.013	0.015	0.015
9.51-10.50	0.028	0.016	0.026	0.018	0.013	0.013	0.000	-

[1] N = number of chromosomes studied.

TABLE 4.5b 'Allele bin' frequencies for 5'HVR on $\alpha\alpha$ chromosomes in the minor South African Indian groups

ALLELE BINS (kb)	MOSLEM URDU (N=13)[1]		MOSLEM MEMON (N=12)[1]		HINDU TELEGU (N=8)[1]	
	Freq	SE	Freq	SE	Freq	SE
<1.9	0.077	0.074	0.000	-	0.000	-
1.91-2.10	0.077	0.074	0.000	-	0.250	0.153
2.11-2.30	0.385	0.135	0.333	0.136	0.250	0.153
2.31-2.50	0.077	0.074	0.500	0.144	0.000	-
2.51-2.70	0.077	0.074	0.083	0.080	0.250	0.153
2.71-2.90	0.154	0.100	0.000	-	0.000	-
2.91-3.20	0.000	-	0.000	-	0.000	-
3.21-3.50	0.000	-	0.000	-	0.000	-
3.51-3.80	0.000	-	0.000	-	0.125	0.117
3.81-4.10	0.000	-	0.000	-	0.000	-
4.11-4.50	0.000	-	0.083	0.080	0.000	-
4.51-4.90	0.000	-	0.000	-	0.000	-
4.91-5.40	0.000	-	0.000	-	0.000	-
5.41-5.90	0.000	-	0.000	-	0.000	-
5.91-6.50	0.000	-	0.000	-	0.000	-
6.51-7.10	0.077	0.074	0.000	-	0.000	-
7.11-7.80	0.000	-	0.000	-	0.000	-
7.81-8.60	0.000	-	0.000	-	0.000	-
8.61-9.50	0.077	0.074	0.000	-	0.125	0.117
9.51-10.50	0.000	-	0.000	-	0.000	-

[1] N = number of chromosomes studied.

TABLE 4.6a 'Allele bin' frequencies for 3'HVR on αα chromosomes in the major South African Indian groups

ALLELE (kb)	MOSLEM GUJERATI (N=108)[1]		HINDU GUJERATI (N=77)[1]		HINDU HINDI (N=75)[1]		HINDU TAMIL (N=71)[1]	
	Freq	SE	Freq	SE	Freq	SE	Freq	SE
<1.9	0.000	-	0.039	0.022	0.013	0.013	0.014	0.014
1.91-2.10	0.009	0.009	0.104	0.035	0.054	0.026	0.042	0.024
2.11-2.30	0.102	0.029	0.208	0.046	0.107	0.036	0.099	0.035
2.31-2.50	0.259	0.042	0.143	0.040	0.160	0.002	0.197	0.047
2.51-2.70	0.130	0.032	0.104	0.035	0.133	0.039	0.183	0.046
2.71-2.90	0.074	0.025	0.052	0.025	0.040	0.022	0.028	0.020
2.91-3.20	0.037	0.018	0.039	0.022	0.133	0.039	0.000	-
3.21-3.50	0.019	0.013	0.039	0.022	0.040	0.022	0.056	0.027
3.51-3.80	0.019	0.013	0.026	0.018	0.040	0.022	0.014	0.014
3.81-4.10	0.056	0.022	0.091	0.033	0.027	0.019	0.028	0.020
4.11-4.50	0.056	0.022	0.026	0.018	0.040	0.022	0.141	0.041
4.51-4.90	0.083	0.027	0.039	0.022	0.027	0.019	0.042	0.024
4.91-5.40	0.074	0.025	0.026	0.018	0.054	0.026	0.085	0.033
5.41-5.90	0.019	0.013	0.052	0.025	0.040	0.022	0.014	0.014
5.91-6.50	0.037	0.018	0.013	0.013	0.027	0.019	0.028	0.020
6.51-7.10	0.019	0.013	0.000	-	0.040	0.022	0.014	0.014
7.11-7.80	0.009	0.009	0.000	-	0.013	0.013	0.014	0.014
7.81-8.60	0.000	-	0.000	-	0.013	0.013	0.000	-

[1] N = number of chromosomes studied.

TABLE 4.6b 'Allele bin' frequencies for 3'HVR on αα chromosomes in the minor South African Indian groups

ALLELE PINS (kb)	MOSLEM URDU (N=17)[1]		MOSLEM MEMON (N=12)[1]		HINDU TELEGU (N=8)[1]	
	Freq	SE	Freq	SE	Freq	SE
<1.9	0.000	-	0.000	-	0.000	-
1.91-2.10	0.000	-	0.167	0.108	0.125	0.117
2.11-2.30	0.059	0.057	0.167	0.108	0.375	0.171
2.31-2.50	0.118	0.078	0.333	0.136	0.000	-
2.51-2.70	0.176	0.092	0.083	0.080	0.000	-
2.71-2.90	0.118	0.078	0.083	0.080	0.000	-
2.91-3.20	0.118	0.078	0.000	-	0.000	-
3.21-3.50	0.000	-	0.000	-	0.000	-
3.51-3.80	0.059	0.057	0.000	-	0.000	-
3.81-4.10	0.000	-	0.000	-	0.000	-
4.11-4.50	0.059	0.057	0.083	0.080	0.000	-
4.51-4.90	0.235	0.103	0.000	-	0.000	-
4.91-5.40	0.000	-	0.083	0.080	0.125	0.117
5.41-5.90	0.000	-	0.000	-	0.000	-
5.91-6.50	0.059	0.057	0.000	-	0.125	0.117
6.51-7.10	0.000	-	0.000	-	0.000	-
7.11-7.80	0.000	-	0.000	-	0.000	-
7.81-8.60	0.000	-	0.000	-	0.125	0.117
8.61-9.50	0.000	-	0.000	-	0.125	0.117

[1] N = number of chromosomes studied.

The 5'HVR had a bimodal distribution, apparent in all the major groups studied (Table 4.6a), with no fragments in the 5.91-6.50kb range. This distribution is due to a polymorphic *Rsa*I site, which is either present or absent. 53 alleles were observed prior to binning and these were pooled into 20 bins. The heterozygosity rate at this locus was observed to be 78.5% prior to binning.

The frequencies of the 'allele bins' were not significantly different among the groups, with the majority of fragments in all groups being associated with the presence of the *Rsa*I site and thus fragment-shifted. The most common 'alleles' in the Moslem Gujerati, Hindu Gujerati and Hindu Tamil appeared to be the two 'bins' in the 2.1-2.5kb range. In the Hindu Hindi the most common 'alleles' were between 2.31kb and 2.70kb, a slight shift to the right from the other major groups. In the Hindu Hindi the distribution was very narrow with no fragments sized between 2.91kb and 6.50kb. In the Hindu Gujerati the two common 'bins' account for greater than 50% of 'alleles', whereas in the Moslem Gujerati there appeared to be a slightly greater spread with the two commonest bins only accounting for 38%. The Hindu Tamil were intermediate with the two common alleles accounting for 44%.

The 3'HVR system also has a bimodal distribution with a less well-defined antimode in the 3.51-3.80kb range (Table 4.6a). The majority of fragments were again in the smaller size range, though the distribution in this case is not due to a superimposed single site RFLP. Prior to binning there were 56 alleles, with 88.7% of individuals being heterozygotes.

No significant differences in 'allele bin' frequencies were observed between the major groups. In all groups except the Hindu Gujerati, the most common 'bin' was in the 2.31-2.50kb range. In the Hindu Gujerati fragments of 2.11-2.30kb were more common, accounting for about 20%, with a second mode appearing at 3.81-4.10kb. In the Moslem Gujerati 25% of alleles were in the 2.31-2.50kb range with a long right tail. The second mode, though present, was not distinct in the Moslem Gujerati and Hindu Hindi. In the Hindu Tamil most fragments occurred between 2.31 and 2.70kb, a larger range and the second mode occurred at 4.11-4.50kb.

The distributions of the 5′HVR and 3′HVR alleles on the $\alpha\alpha$ and $-\alpha$ chromosomes were also compared. The samples from the different religious and language groups were pooled in order to facilitate statistical analysis. The data, shown in Table 4.7, were not significantly different between the two groups for either 5′HVR or 3′HVR.

The major religious and language groups were assessed for heterozygote deficiency at the 5′HVR and 3′HVR loci, using the total number of individuals studied in each group, irrespective of the haplotype or the presence of α globin cluster rearrangements. These data together with the genetic diversities are shown in Tables 4.8 and 4.9 respectively. The minor groups were not assessed as the samples were too small. Heterozygote deficiency was significant in the total sample for both systems ($p<0.001$) and in the Moslem Gujerati for 5′HVR ($p<0.05$). However, if true homozygotes were considered, who had identical 3′HVR and 5′HVR fragments as well as an identical α globin haplotype, no heterozygote deficiency was observed in any group for either system. Genetic diversity was >0.950 in all groups studied for both the 5′HVR and 3′HVR systems.

The distributions of the 5′HVR and 3′HVR alleles were compared among the different haplotype groups. Figures 4.1 and 4.2 show these distributions. The sample was not divided according to the religious and linguistic groups, as the numbers would have been too small for meaningful analysis. For the 5′HVR system group I haplotypes had a significantly different 5′HVR distribution from groups II ($p<0.01$) and III ($p<0.05$). Group I haplotypes were predominantly associated with small alleles and the presence of the *Rsa*I site. Sixty percent of the alleles were between 1.90kb and 2.50kb in size, while <5% were associated with absence of the *Rsa*I site. In groups II and III at least 20% of the alleles were associated with absence of the *Rsa*I site and the most common fragment sizes were between 2.11 and 2.70kb, with a peak between 2.31kb and 2.50kb. Similarly, using the 3′HVR system, the distributions were significantly different between groups I and II ($p<0.0001$), groups II and III ($p<0.001$) and groups I and III ($p<0.05$). Group I alleles were again mainly small, clustering between 2.11kb and 2.50kb with a small tail. Group II fragments were predominantly large, between 3.81kb and 5.40kb, whereas group III appeared to have an intermediate pattern with two peaks, one at 2.31-2.50kb and a second at 4.91-5.40kb.

TABLE 4.7 Distribution of 'allele bins' on αα and -α chromosomes for 5'HVR and 3'HVR in the total South African Indian sample

ALLELE BINS (kb)	5'HVR				3'HVR			
	αα		-α		αα		-α	
	N[1]	%	N[1]	%	N[1]	%	N[1]	%
<1.9	2	0.6	1	2.0	5	1.4	2	3.8
1.91-2.10	44	12.2	7	13.7	19	5.2	2	3.8
2.11-2.30	75	20.7	11	21.6	48	13.0	6	11.3
2.31-2.50	80	22.1	14	27.5	71	19.3	4	7.5
2.51-2.70	46	12.7	7	[illegible]	49	13.3	5	9.4
2.71-2.90	28	7.7	3	[illegible]	20	5.4	4	7.5
2.91-3.20	7	1.9	3	5.9	19	5.2	3	5.7
3.21-3.50	8	2.2	1	2.0	12	3.3	8	15.1
3.51-3.80	4	1.1	0	-	9	2.4	4	7.5
3.81-4.10	4	1.1	0	-	17	4.6	6	11.3
4.11-4.50	3	0.8	0	-	23	6.3	3	5.7
4.51-4.90	2	0.6	0	-	21	5.7	2	3.8
4.91-5.40	3	0.8	0	-	22	6.0	1	1.9
5.41-5.90	1	0.3	0	-	10	2.7	1	1.9
5.91-6.50	0	-	0	-	11	3.0	2	3.8
6.51-7.10	8	2.2	1	2.0	6	1.6	0	-
7.11-7.80	12	3.3	1	2.0	3	0.8	0	-
7.81-8.60	22	6.1	1	2.0	2	0.5	0	-
8.61-9.50	7	1.9	1	2.0	1	0.3	0	-
9.51-10.50	6	1.7	0	-	0	-	0	-
TOTAL	362		51		368		53	

[1] N = number of chromosomes studied.

TABLE 4.8 Numbers of heterozygotes observed and expected and genetic diversities for the 5'HVR system in the South African Indian groups

GROUP	INDIVIDUALS STUDIED	HETEROZYGOTES		HOMOZYGOTES			(χ^2-VALUES)[2]	GENETIC DIVERSITY
		Observed	Expected	Observed	Expected	True[1]		
Moslem Gujerati	65	50	60	15	5	1	5.909*	0.938
Hindu Gujerati	43	35	39	8	4	2	1.550	0.925
Hindu Hindi	45	39	42	6	3	2	1.111	0.933
Hindu Tamil	40	35	37	5	3	0	0.556	0.456
Total[3]	209	164	195	45	14	6	18.969**	0.933

[1] True homozygotes are those who not only have the same 5'HVR on both their chromosomes, but also the same α globin cluster haplotypes and rearrangements.

[2] Chi-squared values refer to the comparisons of observed and expected values, only using 5'HVR, not true homozygotes.
* $p<0.05$,** $p<0.001$.

[3] 'Total' refers to all individuals studied, including those from the minor religious/language groups.

TABLE 4.9 Numbers of heterozygotes observed and expected and genetic diversities for the 3'HVR system in the South African Indian groups

GROUP	INDIVIDUALS STUDIED	HETEROZYGOTES		HOMOZYGOTES			(χ^2-VALUES)[2]	GENETIC DIVERSITY
		Observed	Expected	Observed	Expected	True[1]		
Moslem Gujerati	66	58	63	8	3	1	2.479	0.957
Hindu Gujerati	44	41	42	3	2	2	0.212	0.955
Hindu Hindi	45	41	43	4	2	2	0.714	0.966
Hindu Tamil	40	34	38	6	2	0	2.222	0.936
Total[3]	213	189	204	24	9	6	7.391**	0.959

[1] True homozygotes are those who not only have the same 3'HVR on both their chromosomes, but also the same α globin cluster haplotypes and rearrangements.

[2] Chi-squared values refer to the comparisons of observed and expected values, only using 3'HVR, not true homozygotes.
** $p<0.01$.

[3] 'Total' refers to all individuals studied, including those from the minor religious/language groups.

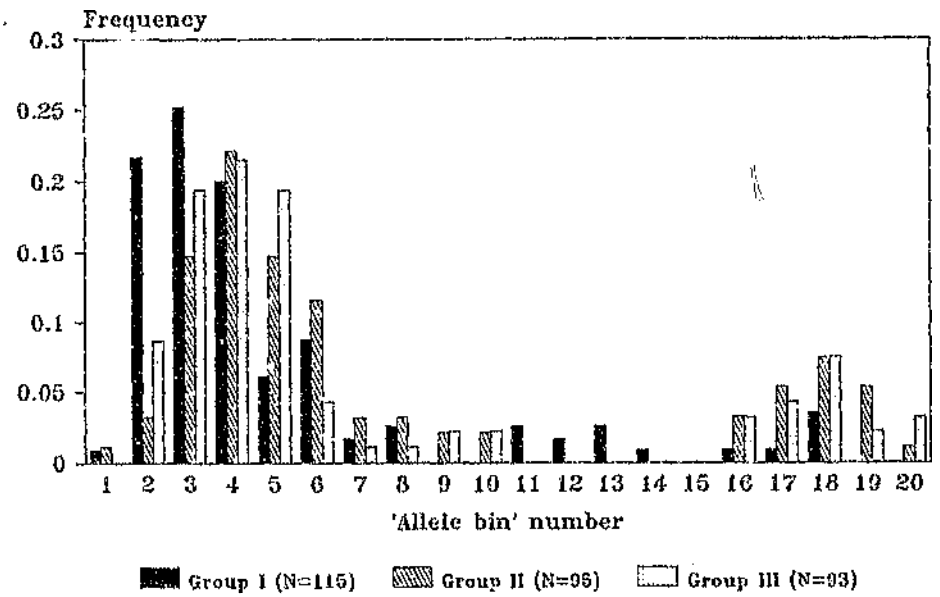


FIGURE 4.1 5'HVR distributions for major haplotype groups

Haplotypes are grouped as shown in Table 4.3.
'Allele bin' numbers correspond to 'allele bins' as follows:
1: ≤1.9kb, 2: 1.91-2.10kb, 3: 2.11-2.30kb, 4: 2.31-2.50kb, 5: 2.51-2.70kb, 6: 2.71-2.90kb, 7: 2.91-3.20kb, 8: 3.21-3.50kb, 9: 3.51-3.80kb, 10: 3.81-4.10kb, 11: 4.11-4.50kb, 12: 4.51-4.90kb, 13: 4.91-5.40kb, 14: 5.41-5.90kb, 15: 5.91-6.50kb, 16: 6.51-7.10kb, 17: 7.11-7.80kb, 18: 7.81-8.60kb, 19: 8.61-9.50kb, 20: 9.51-10.50kb.

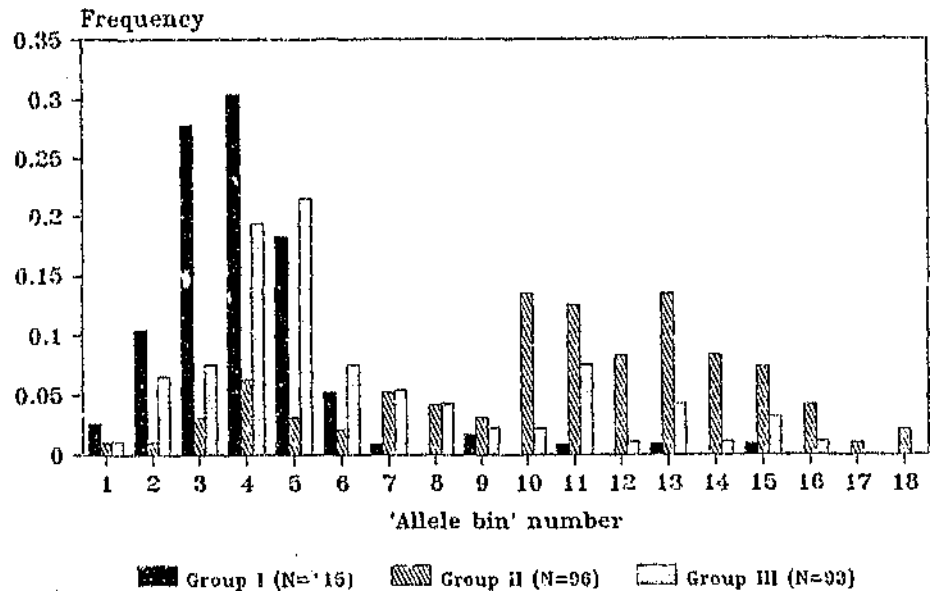


FIGURE 4.2 3'HVR distributions for major haplotype groups

Haplotypes are grouped as shown in Table 4.3.
'Allele bin' numbers correspond to 'allele bins' as follows:
1: <1.9kb, 2: 1.91-2.10kb, 3: 2.11-2.30kb, 4: 2.31-2.50kb, 5: 2.51-2.70kb, 6: 2.71-2.90kb, 7: 2.91-3.20kb, 8: 3.21-3.50kb, 9: 3.51-3.80kb, 10: 3.81-4.10kb, 11: 4.11-4.50kb, 12: 4.51-4.90kb, 13: 4.91-5.40kb, 14: 5.41-5.90kb, 15: 5.91-6.50kb, 16: 6.51-7.10kb, 17: 7.11-7.80kb, 18: 7.81-8.60kb.

4.2 Discussion

4.2.1 Types of -α chromosomes in South African Asian Indians

The -α chromosomes, encountered in almost all racial groups originating in tropical and subtropical regions of the world, are thought to be the result of unequal crossover between homologous subsegments surrounding the α globin genes. The geographical distributions of the $-\alpha^{3.7}$ and $-\alpha^{4.2}$ chromosomes are discussed in Section 1.9.1.

The South African Asian Indian data, consistent with the world trends, show a higher proportion of $-\alpha^{3.7}$ than $-\alpha^{4.2}$ chromosomes in all the groups studied, though the relative contributions of $-\alpha^{4.2}$ chromosomes to the total differ considerably among the groups.

Using the calculated -α haplotype frequencies (Table 3.6) and the relative proportions of $-\alpha^{3.7}$:$-\alpha^{4.2}$ (Table 4.1), the frequency of the $-\alpha^{4.2}$ chromosome could be estimated to be 0.002-0.005 in all the major groups, except the Hindu Hindi. In the latter, the ratio of $-\alpha^{3.7}$:$-\alpha^{4.2}$ chromosomes was only 2:1 and the frequency of $-\alpha^{4.2}$ was estimated to be 0.017. As few studies have been done on the different subgroups in India, it is difficult to compare these data with those of the corresponding parent groups in India. A ratio of 8:1 has been described in a previous mixed Indian sample from Punjab, Gujerat and Kashmir (Hill *et al.* 1985b), in north and north-west India. This is higher than that observed in the Hindu Hindi, but much lower than the ratios observed in the Moslem Gujerati and Hindu Gujerati in this study, who also originate from Gujerat.

As the Hindu Hindi originate from the more central and north-eastern regions of the Indian subcontinent, the higher $-\alpha^{4.2}$ frequency may reflect a south-east Asian influence, as the frequency of the $-\alpha^{4.2}$ chromosome is higher in south-east Asia (Embury *et al.* 1980b). Alternatively it may be the result of genetic drift in the parent population, especially as over 28% of the South African Hindu Hindi originated from a limited geographical area, namely a small number of districts in the United provinces of Agra and Oudh in India (Bhana 1991) and each community maintains a system of intra-caste marriages. In addition, many of the Indians who came to South Africa came in family groups and they intermarried within their caste system, so some rare alleles may have

increased in frequency due to genetic drift or founder effect.

The distribution of $-\alpha^{3.7}$ subtypes in the present study is as expected. The $-\alpha^{3.7I}$ chromosome, commonest in most world populations studied (Higgs *et al.* 1984), occurs most frequently, with a small number of $-\alpha^{3.7II}$ chromosomes, only in the Moslem Gujerati. As expected, the $-\alpha^{3.7III}$ chromosome, restricted to Polynesia and Melanesia (Hill *et al.* 1985b, Hill *et al.* 1987a) and the most prevalent type in Melanesia, was not found in this study.

There has been much speculation on the reasons for the varying frequencies of the different types of $-\alpha$ chromosome and particularly for the higher $-\alpha^{3.7}$ frequency in most populations studied. Though all are derived from crossover events in regions of high homology (Higgs *et al.* 1984), it has been suggested that there may be higher crossover constraints for the $-\alpha^{4.2}$ chromosome (Higgs *et al.* 1989). In the $-\alpha^{3.7}$ subtypes, the frequencies correlate with the length of the homologous region that serves as a target area for the crossover (Higgs *et al.* 1989). Alternatively the $-\alpha^{4.2}$ chromosome may have evolved more recently or the $-\alpha^{3.7}$ chromosome may be at a greater selective advantage (Trent *et al.* 1981b, Higgs *et al.* 1989). The $-\alpha^{4.2}$ chromosome is thought to produce a marginally more severe phenotype than the $-\alpha^{3.7}$ type, as reflected by a higher percentage of Hb Bart's in the cord bloods of $-\alpha^{4.2}$ homozygotes (Bowden *et al.* 1987), though there are no significant differences in the mean cell volumes or mean cell haemoglobin levels of adult heterozygotes (Bowden *et al.* 1985).

A further type of $-\alpha$ chromosome with a 3.5kb deletion, reported in Orissa State, India (Kulozik *et al.* 1988a), was not observed in this study.

4.2.2 Non-deletion α thalassaemia

There is also evidence for the presence of non-deletion α thalassaemia in the South African Indians from one consanguineous Moslem Gujerati family in which two of the children had relatively severe HbH disease. Neither parent had any α globin gene deletions. The children had inherited an identical chromosome 16 from each of their parents, as demonstrated by α globin cluster haplotype analysis, as well as analysis of the

5'HVR and 3'HVR alleles in the family. It is proposed that HbH disease in these individuals is due to homozygosity for an $\alpha 2$ mutation. The $\alpha 2$ gene has been shown to encode a 2-3 fold higher steady state level of mRNA (Liebhaber and Kan 1981, Orkin and Goff 1981, Liebhaber and Cash 1985) and 2-3 fold more protein than $\alpha 1$ (Liebhaber *et al.* 1986). Thus mutations in the $\alpha 2$ gene are likely to have a greater effect than those in $\alpha 1$, as they result in a loss of 37% and 13% of activity respectively (Liebhaber *et al.* 1986).

4.2.3 The -- chromosome in Indians

A deletion removing 22.8-23.7kb of the α globin cluster was originally described in a South African 'Coloured' individual (Vandenplas *et al.* 1987) and subsequently in an individual from Vayra in Gujerat (Drysdale and Higgs 1988) and a South African Indian family (Fei *et al.* 1992). An Indian origin for the deletion was therefore proposed. No individuals with the deletion were ascertained in the present study. In view of the rarity of HbH disease and Hb Bart's hydrops fetalis in Indians, the deletion is unlikely to be common.

4.2.4 Other α globin cluster rearrangements

The geographical distribution of the α globin cluster rearrangements is reviewed in Section 1.9.1.

The $\alpha\alpha\alpha^{anti3.7l}$ was the commonest $\alpha\alpha\alpha$ chromosome in the South African Asian Indians, as it is in most populations (reviewed in Weatherall *et al.* 1988). As expected, it occurs at a lower frequency than the $-\alpha$ chromosome in all the groups, except the Hindu Hindi where it reaches a relatively high frequency of 0.083, higher than the $-\alpha$ frequency of 0.050. Again it is difficult to determine whether this is typical of the parent population in India or is a feature specific to the local group, as discussed for the $-\alpha^{3.7}$:$-\alpha^{4.2}$ ratio (see Section 4.2.1). In a study of Dravidian Indians in Singapore (Tan *et al.* 1991) a $\alpha\alpha\alpha$ frequency of 0.0048 was reported, in a similar range to the minimum Moslem Gujerati frequency.

The ζζζ and -ζ chromosomes occur sporadically in the South African Indians as in most other populations. None of the ζζζ chromosomes in this study was associated with the *Bgl*II polymorphism linked to the ζζζ chromosome in Polynesians and Melanesians (Trent *et al.* 1986, Hertzberg *et al.* 1988).

4.2.5 α globin RFLP frequencies

4.2.5.1 RFLP frequencies on αα chromosomes

The α globin cluster is highly polymorphic (Higgs *et al.* 1986) and only a limited number of the known polymorphic sites have been tested in this study. However, even with these, additional variation has been observed.

The three *PZ* alleles at the PZ/Z locus have been previously described in a study of Indians from Andhra Pradesh (Fodde *et al.* 1991) and in south-east Asia (O'Shaughnessy *et al.* 1990). The *PZ3* allele occurs in the Hindu Gujerati and Hindu Hindi, but not in the Moslem Gujerati or the Hindu Tamil. The new *Bgl*I *H* allele similarly occurs at highest frequency in the former two groups.

Although the allele frequencies of the eight α globin associated RFLPs do not show significant differences among the major religious and language groups of Indians in South Africa on αα chromosomes, there do appear to be some significant differences from previously published Indian data (Higgs *et al.* 1986). The latter sample was drawn predominantly from the Punjab region of northern India and not classified according to religious or language origins (Dr DR Higgs, personal communication 1992). Compared to the sample of Higgs *et al.* (1986), there was a significant increase in the frequency of the Z allele at the PZ/Z locus (site 6) in the Moslem Gujerati ($p<0.001$), Hindu Gujerati ($p<0.05$) and Hindu Tamil ($p<0.01$); in the frequency of the *Bgl*I (site 4) + and *H* alleles in the Moslem Gujerati ($p<0.05$), Hindu Hindi ($p<0.05$) and Hindu Tamil ($p<0.001$) and in the frequency of the *Rsa*I/α (site 7) - allele in the Moslem Gujerati ($p<0.05$) and Hindu Tamil ($p<0.05$). The differences may be due to samples originating from different areas, as none of the South African Indian groups originates from Punjab. South African Indians also differ from Punjabis in studies of the ß globin cluster (see

Sections 5.2.3 and 6.2 for discussion). It is interesting that the *Z* allele at the PZ/Z locus and the *BglI* + allele, partly responsible for the differences have previously been described as common in Papua New Guinea and Melanesia (Higgs *et al.* 1986), though the *Z* allele is seen in most human populations at low frequencies (Hill *et al.* 1985a). Hill *et al.* (1985a) also describe an Indian sample of undefined origin with a frequency of the *Z* allele of 16%, similar to the 14% of Higgs *et al.* (1986). This is in contrast to the findings of the present study where frequencies of between 24 and 37% were observed, the highest being in the Hindu Tamil from the south of the subcontinent.

The South African Indians come from the western, southern and central regions of the Indian subcontinent (see Figure 1.4), regions perhaps less vulnerable to invasion and admixture than the northern areas. The alleles found further south may represent the older, ancestral types, which have been less influenced by types brought by the peoples who have invaded the Indian subcontinent from the north.

On $\alpha\alpha$ chromosomes, the *Z* allele was found in association with a limited number of IZHVR sizes and *SstI* alleles, namely *L*+, *M*- and *M*+, though predominantly with *M*-, perhaps suggesting a limited number of origins of this mutation in India.

In the IZHVR system, the *M* allele predominates in the South African Indians, as in most other populations studied (Goodbourn *et al.* 1984, Higgs *et al.* 1986). There were also significant frequencies of the *S* and *L* alleles. The *S* allele has also been described previously in populations from Melanesia, Papua New Guinea and Asia (Goodbourn *et al.* 1984, Chapman *et al.* 1986, Higgs *et al.* 1986), while the *L* allele is common in Caucasoids and Asians (Goodbourn *et al.* 1984, Higgs *et al.* 1986).

The Moslem Memon appeared to have a high frequency of the *RsaI*/α + allele, though the sample was very small. The individuals of this group came from a few towns and villages in the Kathiawar region of Gujerat, and settled in South Africa in two relatively isolated groups (Bhana and Brain 1990) where they have maintained a high rate of consanguineous and endogamous marriages. The may thus have unusual allele frequencies due to their limited founder numbers.

4.2.5.2 RFLP frequencies on -α chromosomes

Though the numbers of -α chromosomes were small, comparisons among the groups did show some significant differences. The differences in frequencies between the Hindu Hindi -α chromosomes and those of the other groups can be accounted for partly by the relatively high frequency of $-\alpha^{4.2}$ chromosomes and also apparently more limited origins of the $-\alpha^{3.7}$ chromosome in the Hindu Hindi (see Section 4.2.6.2 for discussion).

Differences also occur in allele frequencies on the -α and αα chromosomes for some systems. As the -α chromosome frequency is likely to have increased by natural selection, a different distribution of alleles from the αα chromosomes is not unexpected, the allele frequencies on the -α chromosomes depending on which original subset of parent chromosomes gave rise to the -α chromosomes.

4.2.6 α globin haplotypes

4.2.6.1 Haplotypes on αα chromosomes

The α globin haplotype is characterised by eight dimorphic RFLPs and a ninth marker, with three alleles, for the purposes of analysis. The *Acc*I/α site was excluded from this analysis, because of the high cost of the enzyme. Only 50 out of a possible 768 haplotype combinations, using the nine sites, have been observed in population studies, suggesting linkage disequilibrium of alleles between these polymorphic markers. In all populations the commonest haplotype frequency exceeds that predicted by a factor of 5-50, though the average heterozygosity was high at 0.93 (Higgs *et al.* 1986). In the present study, a total of 27 eight-site haplotypes was observed and the heterozygosity rates in all groups were high, $\geq$0.83. Even with grouping of similar alleles for haplotyping, a high level of polymorphism remains. Type Ia was the most frequent haplotype in all groups, although chromosomes of groups I, II and III each account for similar proportions of the total, about one-third each.

The haplotype distributions in the South African Asian Indian groups did not differ significantly, but all had significantly different haplotype frequencies from a previously

4.2.5.2 RFLP frequencies on -α chromosomes

Though the numbers of -α chromosomes were small, comparisons among the groups did show some significant differences. The differences in frequencies between the Hindu Hindi -α chromosomes and those of the other groups can be accounted for partly by the relatively high frequency of $-\alpha^{4.2}$ chromosomes and also the apparently more limited origins of the $-\alpha^{3.7}$ chromosome in the Hindu Hindi (see Section 4.2.6.2 for discussion).

Differences also occur in allele frequencies on the -α and αα chromosomes for some systems. As the -α chromosome frequency is likely to have increased by natural selection, a different distribution of alleles from the αα chromosomes is not unexpected, the allele frequencies on the -α chromosomes depending on which original subset of parent chromosomes gave rise to the -α chromosomes.

4.2.6 α globin haplotypes

4.2.6.1 Haplotypes on αα chromosomes

The α globin haplotype is characterised by eight dimorphic RFLPs and a ninth marker, with three alleles, for the purposes of analysis. The *Acc*I/α site was excluded from this analysis, because of the high cost of the enzyme. Only 50 out of a possible 768 haplotype combinations, using the nine sites, have been observed in population studies, suggesting linkage disequilibrium of alleles between these polymorphic markers. In all populations the commonest haplotype frequency exceeds that predicted by a factor of 5-50, though the average heterozygosity was high at 0.93 (Higgs *et al.* 1986). In the present study, a total of 2^7 eight-site haplotypes was observed and the heterozygosity rates in all groups were high, $\geq$0.83. Even with grouping of similar alleles for haplotyping, a high level of polymorphism remains. Type Ia was the most frequent haplotype in all groups, although chromosomes of groups I, II and III each account for similar proportions of the total, about one-third each.

The haplotype distributions in the South African Asian Indian groups did not differ significantly, but all had significantly different haplotype frequencies from a previously

described Indian sample ($p < 0.01$ in all except the Hindu Gujerati group where $p < 0.02$) (Higgs *et al.* 1986). These data reinforce the findings of the RFLP data. The differences were mainly due to an increase in the frequency of group III haplotypes, particularly types IIIa and IIIb in the present study. Group IV haplotypes were also observed in all groups in this study but were absent from the previous sample (Higgs *et al.* 1986). More recently, Higgs *et al.* (1989) have noted that types Ia, IIa and IIIa are present in India, but that IVa has not been recorded.

The haplotypes in the sample of Higgs *et al.* (1986) were determined from homozygotes and thus the distributions noted were not true frequencies, since commoner haplotypes would have been over-represented. In addition, haplotypes were typed Ia if they had an *Rsa*I + allele, but there are other group I haplotypes, which may therefore have been underestimated.

Five haplotypes predominate in the world, Ia, IIa, IIIa, IVa, VIIa, with Ia commonest in Caucasoids (British, Mediterranean, Saudi Arabia and Indian) (Higgs *et al.* 1986). Types Ia, IIa and IId/e constitute 90% of the haplotypes in south-east Asia, while in Melanesia types IIIa, IVa and Vc predominate (O'Shaughnessy *et al.* 1990) and in Polynesia Ia, IIa and IVa are commonest (reviewed in Flint *et al.* 1993b). Type IIIa was shown to be the second commonest haplotype in the Hindu Gujerati in this study, as it was in a study of tribal populations of Andhra Pradesh (Fodde *et al.* 1991). Type IIIa is the third most common haplotype in Sri Lanka and Saudi Arabia (reviewed in Flint *et al.* 1993b). In the Moslem Gujerati and the Hindu Tamil type IIb was the second most common. Thus the groups from India appear to represent a mixture of Caucasoid and south-east Asian types. The origin of the high frequency of group III haplotypes is difficult to determine, perhaps the result of founder effect and genetic drift in the relative isolation of the Indian subcontinent. Alternatively, group III haplotypes may represent the ancestral haplotype on the Indian subcontinent.

The exact mechanisms underlying the differences between populations are still poorly understood, however it has been proposed that the common haplotypes must have preceded human racial divergence and have been inherited as stable linkage groups since. Rarer haplotypes may have arisen in different populations by recombination or mutation

(Higgs *et al.* 1986). The association of particular haplotypes and *PZ* alleles reinforces the concept of stable linkage groups, as does the presence of the ψζ2-ζ1 polymorphism in all racial groups (Hill *et al.* 1985a). Similar associations to those in this study between *PZ* types and haplotypes have also been observed in south-east Asia (O'Shaughnessy *et al.* 1990), though Fodde *et al.* (1991) also observed *PZ1* in type I haplotypes.

Similarly the insertion in ζ2 reported in the Hindu Gujerati and Hindu Hindi groups in this study has only been observed in association with a single haplotype (IIj), suggesting a single origin. In addition, the haplotype appears to be associated exclusively with the *PZ3* allele and the *BglI H* allele. The insertion has been previously observed in Indians of unspecified origin and Sri Lankans (Dr DR Higgs, personal communication 1992), though no haplotype associations have been described.

4.2.6.2 α thalassaemia haplotypes

The -α chromosome haplotypes were studied in an attempt to determine the number of mutations which have occurred. The -α chromosomes are assumed to be the result of interchromosomal crossover and the resultant associated haplotype is the result of recombination between two haplotypes. It is difficult to determine the derivative chromosomes of the -α haplotypes unequivocally. The 5′ ends are not necessarily unique to any particular haplotype (see Table 4.3a), while at the 3′ end minimal information is available, because of deleted sites (the *RsaI*/α site in the crossover resulting in the $-\alpha^{4.2}$ chromosome and the *PstI*/α site in the crossover resulting in the $-\alpha^{3.7}$ chromosome). Further, the remaining 3′ sites have low frequencies of the rarer allele and thus provide little information. There is also no fixed reference point in the cluster, like the framework in the ß globin cluster. In the assignment of $-\alpha^{3.7}$ chromosomes to haplotypes, the *RsaI*/α polymorphism is useful in defining the 5′ origin of the haplotypes, but the 3′ end is still equivocal.

Each different -α haplotype may represent a new mutation. Alternatively, each of the types and subtypes of -α mutation may have occurred once and the present day haplotypes may be the result of an accumulation of mutation and recombination events on these chromosomes. Natural selection would then have increased the frequencies of the different

-α chromosomes. The true number of -α mutational events is probably somewhere between these two theories, but is difficult to estimate.

Twelve different haplotypes (with PZ subtyping included) were associated with the $-\alpha^{3.7\text{I}}$ chromosomes, one with the $-\alpha^{3.7\text{II}}$ and two with the $-\alpha^{4.2}$ chromosomes, thus providing an estimate of at least 15 origins for the -α mutations in South African Asian Indians. This number could be expanded if the different sized IZHVR alleles were included. In addition, if one includes the 3'HVR and 5'HVR alleles at either end of the haplotype, each of the haplotypes could be divided further. However, if one attempts to determine a minimum number of origins, it can be assumed that different size VNTR alleles are due to recent mutation, although this is not necessarily supported by analysis of haplotype/VNTR association, which seems to have been maintained over a long period (see discussion in Section 4.2.7). Further, those haplotypes that differ by a single restriction enzyme site may be assumed to differ because of recent mutation. All the $-\alpha^{3.7\text{I}}$ haplotypes observed in South African Indians could therefore be derived from three or four original mutations, if recombination events and point mutations were important in contributing to the current variability. Recombination could decrease this number further. Although the two $-\alpha^{4.2}$ haplotypes cannot be derived from one another by a single mutation event, a single crossover 3' to IZHVR could explain a single origin for the two haplotypes. Thus there could be as few as five or six origins of -α chromosomes in South African Indians.

There is some evidence for increased crossover on chromosome 16p in general (Donis-Keller *et al.* 1987). There is increased male recombination at both chromosome 16 telomeres (Julier *et al.* 1990) and increased chiasmata in the region of 16p (Saadallah and Hultén 1983). Areas with an increased GC content and hypomethylation, which have been implicated in prokaryotes as regions of high recombination (Fischel-Ghodsian *et al.* 1987a), also exist in the region of the α globin cluster. In addition, it is thought that tandem repeats, which are prevalent in the cluster, may promote recombination events (Jeffreys *et al.* 1985, Jarman *et al.* 1986).

A hotspot for recombination has been proposed in the α cluster, between IZHVR and IVS1, in exon 1 of ζ1, though none has been definitively demonstrated and no linkage

disequilibrium is observed on either side of IZHVR (Higgs *et al.* 1986, Higgs *et al.* 1989, Fodde *et al.* 1991). A breakpoint cluster region has also been proposed upstream of 3'HVR (Nicholls *et al.* 1987) and the presence of $\alpha\alpha\alpha/-\alpha$ and $\zeta\zeta\zeta/-\zeta$ chromosomes have been used as evidence for the occurrence of homologous interchromosomal crossover (Winichagoon *et al.* 1982). There is also some evidence for gene conversion (Hill *et al.* 1985a) and interchromatid or intrachromosomal events (Higgs *et al.* 1986). On the other hand, it has been suggested that recombination may be somewhat suppressed in the cluster (Ramsay and Jenkins 1988). No recombination events between 5'HVR and 3'HVR within the cluster were observed in 157 paternal and 157 maternal informative meioses in this study or by Jarman and Higgs (1988) and Higgs *et al.* (1989) in 500 meioses.

The ubiquity of the $-\alpha$ chromosome suggests that it may be a very old mutation (Ramsay and Jenkins 1984). In addition, it is known that natural selection has acted to increase its frequency. However, the true number of origins remains difficult to determine. It seems likely that the true number of mutations lies somewhere between the minimum and maximum estimates. Further it is also uncertain whether the same mutation in different populations suggests common ancient origin or recurrent mutation (Chapman *et al.* 1986).

Evidence for the association of multiple haplotypes with $-\alpha$ was first demonstrated by showing no consistent linkage between $-\alpha$ types and IZHVR (Goodbourn *et al.* 1984). Other evidence for multiple mutation events comes from the association of $-\alpha$ and different haemoglobin variants and HVR alleles (Goodbourn *et al.* 1984, Flint *et al.* 1986).

The findings in this study are in keeping with studies elsewhere, where the $-\alpha$ chromosome is associated with a number of haplotypes and thus thought to have multiple origins. Multiple origins have also been proposed for the $-\alpha$ chromosomes of Kenyans (Ojwang *et al.* 1987), Australian Aborigines (Tsintsof *et al.* 1990) and southern Italians (Di Rienzo *et al.* 1986). Although tribal Indians in Orissa and Andhra Pradesh appeared to have a single origin (judged only by the *Rsa*I/α site) (Labie *et al.* 1989), a study in forest tribal communities from Andhra Pradesh in India showed genetic heterogeneity of α thalassaemia haplotypes. This was ascribed to multiple recombination events and natural selection (Fodde *et al.* 1991). The wide geographical distribution and multiple haplotypes

associated with the $-\alpha^{3.7I}$ chromosome are suggestive of multiple origins (Weatherall *et al.* 1988). In contrast, the $-\alpha^{3.7III}$ in Melanesia has a single haplotype associated with it (Hill *et al.* 1985b, Flint *et al.* 1986, O'Shaughnessy *et al.* 1990), as does the $-\alpha$ chromosome in Sardinia (Di Rienzo *et al.* 1985) and the different -- chromosomes (Goodbourn *et al.* 1984), which have increased due to selection and disseminated. A limited diversity of α globin haplotypes has been found in Polynesia and is thought to be due to common ancestral founders with genetic drift resulting in group differences (Hertzberg *et al.* 1988).

Fodde *et al.* (1991) proposed three origins for the $-\alpha^{4.2}$ chromosome in Andhra Pradesh. One of the $-\alpha^{4.2}$ haplotypes is similar in this study and in that of Fodde *et al.* (1991). Three of four south-east Asian and Polynesian $-\alpha^{4.2}$ haplotypes differ from those in India (Flint *et al.* 1986, Hill 1986). All three $-\alpha^{4.2}$ haplotypes described in Melanesia (Hill *et al.* 1989) are different from those in South African Indians.

4.2.6.3 Haplotypes associated with α globin cluster rearrangements

The $-\zeta$, $\zeta\zeta\zeta$ and $\alpha\alpha\alpha$ chromosomes may provide better information than the $-\alpha$ chromosomes on genetic affinities between populations as they are under no selective pressure but the data are relatively limited at present.

At least three different $\alpha\alpha\alpha$ chromosomes were detected in this study, all different from those previously described in Polynesia (Hertzberg *et al.* 1988). Because of the difficulties in ascertaining which haplotypes were involved in the crossover events to form the $-\alpha$ and $\alpha\alpha\alpha$ chromosomes, it is not possible to determine whether any of the mutations represent reciprocal chromosomes.

One of the two $\zeta\zeta\zeta$ chromosomes studied was of the south-east Asian type (Hill *et al.* 1985a, Hill *et al.* 1987a, Hertzberg *et al.* 1988). This chromosome is thought to have spread by migration through south-east Asia, Melanesia, Micronesia and Polynesia (Hill *et al.* 1989). The second $\zeta\zeta\zeta$ chromosome in this study has a different haplotype, which is unlikely to have arisen from the south-east Asian type. The $\zeta\zeta\zeta$ chromosomes were again unusual as the haplotypes from which they appear to be derived, IId/e and IIf, are

rare in India. A single south-east Asian ancestral origin of ζζζ has been proposed (Hertzberg *et al.* 1988, Trent *et al.* 1988, O'Shaughnessy *et al.* 1990), though the -ζ does not appear to be a reciprocal chromosome and is probably the result of a separate crossover event. The south-east Asian ζζζ chromosome is associated with the acquisition of a *Bgl*II polymorphism on the chromosome in Melanesia and Polynesia (Hill *et al.* 1987a). Similarly the second ζζζ chromosome in this study suggests at least two mutational events have given rise to the ζζζ chromosome. Even in Polynesia where each of the ααα, -α, ζζζ, -ζ chromosomes is associated with a single haplotype, at least four crossover events have to be proposed to explain the data (Hertzberg *et al.* 1988).

The single -ζ chromosome observed was interesting in that it had a (- + +) at the 3′ end typical of a Type IV haplotype, which is rare in Indians, but is the type associated with -ζ in Polynesia (Hertzberg *et al.* 1988). The 5′ end was uninformative as a number of the sites were deleted and two were indeterminate. It did however have a different *Xba*I site from the Polynesian chromosome. There is further evidence for multiple origins of -ζ as it occurs with and without the *Xba*I site and on chromosomes with and without α thalassaemia (Weatherall *et al.* 1988), and is associated with four haplotypes in American Blacks (Felice *et al.* 1986).

Some of the -ζ and ζζζ chromosomes in India appear to have a common origin with the south-east Asian types. They may represent old Asian mutations which have remained at low frequency in most of the populations as no natural selection has operated on them.

4.2.7 VNTR analysis

The region of the α cluster contains several loci with tandemly repeated segments of apparently related GC-rich sequences (Higgs *et al.* 1989), including the 5′HVR and the 3′HVR, which have a 57bp and a 17bp tandem repeat sequence respectively. The number of repeats may be altered by unequal genetic exchange at meiosis or mitosis or by DNA slippage during replication thus producing highly polymorphic regions.

In the VNTR systems studied the sizes of the repeats are small and thus alleles which differ by only a few repeats are not resolved using Southern blotting methods.

Nevertheless, a large number of alleles were observed at both of the VNTR loci, with high heterozygosities. The distributions are similar to those previously noted for these two systems (Jarman and Higgs 1988, Allen *et al.* 1989). There were no significant differences between the major groups, although the importance of small variations between groups is difficult to determine. The narrow 5'HVR distribution in the Hindu Hindi is perhaps suggestive of a small number of more closely related founders than in other groups.

Both the 5'HVR and 3'HVR distributions were skewed to the left, with smaller fragments predominating. This has been thought to reflect a balance between unequal interchromosomal exchange producing both large and small alleles and intramolecular deletions generating smaller alleles (Jarman and Higgs 1988). Others have proposed hitchhiking as the explanation, particularly as the distributions differ markedly from those of VNTRs at other loci (Deka *et al.* 1991). The large number of alleles at these loci together with the low sequence divergence of the repeat sequences is thought to point to rapid concerted evolution (Jarman *et al.* 1986). The mutation rate of 3'HVR has been shown to be high *in vitro* (Jarman *et al.* 1986), while 5'HVR is more stable (Jarman and Higgs 1988). These findings have not been verified *in vivo*, however. Estimates of the mutation rate in VNTR loci have varied, though the mean is estimated at 0.004 per DNA fragment per gamete (Jeffreys *et al.* 1988). In 314 informative meioses in this study, there were no paternal or maternal exclusions, which could be ascribed to new mutations, though the high heterozygosity rates suggest at least a 10-50 times higher mutation rate to account for the increased genetic variation (Deka *et al.* 1991). As the function of VNTRs is unknown, the significance of this is uncertain. They may play a role in recombination or other cellular functions (Higgs *et al.* 1989).

VNTR systems have been used widely in genetic studies in, for example, paternity testing and linkage studies. The 5'HVR and 3'HVR systems together provide a very powerful tool because of their high combined heterozygosity, with no high degree of linkage disequilibrium between them (Jarman and Higgs 1988).The 3'HVR system was used to localise the autosomal dominant polycystic kidney disease (ADPKD) locus (Reeders *et al.* 1985) and the familial Mediterranean fever locus (Shohat *et al.* 1992) to chromosome 16p. It was also used to exclude the presence of uniparental disomy as a cause of

Rubinstein-Taybi syndrome (Hennekam *et al.* 1993), after it was shown to be due to a submicroscopic interstitial deletion within 16p13.3 in 25% of patients (Breuning *et al.* 1993). In this study, the VNTR markers, together with the α globin haplotypes, were used to show that non-deletion HbH disease in the offspring of a consanguineous marriage was associated with homozygosity for 16p and probably with homozygosity for a point mutation in the α2 gene. At present, population studies using VNTRs are difficult because intra-population variation is greater than inter-population variation. They are potentially extremely useful, however, and may be able to provide some depth to other mutational events and haplotype origins (Higgs *et al.* 1986).

In this study a correlation was demonstrated between the haplotype groups and the allele sizes at the 5'HVR and 3'HVR loci (Figures 4.1 and 4.2). The haplotypes are classed according to their 3' ends and it is not surprising that there is a greater correlation with 3'HVR than 5'HVR allele sizes. In addition, the lesser correlation between 5'HVR and haplotypes may be due to the greater distance of 5'HVR from the cluster, as more distant markers show progressively less correlation with the established α globin haplotypes (Higgs *et al.* 1989). Higgs *et al.* (1986) have previously noted the uniformity in 3'HVR fragments with the Ia haplotype, findings which were less obvious with IIa and IIIa haplotypes, but the sample sizes were small. In this study distributions of allele sizes were compared in haplotype groups, rather than in individual haplotypes. Clear trends in the allele sizes in different groups were observed. Higgs *et al.* (1989) also showed an association between some haplotypes and the presence or absence of the *Rsa*I site in the 5'HVR region. In this study, all groups had 5'HVR alleles with and without the restriction enzyme site, though group I haplotypes do appear to have a lower proportion of fragments lacking the *Rsa*I site.

The differences in allele distribution in the different haplotype groups reinforce the idea that the haplotypes may be old and have relatively few origins in this population. The variation in VNTR size is distributed around the size of the founder chromosome. The limited distributions of VNTR allele sizes around distinct modes in each of the haplotype groups may provide some evidence to suggest that forward and backward mutation due to slippage and resulting in small allele size changes may be more common than larger allele size changes due to unequal recombination or sister chromatid exchange. The more

frequent alleles are often shared between populations and, it is proposed, may precede geographical dispersal (Deka *et al.* 1991). Detailed studies would be required to determine whether the same haplotypes in different populations are associated with similar allele distributions. If so, this would support the theory that haplotypes are ancient and have common origins in all populations. Their size correlation with common haplotypes remains to be determined in other populations.

All populations are thought to maintain numerous alleles and high heterozygosities for VNTRs, unless they have been through recent bottlenecks. This is well demonstrated in this study where all groups had heterozygosities greater than 0.9. One might also expect to find a deficiency of heterozygotes in VNTR analysis due to poor resolution of similarly sized fragments. This phenomenon was found in the total sample, but not in the individual groups, though in all groups the expected number of homozygotes was less than the observed number. The small numbers of observations may explain why the differences do not reach significance. Fragments may, in theory, differ by as little as one repeat sequence (17bp for 3'HVR or 57bp for 5'HVR), and thus fragments which are measured on a Southern blot to be the same size may actually be different. If one looks at the true homozygotes, individuals who share a complete α globin haplotype in addition to the same sized VNTR fragments, no heterozygote deficiency exists, reinforcing the idea that different alleles are being pooled in the analysis. An additional reason for excess homozygosity at a locus may be forward and backward mutation to similar sized alleles, though these would be distinguished by comparing the rest of the haplotype.

In some of the South African Indian groups the rate of consanguineous marriage may be as high as 30%. This could raise the number of individuals homozygous at a locus considerably. The study sample was not representative in that families in which the parents were consanguineous were excluded, so the true rate of homozygosity in the groups may be higher than that observed here. It is interesting that the high rate of intra-caste marriage does not seem to have decreased the heterozygosity markedly, perhaps because of a high new mutation rate, although the narrow Hindu Hindi distribution may in part be explained by the intra-caste marriage system.

CHAPTER 5 - CHARACTERISATION OF THE ß GLOBIN CLUSTER IN SOUTH AFRICAN ASIAN INDIANS

This chapter deals with the ß globin cluster polymorphisms and rearrangements that occur in South African Indians. The allele frequencies for the common ß globin cluster RFLPs are presented for chromosomes with normal ß globin genes ($ß^A$) and those carrying a mutation causing ß thalassaemia ($ß^T$). ß globin haplotypes were constructed, using family studies, for chromosomes carrying $ß^A$, $ß^S$, $ß^E$, $ß^D$, and $ß^T$ globin genes. The data were used to compare the different Indian groups and to try to define the origins of the different haemoglobin variants, excluding the ß thalassaemias, in Indians. The ß thalassaemia mutations and their origins are presented in Chapter 6. In addition the *Xmn*I site 5′ to $^G\gamma$ has been characterised and its relationship to the haplotypes is described. A new *Xmn*I variant is also reported.

Data presented for the major groups, namely Moslem Gujerati, Hindu Gujerati, Hindu Hindi and Hindu Tamil, were analysed statistically. For the minor groups, namely Moslem Urdu, Moslem Memon and Hindu Telegu, data are presented for completeness, but the samples were too small for statistical analysis in most cases.

5.1 Results

5.1.1 ß globin cluster rearrangements

The most common rearrangements in the ß globin cluster involve the two γ genes. The γγγ rearrangement was observed in three individuals, one Moslem Gujerati out of 70 individuals studied and two Hindu Hindi out of 45 studied. The frequencies of the rearrangement were thus estimated to be 0.007 ± 0.007 and 0.022 ± 0.015 in the Moslem Gujerati and Hindu Hindi, respectively. All occur on chromosomes with normal ß globin genes and appear to be of the $^G\gamma^G\gamma^A\gamma$ type, as there is no 4.2kb fragment after *Pst*I digestion and the 5.1kb fragment is increased in intensity (Trent *et al.* 1981a). The rearrangement was not found in 47 Hindu Gujerati nor in 34 Hindu Tamil individuals studied.

The -γ rearrangement of the $^{A}\gamma$ type was observed in one Hindu Tamil individual out of 39 studied, on a chromosome with a normal ß globin gene. The frequency of the rearrangement in this group was thus estimated to be 0.013 ± 0.013. It was not found in 70 Moslem Gujerati, 47 Hindu Gujerati or 45 Hindu Hindi individuals studied.

All the individuals with γγγ and -γ rearrangements had normal FBCs. Unfortunately HbF levels and $^{G}\gamma$:$^{A}\gamma$ ratios were not determined.

No $^{G}\gamma^{G}\gamma$ or $^{A}\gamma^{A}\gamma$ rearrangements were found in any of the groups.

5.1.2 ß globin cluster RFLPs

The allele frequencies for the nine common ß globin cluster RFLPs studied on $ß^{A}$ and $ß^{T}$ chromosomes are shown in Tables 5.1a and b for the major and minor South African Indian groups respectively.

For the $ß^{A}$ chromosomes, the allele frequencies differed significantly ($p<0.05$) between the Moslem Gujerati and Hindu Hindi groups for *Hinc*II/ε (System 1), *Hin*dIII/$^{G}\gamma$ (System 3), *Hin*dIII/$^{A}\gamma$ (System 4) and *Hinc*II/pP3.9 3′ site (System 6). The Hindu Hindi also differed significantly from the Hindu Tamil at *Hin*dIII/$^{G}\gamma$ ($p<0.05$) and *Hinc*II/pP3.9 5′ and 3′ sites (Systems 5 and 6) ($p<0.05$ and $p<0.01$ respectively). The Hindu Hindi and the Hindu Gujerati had significantly different allele frequencies at the *Hinc*II/pP3.9 3′ site ($p<0.05$). In addition, the Moslem Urdu differ from the Moslem Gujerati at *Hinc*II/ε ($p<0.05$), and the Hindu Tamil from the Hindu Telegu at *Hinc*II/pP3.9 ($p<0.05$).

The allele frequencies on the $ß^{T}$ chromosomes differed significantly ($p<0.05$) between the Hindu Hindi and Hindu Gujerati chromosomes at the *Hinc*II/ε, *Hin*dIII/$^{G}\gamma$ and *Hinc*II/pP3.9 3′ sites, between the Moslem Gujerati and Hindu Gujerati at the *Bam*HI/ß site ($p<0.05$) and between the Hindu Hindi and Moslem Memon at the *Hinc*II/ε, *Hin*dIII/$^{G}\gamma$ and *Hinc*II/pP3.9 5′ and 3′ sites ($p<0.05$ for all sites). The sample sizes are relatively small, however.

The allele frequencies on the $ß^{A}$ and $ß^{T}$ chromosomes were also compared in each of the

TABLE 5.1a Allele frequencies of the common β globin cluster polymorphisms for β^A and β^T chromosomes in the major South African Indian groups

SYSTEM[1]	ALLELE[2]	CHROMOSOME	MOSLEM GUJERATI			HINDU GUJERATI			HINDU HINDI			HINDU TAMIL		
			N[3]	Freq	SE	N[3]	Freq	SE	N[3]	Freq	SE	N[3]	Freq	SE
1	+	β^A	108	0.685	0.045	66	0.621	0.060	68	0.529	0.061	57	0.632	0.064
		β^T	22	0.636	0.103	8	0.500	0.177	10	1.000	-	6	0.500	0.204
2	+	β^A	108	0.269	0.043	65	0.169	0.046	68	0.294	0.055	57	0.175	0.050
		β^T	22	0.227	0.089	8	0.375	0.171	10	0.000	-	6	0.333	0.192
3	+	β^A	108	0.380	0.047	66	0.409	0.061	68	0.559	0.060	57	0.368	0.064
		β^T	22	0.364	0.103	8	0.500	0.177	10	0.000	-	6	0.500	0.204
4	+	β^A	108	0.111	0.030	66	0.182	0.047	68	0.250	0.053	57	0.211	0.054
		β^T	22	0.136	0.073	8	0.125	0.117	10	0.900	-	6	0.167	0.152
5	+	β^A	108	0.278	0.043	66	0.197	0.049	68	0.309	0.056	57	0.158	0.048
		β^T	22	0.227	0.089	8	0.375	0.171	10	0.000	-	6	0.333	0.192
6	+	β^A	108	0.389	0.047	66	0.379	0.060	68	0.559	0.060	57	0.316	0.062
		β^T	22	0.364	0.103	8	0.500	0.177	1	0.000	-	6	0.500	0.204
7	+	β^A	108	0.676	0.045	66	0.636	0.059	68	0.544	0.060	57	0.684	0.062
		β^T	22	0.273	0.095	8	0.500	0.177	10	0.100	0.095	6	0.167	0.152
8	+	β^A	108	0.495	0.048	66	0.500	0.062	68	0.485	0.061	57	0.614	0.064
		β^T	22	0.136	0.073	8	0.250	0.153	10	0.100	0.095	6	0.000	-
9	+	β^A	106	0.815	0.037	66	0.742	0.054	68	0.779	0.050	57	0.842	0.048
		β^T		1.000	-	8	0.500	0.177	10	0.900	0.095	6	1.000	-

[1] The numbers of the systems correspond to the sites labelled in Figure 2.2 and the RFLPs described in Table 2.5.
[2] All systems are biallelic and the frequency of the '+' allele is shown.
[3] N = number of chromosomes studied.

TABLE 5.1b Allele frequencies of the common β globin cluster polymorphisms for β^A and β^T chromosomes in the minor South African Indian groups

SYSTEM[1]	ALLELE[2]	CHROMOSOME	MOSLEM URDU			MOSLEM MEMON			HINDU TELEGU		
			N[3]	Freq	SE	N[3]	Freq	SE	N[3]	Freq	SE
1	+	β^A	10	0.300	0.145	6	0.500	0.204	6	0.333	0.192
		β^T	6	0.667	0.192	5	0.400	0.219	1	1.000	-
2	+	β^A	8	0.250	0.153	6	0.167	0.152	6	0.500	0.204
		β^T	5	0.200	0.179	5	0.600	0.219	1	0.000	-
3	+	β^A	10	0.600	0.155	6	0.167	0.152	6	0.833	0.152
		β^T	6	0.333	0.192	5	0.600	0.219	1	0.000	-
4	+	β^A	10	0.300	0.145	6	0.167	0.152	6	0.333	0.192
		β^T	6	0.167	0.152	5	0.000	-	1	0.000	-
5	+	β^A	10	0.300	0.145	6	0.167	0.152	6	0.500	0.204
		β^T	6	0.167	0.152	5	0.600	0.219	1	0.000	-
6	+	β^A	10	0.600	0.155	6	0.333	0.192	6	0.833	0.152
		β^T	6	0.333	0.192	5	0.600	0.219	1	0.000	-
7	+	β^A	10	0.500	0.158	6	0.833	0.152	6	0.667	0.192
		β^T	6	0.167	0.152	5	0.600	0.219	1	0.000	-
8	+	β^A	10	0.500	0.158	6	0.833	0.152	6	0.500	0.204
		β^T	6	0.000	-	5	0.000	-	1	0.000	-
9	+	β^A	10	0.800	0.126	6	0.667	0.192	6	1.000	-
		β^T	6	1.000	-	5	1.000	-	1	1.000	-

[1] The numbers of the systems correspond to the sites labelled in Figure 2.2 and the RFLPs described in Table 2.5.
[2] All systems are biallelic and the frequency of the '+' allele is shown.
[3] N = number of chromosomes studied.

major groups. In the Hindu Hindi frequencies were significantly different for the *Hin*cII/ε ($p<0.05$). *Hin*dIII/$^{G}\gamma$ ($p<0.01$), pP3.9 3′ site ($p<0.01$) and *Ava*II/ß ($p<0.05$) systems. The *Ava*II/ß and *Hin*dIII/3′ß allele frequencies differed significantly in both the Moslem Gujerati ($p<0.01$) and the Hindu Tamil ($p<0.05$). In the minor groups, the numbers were too small for statistical analysis, but nevertheless in the Moslem Memon, the *Hin*dIII/3′ß frequencies differed between the $ß^{A}$ and $ß^{T}$ chromosomes ($p<0.05$).

A number of $ß^{S}$, $ß^{D}$ and $ß^{E}$ chromosomes were also studied. The allele frequencies were not calculated as the numbers in each religious and linguistic group were too small for comparison. The data are presented as haplotypes in Sections 5.1.3.3-5.1.3.5.

5.1.2.1 A new *Xmn*I variant

In contrast to the α globin cluster where the allele sizes vary considerably due to the presence of tandem repeats, the ß globin RFLPs are single site polymorphisms and thus alleles are generally of fixed size. Additional bands were observed, however, with *Xmn*I digestion and the $p^{A}\gamma$ probe in individuals from three families, one Moslem Gujerati and two Hindu Gujerati.

When DNA is digested with *Xmn*I and hybridised to the probe $p^{A}\gamma$ an 8.2kb fragment is produced when the polymorphic site at -158 to $^{G}\gamma$ is absent and a 7kb fragment when it is present (Gilman and Huisman 1985). In the Moslem Gujerati family additional fragments of 4.9kb and 2.1kb were seen and in both Hindu Gujerati families, fragments of 6.1 and 2.1kb were seen (see Figure 5.1).

The *Hin*dIII/$^{G}\gamma$ and *Hin*dIII/$^{A}\gamma$ polymorphisms had previously been characterised in these families and no aberrant bands had been noted. Aberrant bands were also not seen with the $p^{A}\gamma$ probe and the enzymes *Bgl*II or *Pst*I in these families, suggesting that the variants were point mutations altering sites for the enzyme *Xmn*I.

In an attempt to define these variants further, double digests with *Xmn*I and *Hin*dIII were done and a restriction map constructed. This is shown in Figure 5.2. As *Hin*dIII cuts between the two γ genes, it was hoped that this enzyme would help to localise the variant

FIGURE 5.1 Autoradiograph following *Xmn*I digestion and $p^A\gamma$ hybridisation, demonstrating the normal *Xmn*I/γ polymorphism and the variants observed in this study

When DNA is digested with *Xmn* I and hybridised to the probe $p^A\gamma$ an 8.2kb fragment is produced when the polymorphic site at -158 to $^G\gamma$ is absent and a 7kb fragment when it is present. Individuals I-1 in family A, I-2 in family B and I-1 in family C are all heterozygotes for this polymorphism, while II-1 in family A and II-2 in family B are homozygotes for the 8.2kb and 7kb fragments, respectively.

Family A is a Moslem Gujerati family, in which additional fragments of 4.9kb and 2.1kb were seen (individuals I-1 and II-2).

Families B and C are both Hindu Gujerati families, in which additional fragments of 6.1 and 2.1kb were seen (Family B, individuals I-1, II-1 and II-3; Family C, individuals I-2 and II-1).

Lanes marked M1 and M2 represent the *Hin*dIII and *Hin*dIII/*Eco*RI molecular weight markers, respectively.

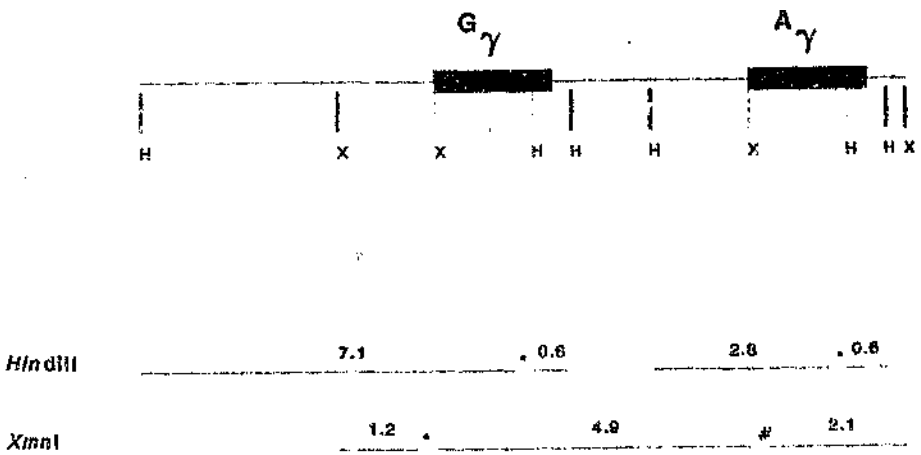


FIGURE 5.2 Restriction map of the $^{G}\gamma$ and $^{A}\gamma$ genes, showing the *Hind*III and *Xmn*I sites

Constant sites are indicated with solid lines. Polymorphic sites are marked with an asterisk and indicated with dotted lines. The proposed position of the variant site observed in this study, at position -158 to the $^{A}\gamma$ gene, is shown with a # and a broken line.

In the Moslem Gujerati family studied, the *Xmn*I site 5' to $^{G}\gamma$ is also present, resulting in the 4.9kb fragment, whereas in the 2 Hindu Gujerati families it is absent, resulting in the 6.1kb fragment.

to one of the two genes. Both *Xmn*I variants were shown to be due to the presence of an additional *Xmn*I site 5′ to $^{A}\gamma$ gene. In view of the high homology between the $^{A}\gamma$ and $^{G}\gamma$ genes, it is proposed that it is the same site in all the families and that it is at position -158 to $^{A}\gamma$, in the homologous position to the *Xmn*I site 5′ to $^{G}\gamma$. Definitive proof must await sequencing. In the case of the Moslem Gujerati family the site 5′ to $^{G}\gamma$ is also present, resulting in the 4.9kb fragment, whereas in the two Hindu Gujerati families it is absent, resulting in the 6.1kb fragment, as shown in Figure 5.2. In both Hindu Gujerati families the variant was on a β^{T} chromosome, carrying the frameshift mutation at codon 41/42, while in the Moslem Gujerati it was on a β^{A} chromosome.

In the Moslem Gujerati group, the variant was observed on one of 136 chromosomes studied, giving a gene frequency of 0.007 ± 0.007. In the Hindu Gujerati, the frequency of the variant could not be determined as it occurred on β^{T} chromosomes and the two families in which it was found were ascertained non-randomly through their thalassaemia major child.

The HbF levels in the two members of the Moslem Gujerati family with the variant were 1.9% and 1.2%, respectively. In one Hindu Gujerati family, HbF levels of 1.6% and 0.8% were found in two members. In the second family HbF levels were not available. The ratio of $^{G}\gamma$:$^{A}\gamma$ chains was not determined.

5.1.3 ß globin haplotypes

ß globin haplotypes were determined using eight RFLPs (sites 1, 3-9 shown in Figure 2.2 and described in Table 2.5). The linkage phase was assigned with family studies. A total of 445 chromosomes were studied in 111 families of different religious and language groups and 403 haplotypes could be unequivocally determined.

No consistent nomenclature has been formulated for the ß globin haplotypes. In addition, many studies have excluded the *Hin*dIII/3′ß (Site 8) from their analysis making complete comparisons difficult. A hotspot for recombination has been shown to exist between the δ and ß genes (Antonarakis *et al.* 1982a, Chakravarti *et al.* 1984), and thus, for the purposes of analysis, the haplotypes are grouped according to their three 3′ sites, which

lie 3' to the hotspot. The haplotypes are numbered consecutively and these numbers will be used in the discussion and for comparison with previous studies.

5.1.3.1 β^A haplotypes

A total of 320 β^A haplotypes has been determined. The β^A haplotypes and their frequencies are shown in Tables 5.2a and 5.2b for the major and minor groups respectively. Of the 29 haplotypes observed, nine have been described previously on β^A chromosomes and three on β^T chromosomes in Asian Indians (Kazazian *et al.* 1984a). In addition, a further eight haplotypes were similar at all sites studied (Old *et al.* 1984, Thein *et al.* 1984b,1988, Varawalla *et al.* 1992) but the *Hin*dIII/3'ß was not defined and may therefore differ. Thus nine haplotypes not previously described in Asian Indians are reported. Five haplotypes previously observed at low frequency were not found in this study (Varawalla *et al.* 1992).

In each major group four common haplotypes account for 57-61.5% of the chromosomes. There were no significant differences in the overall haplotype frequencies between the major groups, though the distributions of the haplotypes vary somewhat. Haplotypes 1 and 16 occur at the highest frequencies in all groups, each accounting for between 13.6% and 26.3% of chromosomes. In the Moslem Gujerati and Hindu Tamil haplotype 16 was the commonest, whereas haplotype 1 was the most common in the Hindu Gujerati and Hindu Hindi. Together the two haplotypes account for between 34.8% and 47.4% of chromosomes in all the major groups. Haplotypes 7 and 22 were the next most common in the Moslem Gujerati, while in the Hindu Gujerati haplotypes 17 and 22, in the Hindu Hindi haplotypes 2 and 24 and in the Hindu Tamil haplotypes 18 and 22 make up the remainder of the common haplotypes. The distributions of these haplotypes in the major groups are shown in Figure 5.3.

It is difficult to comment on the haplotype frequencies in the minor groups as the numbers of chromosomes studied were very small. In general, similar haplotypes were seen, although a new haplotype (haplotype 29), not observed in any other group, was described in a Moslem Memon individual.

TABLE 5.2a Frequencies of β globin haplotypes on β^A chromosomes in the major South African Indian groups

HAPLOTYPE									MOSLEM GUJERATI		HINDU GUJERATI		HINDU HINDI		HINDU TAMIL	
Type[1]	Description[2]								(N=107)[3]		(N=66)[3]		(N=68)[3]		(N=57)[3]	
									Freq	SE	Freq	SE	Freq	SE	Freq	SE
1	+	−	−	−	−	−	−	+	0.168	0.036	0.212	0.050	0.206	0.049	0.211	0.054
2	−	+	−	+	+	−	−	+	0.093	0.028	0.076	0.033	0.088	0.034	0.035	0.024
3	−	+	+	−	+	−	−	+	0.019	0.013	0.061	0.029	0.074	0.032	0.053	0.030
4*	+	+	−	+	+	−	−	+	0.047	0.020	0.000	-	0.059	0.029	0.018	0.018
5#	−	+	+	−	−	−	−	+	0.000	-	0.015	0.015	0.015	0.015	0.000	-
6	−	+	−	−	+	−	−	+	0.000	-	0.000	-	0.015	0.015	0.000	-
7	+	−	−	−	−	+	−	+	0.112	0.030	0.076	0.033	0.015	0.015	0.053	0.030
8	−	+	−	+	+	+	−	+	0.019	0.013	0.000	-	0.029	0.020	0.000	-
9	−	+	+	−	+	+	−	+	0.009	0.009	0.015	0.015	0.000	-	0.000	-
10#	+	+	−	+	+	+	−	+	0.009	0.009	0.000	-	0.015	0.015	0.000	-
11#	−	+	+	−	−	+	−	+	0.009	0.009	0.000	-	0.000	-	0.000	-
12*	−	+	−	−	+	+	−	+	0.000	-	0.015	0.015	0.000	-	0.000	-
13*	−	−	−	−	+	+	−	+	0.009	0.009	0.000	-	0.000	-	0.000	-
14*	+	−	−	−	+	+	−	+	0.009	0.009	0.000	-	0.000	-	0.000	-
15*	+	−	+	−	+	+	−	+	0.000	-	0.000	-	0.000	-	0.018	0.018
16	+	−	−	−	−	+	+	+	0.224	0.040	0.136	0.042	0.162	0.045	0.263	0.058
17	−	+	−	+	+	+	+	+	0.056	0.022	0.091	0.035	0.059	0.029	0.053	0.030
18	−	+	+	−	+	+	+	+	0.037	0.018	0.030	0.021	0.044	0.025	0.070	0.034
19#	+	+	−	+	+	+	+	+	0.000	-	0.015	0.015	0.000	-	0.000	-
20#	−	+	+	−	+	+	+	+	0.000	-	0.000	-	0.000	-	0.053	0.030
21*	−	+	−	−	−	+	+	+	0.000	-	0.000	-	0.000	-	0.018	0.018
22	+	−	−	−	−	+	+	−	0.103	0.029	0.136	0.042	0.059	0.029	0.070	0.034
23	−	+	−	+	+	+	+	−	0.047	0.020	0.000	-	0.044	0.025	0.053	0.030
24#	−	+	+	−	+	+	+	−	0.019	0.013	0.061	0.029	0.118	0.039	0.018	0.018
25#	+	+	−	+	+	+	+	−	0.009	0.009	0.015	0.015	0.000	-	0.000	-
26*	−	+	−	−	−	+	+	−	0.000	-	0.015	0.015	0.000	-	0.000	-
27*	−	−	−	−	−	+	+	−	0.000	-	0.000	-	0.000	-	0.018	0.018
28[2]	+	−	−	−	−	+	−	−	0.000	-	0.030	0.021	0.000	-	0.000	-

[1] Unmarked haplotypes have been described previously, though 6, 7 and 23 were described on β^T chromosomes (Kazazian *et al.* 1984a).

* Denotes a haplotype not previously described in Asian Indians.

\# Denotes a haplotype which resembles a previously described haplotype at all sites except site 8, *Hin*dIII/pRK29, which was not determined in the previous studies (Old *et al.* 1984, Thein *et al.* 1984b, 1988, Varawalla *et al.* 1992).

[2] Sites 1, 3-9 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes.

[3] N = number of chromosomes on which haplotypes could be determined. In each group a number of haplotypes were indeterminate and were excluded from the analysis.

TABLE 5.2b Frequencies of β globin haplotypes on β^A chromosomes in the minor South African Indian groups

HAPLOTYPE									MOSLEM URDU		MOSLEM MEMON		HINDU TELEGU	
Type[1]	Description[2]								(N=10)[3]		(N=6)[3]		(N=6)[3]	
									Freq	SE	Freq	SE	Freq	SE
1	+	−	−	−	−	−	−	+	0.100	0.009	-	-	-	-
2	−	+	−	+	+	−	−	+	0.200	0.126	-	-	0.167	0.152
3	−	+	+	−	+	−	−	+	0.200	0.126	-	-	0.167	0.152
29*	−	−	−	−	−	−	−	+	-	-	0.167	0.152	-	-
9	−	+	+	−	+	+	−	+	-	-	-	-	0.167	0.152
16	+	−	−	−	−	+	+	+	0.200	0.126	0.167	0.152	0.167	0.152
17	−	+	−	+	+	+	+	+	0.100	0.009	0.167	0.152	0.167	0.152
18	−	+	+	−	+	+	+	+	-	-	0.167	0.152	-	-
19#	+	+	−	+	+	+	+	+	-	-	-	-	0.167	0.152
22	+	−	−	−	−	+	+	−	-	-	0.333	0.192	-	-
24#	−	+	+	−	+	+	+	−	0.100	0.009	-	-	-	-
27*	−	−	−	−	−	+	+	−	0.100	0.009	-	-	-	-

[1] Unmarked haplotypes have been described previously, though 6, 7 and 23 were described on β^T chromosomes (Kazazian *et al.* 1984a).

* Denotes a haplotype not previously described in Asian Indians.

\# Denotes a haplotype which resembles a previously described haplotype at all sites except site 8, *Hin*dIII/3'β, which was not determined in the previous studies (Old *et al.* 1984, Thein *et al.* 1984b, 1988, Varawalla *et al.* 1992).

[2] Sites 1, 3-9 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes.

[3] N = number of chromosomes on which haplotypes could be determined. In each group a number of haplotypes were indeterminate and were excluded from the analysis.

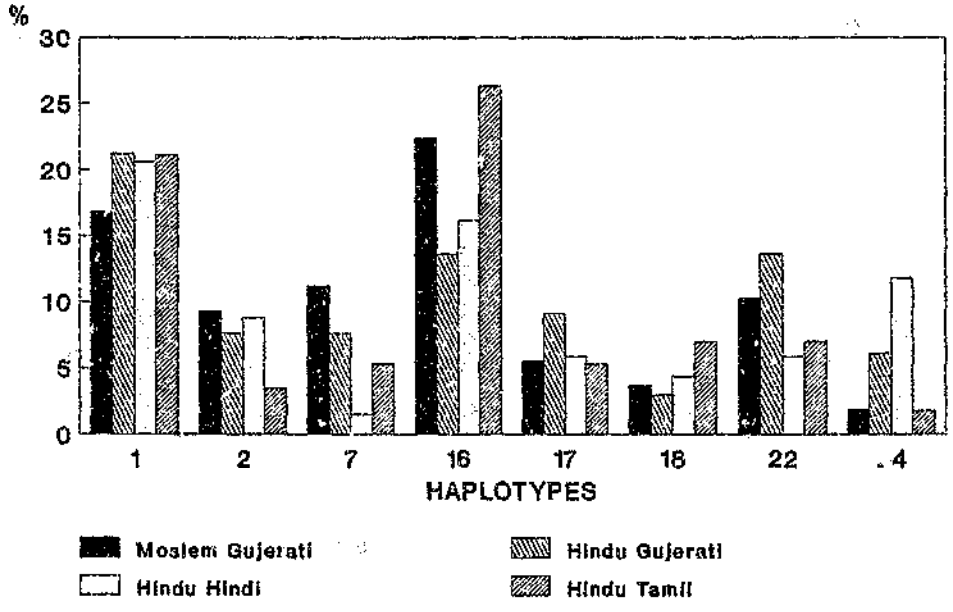


FIGURE 5.3 The common β^A haplotypes in the major South African Indian groups.

The haplotype numbers correspond to those in Table 5.2a.
Haplotypes 1 and 16 are the two commonest haplotypes in all the groups.
Haplotypes 2, 7, 17, 18, 22 or 24 are the third or fourth most common haplotypes in at least one of the major groups.

As there is a hotspot for recombination betw[illegible]e δ and ß genes (Antonarakis *et al.* 1982a, Chakravarti *et al.* 1984), the 5′ and 3′ parts of the haplotypes were also analysed separately. The distributions of the 5′ and 3′ haplotypes are shown in Tables 5.3 and 5.4, respectively. Eleven of the 32 possible 5′ haplotypes were observed on $ß^A$ chromosomes with three accounting for between 87.6 and 92.5% of all chromosomes in the major groups. There were no statistically significant differences between the major groups. The (+---) haplotype was the commonest in all the major groups, accounting for at least 41% of chromosomes. In the Moslem Gujerati (-+-++) was next, followed by (-++-+), with frequencies of 21.4% and 8.4% respectively. In the other groups, however, the latter two types had almost equal frequencies of between 14 and 24%. Two of the rarer 5′ haplotypes have not been described previously in Asian Indians (Kazazian *et al.* 1984a, Old *et al.* 1984, Thein *et al.* 1988, Varawalla *et al.* 1992, Flint *et al.* 1993a,b).

Five of the eight possible combinations of 3′ haplotypes were observed and all three frameworks were represented, as shown in Table 5.4. In all the groups haplotypes (--+) and (+++) were the two most common. They occur at almost equal frequencies in the Moslem Gujerati, while in the Hindu Gujerati and Hindu Hindi (--+) was more common and in the Hindu Tamil (+++) was more common. The differences between the major groups are not statistically significant, however.

Genetic diversities (Nei 1987) or average heterozygosities for the haplotype distributions on $ß^A$ chromosomes in the major groups were:

Moslem Gujerati	0.887
Hindu Gujerati	0.902
Hindu Hindi	0.901
Hindu Tamil	0.875

5.1.3.2 $ß^T$ haplotypes

A total of 58 $ß^T$ haplotypes could be determined unequivocally. The $ß^T$ haplotypes and their frequencies are shown in Table 5.5 for the major and minor South African Indian groups. Of the 11 haplotypes observed, five have been described previously on $ß^T$

TABLE 5.3 Distribution of 5′ β haplotypes on β^A chromosomes in South African Indians

5′ HAPLOTYPE[1,2]					MOSLEM GUJERATI (N=107)[3]		HINDU GUJERATI (N=66)[3]		HINDU HINDI (N=68)[3]		HINDU TAMIL (N=57)[3]		MOSLEM URDU (N=10)[3]		MOSLEM MEMON (N=6)[3]		HINDU TELEGU (N=6)[3]	
					N	%	N	%	N	%	N	%	N	%	N	%	N	%
+	−	−	−	−	65	60.7	39	59.1	30	44.1	34	59.6	3	30.0	3	50.0	1	16.7
−	+	−	+	+	23	21.5	11	16.7	15	22.1	8	14.0	3	30.0	1	16.7	2	33.3
−	+	+	−	+	9	8.4	11	16.7	16	23.5	8	14.0	3	30.0	1	16.7	2	33.3
+	+	−	+	+	7	6.5	2	3.0	5	7.4	1	1.8	-	-	-	-	1	16.7
−	+	+	−	−	1	0.9	1	1.5	1	1.5	3	5.3	-	-	-	-	-	-
−	+	−	−	+	-	-	1	1.5	1	1.5	-	-	-	-	-	-	-	-
−	−	−	−	+*	1	0.9	-	-	-	-	-	-	-	-	-	-	-	-
+	−	−	−	+*	1	0.9	-	-	-	-	-	-	-	-	-	-	-	-
+	−	+	−	+	-	-	-	-	-	-	1	1.8	-	-	-	-	-	-
−	+	−	−	−	-	-	1	1.5	-	-	1	1.8	-	-	-	-	-	-
−	−	−	−	−	-	-	-	-	-	-	1	1.8	1	10.0	1	16.7	-	-

[1] Sites 1, 3-6 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes.

[2] * Denotes a 5′ haplotype not previously described in Asian Indians (Kazazian *et al.* 1984a, Old *et al.* 1984, Thein *et al.* 1988, Varawalla *et al.* 1992).

[3] N = number of chromosomes on which haplotypes could be determined. In each group a number of haplotypes were indeterminate and were excluded from the analysis.

TABLE 5.4 Distribution of 3′ β haplotypes on β^A chromosomes in South African Indians

3′ HAPLOTYPE[1,2]			MOSLEM GUJERATI (N=107)[3]		HINDU GUJERATI (N=66)[3]		HINDU HINDI (N=68)[3]		HINDU TAMIL (N=57)[3]		MOSLEM URDU (N=10)[3]		MOSLEM MEMON (N=6)[3]		HINDU TELEGU (N=6)[3]	
			N	%	N	%	N	%	N	%	N	%	N	%	N	%
−	−	+[4]	35	32.7	24	36.4	31	45.6	18	31.6	5	50.0	1	16.7	2	33.3
+	−	+[5]	19	17.8	7	10.6	4	5.9	4	7.0	-	-	-	-	1	16.7
+	+	+[5]	34	31.8	18	27.3	18	26.5	26	45.6	3	30.0	3	50.0	3	50.0
+	+	−[6]	19	17.8	15	22.7	15	22.1	9	15.8	2	20.0	2	33.3	-	-
+	−	−[*6]	-	-	2	3.0	-	-	-	-	-	-	-	-	-	-

[1] Sites 7-9 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes.

[2] * Denotes a 3′ haplotype not previously described in Asian Indians (Kazazian *et al.* 1984a).

[3] N = number of chromosomes on which haplotypes could be determined. In each group a number of haplotypes were indeterminate and were excluded from the analysis.

[4] Framework 3

[5] Framework 1

[6] Framework 2

TABLE 5.5 Frequencies of β globin haplotypes on β^T chromosomes in the major and minor South African Indian groups

HAPLOTYPE									MOSLEM GUJERATI		HINDU GUJERATI		HINDU HINDI		HINDU TAMIL		MOSLEM URDU		MOSLEM MEMON		HINDU TELEGU	
Type[1]	Description[2]								(N=22)[3]		(N=8)[3]		(N=10)[3]		(N=6)[3]		(N=6)[3]		(N=5)[3]		(N=1)[3]	
									Freq	SE	Freq	SE	Freq	SE	Freq	SE	Freq	SE	Freq	SE	Freq	SE
1	+	-	-	-	-	-	-	+	0.591	0.105	0.125	0.117	0.900	0.095	0.500	0.200	0.667	0.192	0.400	0.219	1.000	-
2	-	+	-	+	+	-	-	+	0.045	0.044	0.375	0.171	-	-	0.333	0.037	-	-	-	-	-	-
3	-	+	+	-	+	-	-	+	0.091	0.061	-	-	-	-	-	-	0.167	0.152	-	-	-	-
8	-	+	-	+	+	+	-	+	0.136	0.073	-	-	-	-	-	-	0.167	0.152	0.600	0.219	-	-
9	-	+	+	-	+	+	-	+	-	-	-	-	-	-	0.167	0.139	-	-	-	-	-	-
16	+	-	-	-	-	+	+	+	0.045	0.044	-	-	-	-	-	-	-	-	-	-	-	-
17	-	+	-	+	+	+	+	+	0.045	0.044	-	-	-	-	-	-	-	-	-	-	-	-
18	-	+	+	-	+	+	+	+	0.045	0.044	-	-	-	-	-	-	-	-	-	-	-	-
22	+	-	-	-	-	+	+	-	-	-	0.250	0.153	0.100	0.090	-	-	-	-	-	-	-	-
28#	+	-	-	-	-	+	-	-	-	-	0.125	0.117	-	-	-	-	-	-	-	-	-	-
30#	-	+	+	-	+	+	-	-	-	-	0.125	0.117	-	-	-	-	-	-	-	-	-	-

[1] Unmarked haplotypes have been described previously, though 2, 9, 16 and 17 were described on β^N chromosomes (Kazazian *et al.* 1984a).
# Denotes a haplotype which resembles a previously described haplotype at all sites except site 8, *Hind*III/3'β, which was not determined in the previous studies (Old *et al.* 1984, Thein *et al.* 1988, Varawalla *et al.* 1992).

[2] Sites 1, 3-9 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes.

[3] N = number of chromosomes on which haplotypes could be determined. In some groups haplotypes were indeterminate and were excluded from the analysis.

chromosomes and a further four on ß^A chromosomes (numbers 2, 9, 16 and 17) in Asian Indians (Kazazian *et al.* 1984a). For the latter four haplotypes, ß^T haplotypes have been described which were the same at all sites studied (sites 1, 3-7, 9 in Figure 2.2) (Old *et al.* 1984, Thein *et al.* 1988, Varawalla *et al.* 1992). The *Hin*dIII/3′ß site (site 8) was again not studied and may differ. Haplotype 28 was only found in the Hindu Gujerati on ß^A and ß^T chromosomes. Haplotype 30 was not found on ß^A chromosomes in this study or that of Kazazian *et al.* (1984a), but a similar haplotype without the *Hin*dIII/3′ß site was described on ß^A and ß^T chromosomes (Varawalla *et al.* 1992). Four haplotypes observed by Varawalla *et al.* (1992) on ß^T chromosomes at frequencies ≤1% were not found in this study.

The haplotype frequencies on ß^T chromosomes were compared between the major groups. Haplotype 1 appears to be the most common in all groups except the Hindu Gujerati, although the numbers of chromosomes studied were small. The differences were not statistically significant. The associations of the ß^T haplotypes with ß thalassaemia mutations are presented and discussed in Chapter 6.

The 5′ and 3′ ß^T haplotypes were also analysed separately. The distributions of the 5′ and 3′ haplotypes are shown in Tables 5.6 and 5.7 respectively. Three of the 11 5′ ß globin haplotypes observed on ß^A chromosomes were observed on ß^T chromosomes. The same three haplotypes were most common on ß^A and ß^T chromosomes. Further, the distributions of the 5′ haplotypes on ß^T chromosomes do not differ significantly between the major groups. In the Hindu Hindi only one 5′ haplotype was observed.

As on the ß^A chromosomes, five of the eight possible 3′ haplotypes were observed, representing the three frameworks. In all the major groups the (--+) (Fw3) haplotype appears to be most common. In the Moslem Gujerati, the rest of the chromosomes either had the (+-+) or (+++) haplotype, both Fw1, the former also occurring in the Hindu Tamil. In the Hindu Gujerati haplotypes (++-) and (+--), both Fw2, occur, while in the Hindu Hindi the former was present. In each group only two of the three frameworks were observed on ß^T chromosomes. The distributions between the groups were not significantly different.

TABLE 5.6 Distribution of 5′ β haplotypes on β^T chromosomes in South African Indians

5′ HAPLOTYPE[1]	MOSLEM GUJERATI (N=22)[2]		HINDU GUJERATI (N=8)[2]		HINDU HINDI (N=10)[2]		HINDU TAMIL (N=6)[2]		MOSLEM URDU (N=6)[2]		MOSLEM MEMON (N=5)[2]		HINDU TELEGU (N=1)[2]	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
+ − − − −	14	63.6	4	50.0	10	100.0	3	50.0	4	66.7	2	40.0	1	100.0
− + − + +	5	22.7	3	37.5		-	2	33.3	1	16.7	3	60.0	-	-
− + + − +	3	13.6	1	12.5	-	-	1	16.7	1	16.7	-	-	-	-

[1] Sites 1, 3-6 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes.
[2] N = number of chromosomes on which haplotypes could be determined. In some groups a number of haplotypes were indeterminate and were excluded from the analysis.

TABLE 5.7 Distribution of 3′ β haplotypes on β^T chromosomes in South African Indians

3′ HAPLOTYPE[1,2]			MOSLEM GUJERATI (N=22)[3]		HINDU GUJERATI (N=8)[3]		HINDU HINDI (N=10)[3]		HINDU TAMIL (N=6)[3]		MOSLEM URDU (N=6)[3]		MOSLEM MEMON (N=5)[3]		HINDU TELEGU (N=1)[3]	
			N	%	N	%	N	%	N	%	N	%	N	%	N	%
−	−	+[4]	16	72.7	4	50.0	9	90.0	5	83.3	5	83.3	2	40.0	1	100.0
+	−	+[5]	3	13.6	-	-	-	-	1	16.7	1	16.7	3	60.0	-	-
+	+	+[5]	3	13.6	-	-	-	-	-	-	-	-	-	-	-	-
+	+	−[6]	-	-	2	25.0	1	10.0	-	-	-	-	-	-	-	-
+	−	−[6]	-	-	2	25.0	-	-	-	-	-	-	-	-	-	-

[1] Sites 7-9 in shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes.

[2] * Denotes a 3′ haplotype not previously described in Asian Indians (Kazazian *et al.* 1984a).

[3] N = number of chromosomes on which haplotypes could be determined. In some groups a number of haplotypes were indeterminate and were excluded from the analysis.

[4] Framework 3

[5] Framework 1

[6] Framework 2

5.1.3.3 Comparison of β^A and β^T haplotypes

The β^T genes appear on a more limited subset of haplotypes than do the β^A genes (Tables 5.2a and 5.5), and the frequency distributions of the haplotypes on β^T chromosomes appear to differ from those on β^A chromosomes. In the Moslem Gujerati haplotypes 1 and 8 represent 72.7% of the β^T chromosomes, while only forming 18.7% of β^A chromosomes. Haplotype 16 which represents 22.4% of β^A chromosomes only occurs on 4.5% of β^T chromosomes. In the Hindu Gujerati haplotypes 2 and 22 were commonest on the β^T chromosomes studied. These types account for 62.5% of β^T chromosomes, but only 21.2% of β^A chromosomes. In the Hindu Hindi haplotype 1 was commonest on both chromosome types. In the Hindu Tamil haplotype 1 represents 50.0% of β^T chromosomes, but 21.1% of β^A chromosomes. Haplotype 16 which occurs on 26.3% of β^A chromosomes was not represented on the β^T chromosomes studied. The differences between the haplotype distributions reach significance in the Moslem Gujerati ($p<0.01$), but not in the other groups. In the minor groups differences were also apparent, but the numbers were too small for any comparisons. The distributions between β^A and β^T chromosomes in the pooled sample were statistically significant ($p<0.001$).

When the 5′ and 3′ haplotypes were compared separately, no significant differences were observed in the 5′ haplotype between β^A and β^T chromosomes. The 3′ haplotype distributions, however, differed significantly in the Moslem Gujerati ($p< 0.01$).

5.1.3.4 β^S haplotypes

Eleven β^S chromosomes were studied: two Moslem Gujerati, three Hindu Gujerati, three Hindu Hindi, one Hindu Tamil, one Hindu Telegu and one Moslem Urdu. All were shown to be haplotype 25 (++-+ + ++-).

5.1.3.5 β^E haplotypes

The β^E gene was associated with haplotype 23 (-+-++ ++-) in one Moslem Gujerati individual and haplotype 22 (+--- ++-) in one Hindu Hindi.

5.1.3.6 β^D haplotypes

The β^D gene occurs in association with haplotype 16 (+---- +++) on the two chromosomes studied, from a Moslem Gujerati and a Moslem Memon individual, respectively.

5.1.3.7 Haplotypes associated with γ globin rearrangements

Haplotypes could be determined on two of the γγγ chromosomes. In the Moslem Gujerati individual the haplotype was (+[±±±]++ ±±+) and in the Hindu Hindi (+[±±±]++ --+). The haplotype on the -γ chromosome was (-[-]±± --+).

5.1.3.8 Haplotypes associated with the *Xmn*I variant

In the Moslem Gujerati family, the *Xmn*I variant was associated with haplotype 4 (++-++ --+) on a β^A chromosome, whereas in the two Hindu Gujerati families it was associated with haplotype 22 (+---- ++-) on a β^T chromosome.

5.1.4 Correlation of ß globin haplotypes and the *Xmn*I polymorphic site

The association of the presence or absence of the *Xmn*I site with the different haplotypes has been studied on 316 β^A chromosomes, 58 β^T chromosomes, 11 β^S chromosomes, 2 β^D chromosomes and 2 β^E chromosomes. Since the *Xmn*I site is situated in the 5′ part of the ß globin cluster (site 2 in Figure 2.2), the haplotypes were grouped according to their 5′ haplotypes for analysis. The results are shown in Table 5.8.

There was a strong, but not absolute, correlation between the presence or absence of the *Xmn*I site and the 5′ haplotype. Thus, for example, 174/176 or 98.9% of the β^A chromosomes and 100% of the β^T chromosomes with a (+----) 5′ haplotype lacked the polymorphic *Xmn*I site. The β^E chromosome in this group represents an interesting exception as it was associated with the presence of the *Xmn*I site. Similarly for the 5′ haplotype (-+-++), 93.4% of β^A chromosomes and 100% of β^T chromosomes had the polymorphic *Xmn*I site. The two β^D chromosomes again represent an interesting exception

TABLE 5.8 Association of β globin haplotypes with the *Xmn*I polymorphism

HAPLOTYPE						β^S CHROMOSOMES		β^T, β^S, β^D AND β^E CHROMOSOMES[3]	
5' HAPLOTYPE[1]					NUMBER[2]	-	+	-	+
+	-	-	-	-	1	60	-	32	-
					7	20	1	-	-
					16	61	1	2	-
					22	31	-	3	1(E)
					28	2	-	1	-
-	+	-	+	+	2	1	24	2(D)	6
					8	-	4	-	7
					17		19	-	1
					23	1	10	-	1(E)
-	+	+	-	+	3	16		2	1
					9	3	-	1	-
					18	14	-	1	-
					24	16	-	-	-
					30	-	-	1	-
+	+	-	+	+	4	-	10	-	-
					10	-	2	-	-
					19	-	2	-	-
					25	1	1	-	11(S)
-	+	+	-	-	5	1	-	-	-
					11	1	-	-	-
					20	3	-	-	-
-	+	-	-	-	21	-	1	-	-
					26	-	1	-	-
-	+	-	-	+	6	1	-	-	-
					12	1	-	-	-
-	-	-	-	-	27	2	-	-	-
					29	1	-	-	-
-	-	-	-	+	13	1	-	-	-
+	-	-	-	+	14	1	-	-	-
+	-	+	-	+	15	1	-	-	-

[1] Sites 1, 3-6 shown in Figure 2.2 and described in Table 2.5 were used to construct the 5' haplotype.

[2] The numbers of the haplotypes correspond to those in Tables 5.2a,b and 5.5.

[3] All unmarked numbers represent β^T chromosomes.

as they did not have the site. The majority of β^T chromosomes lack the site. Of the 26 thalassaemia major individuals studied, 14 were (-/-), nine (-/+) and three (+/+) for the *Xmn*I site.

All the β^S chromosomes were associated with haplotype 25 and the presence of the *Xmn*I site. Similarly both β^D chromosomes were associated with haplotype 2 and the absence of the *Xmn*I site. Both β^E chromosomes had the *Xmn*I site, though they were associated with different haplotypes (22 and 23), which differ at their 5′ end.

The *Xmn*I sites have also been determined on the chromosomes with the triple rearrangements. The $^{G}\gamma^{G}\gamma^{A}\gamma$ chromosomes all lack the *Xmn*I site 5′ to both $^{G}\gamma$ genes.

5.2 Discussion

5.2.1 ß globin cluster rearrangements

The most common rearrangements in the ß globin cluster involve the two γ genes. As a result of unequal, but homologous crossover, chromosomes with one and three γ genes occur, the single or middle γ genes being hybrids $^{G}\gamma^{A}\gamma$ or $^{A}\gamma^{G}\gamma$, respectively (Trent *et al.* 1981a, Sukumaran *et al.* 1983, Hill *et al.* 1986, Shimasaki and Iuchi 1986). Individuals with four or five γ genes also occur (Harano *et al.* 1985a, Hill *et al.* 1986, Fei *et al.* 1988b). Chromosomes with two $^{G}\gamma$ or two $^{A}\gamma$ genes only have been reported as well (Powers *et al.* 1984). The γ gene rearrangements are thought to be selectively neutral, so that differences between populations represent genetic drift (Hill *et al.* 1986).

The triple γ rearrangement has been detected in numerous populations at low frequencies (reviewed in Huisman 1987). It occurs at polymorphic frequencies in the Japanese (Harano *et al.* 1985a), in Melanesia and Polynesia (Flint *et al.* 1986, Hill *et al.* 1986, Trent *et al.* 1986, Hill *et al.* 1987a), Fiji (Trent *et al.* 1988) and in Micronesia and south-east Asia (O'Shaughnessy *et al.* 1990). The populations of the Cook Islands and western Samoa in Polynesia have the highest reported frequencies of 16% and 14% respectively (Trent *et al.* 1986).

The frequency appears to be low in all the major South African Indian groups studied, with the possible exception of the Hindu Hindi. In the latter group, the triple γ rearrangement was found on 2.2% of chromosomes (although the standard error was large). This rearrangement may represent another example of a founder effect in this group, which originated from a limited number of districts in India (Bhana 1991) and has been shown to have a number of unusual features including a high frequency of the $-\alpha^{4.2}$ and $\alpha\alpha\alpha$ haplotypes (discussed in Sections 4.2.1 and 4.2.4). The γγγ rearrangement had been previously reported in an Indian subject from the Punjab (Thein *et al.* 1984a) and in the tribal populations of Andhra Pradesh (Fodde *et al.* 1988), though no frequencies for the rearrangement were estimated.

All the γγγ rearrangements in the present study were of the $^{G}\gamma^{G}\gamma^{A}\gamma$ type, as are the majority of γγγ chromosomes, though considerable heterogeneity has since been shown in the location of the crossover (Huisman 1987, Liu *et al.* 1988). At least four independent unequal crossover events have occurred (Liu *et al.* 1988).

In the first γγγ rearrangement, described in Vanuatu in Melanesia, the expression of the γ genes was shown to be affected only marginally (Trent *et al.* 1981a). Further studies have shown that heterozygotes for the γγγ rearrangement have slightly higher mean HbF levels than normal individuals, although individuals may have HbF levels in the normal range (Thein *et al.* 1984a, Hill *et al.* 1986). Individuals homozygous for the γγγ rearrangement have also been shown to have normal haematological indices and HbF levels. The additional γ genes do, however, appear to be expressed, as reflected in an increased $^{G}\gamma:^{A}\gamma$ ratio (Hill *et al.* 1986). In a Turkish family heterozygous for a $^{G}\gamma^{G}\gamma^{G}\gamma^{A}\gamma$ rearrangement, the $^{G}\gamma:^{A}\gamma$ ratio was 95:5 (Yang *et al.* 1986). More recent studies have shown four categories of triplication with different HbF levels and $^{G}\gamma:^{A}\gamma$ ratios, depending on the presence or absence of *Xmn*I restriction sites (Liu *et al.* 1988). These are discussed further in Section 5.2.4.

The -γ rearrangement similarly has been reported in a number of populations. It is found at frequencies of 2-12% in Melanesia, but occurs in less than 1% of Polynesians and south-east Asians (Hill *et al.* 1986, Trent *et al.* 1986, Hill *et al.* 1987a, O'Shaughnessy *et al.* 1990), and was not found in Micronesia (O'Shaughnessy *et al.* 1990). It has also

been reported to occur in Japan (Shimasaki and Iuchi 1986), Fiji and Papua New Guinea (Trent *et al.* 1988), India and China (Sukumaran *et al.* 1983).

The -γ rearrangement was only found in one Hindu Tamil individual, suggesting that the frequency of this rearrangement is probably relatively low in all the major Indian groups studied. This is consistent with previous observations where γ chain composition studies were used to calculate a frequency of 0.012 in Indians (Huisman *et al.* 1983). Sukumaran *et al* (1983) reported an Indian newborn homozygous for this rearrangement. The child was, however, the product of a consanguineous marriage. The rearrangement has also been described in tribal Indians of Andhra Pradesh (Fodde *et al.* 1988).

The remaining γ gene is a hybrid $^{G}\gamma^{A}\gamma$ gene, whose product is an $^{A}\gamma$ chain since the 3′ end includes the codon for alanine at position 136. It is, however, synthesised at birth at $^{G}\gamma$ levels because of regulatory signals originating 5′ to $^{G}\gamma$. The most likely location for the crossover is thus within the transcribed portion of the γ genes 5′ to codon 136 (Huisman *et al.* 1983, Sukumaran *et al.* 1983).

Heterozygosity for the -γ rearrangement does not appear to alter the HbF levels markedly, though heterozygotes do have slightly lower mean HbF levels than normal individuals. In addition, a low $^{G}\gamma$:$^{A}\gamma$ ratio is observed in newborns and adults with the condition. The homozygotes, however, appear to have low HbF levels at birth, with only $^{A}\gamma$ chains, and a rapid decrease in HbF levels during the first few months of life. No overt disease or anaemia appeared to be associated with the condition. Two adult homozygotes studied had HbF levels in the normal range (Huisman *et al.* 1983, Sukumaran *et al.* 1983, Hill *et al.* 1986).

The absence of the $^{G}\gamma^{G}\gamma$ and $^{A}\gamma^{A}\gamma$ rearrangements in the South African Indian sample is not unexpected, since these rearrangements have been reported at low frequency in most populations studied (Powers *et al.* 1984, Harano *et al.* 1985b, Hattori *et al.* 1986a).

5.2.2 ß globin RFLPs

At least 18 polymorphic sites have been described in the ß globin cluster (reviewed in

Boehm and Kazazian 1989), but only nine of these sites have been studied in the South African Asian Indians.

5.2.2.1 RFLP frequencies on ß^A chromosomes

When the frequencies of the nine RFLP sites in the ß globin cluster were compared, the Hindu Hindi appear to differ most from the other groups. They differ significantly from the Moslem Gujerati at four sites, from the Hindu Tamil at three and from the Hindu Gujerati at one. If the frequencies in the different groups are arranged in ascending or descending order, the Hindu Hindi are at one extreme in eight of the nine systems studied. Thus, again the Hindu Hindi had features which appeared to distinguish them from all the other groups. In contrast, none of the other groups had significantly different frequencies from each other at any of the nine sites, and it is thus difficult to determine which are more closely related. The numbers of chromosomes studied in the minor groups were too small for meaningful statistical analysis.

The frequencies in the South African Indians appear to differ from those previously published in an Asian Indian sample of undefined geographical origin, studied in the United States (Orkin *et al.* 1984). The Hindu Hindi frequencies again appear to differ the most. Unfortunately insufficient information is provided in the published data for statistical comparison.

Although the frequencies observed in South African Indians differ from those previously published, when the RFLP site frequencies were compared with a number of other populations (reviewed in Chen *et al.* 1990), the South African Asian Indians had frequencies consistent with the observed trends for the sites studied. For example, site 1 (*Hin*cII/ε) occurs most commonly in Polynesians and Micronesians at frequencies of 0.89-0.99 (Chen *et al.* 1990), while Africans have the lowest frequencies (0.00-0.23) (Ramsay and Jenkins 1987, Long *et al.* 1990). Frequencies in other Asian Pacific populations are 0.54-0.79 (Chen *et al.* 1990), consistent with those observed here of 0.53-0.69 in the major groups.

Most Asian Indian groups studied in other studies have originated from the north-western

and western regions of the subcontinent, as these have been the groups who have emigrated to the United Kingdom (Varawalla *et al.* 1992) or the United States (Kazazian *et al.* 1984a). In the present study, most groups originate from the western, central or southern regions of the subcontinent (see Figure 1.4b). Frequencies of the different polymorphisms may show clines from west to east across the subcontinent and the different frequencies in the Hindu Hindi may, in part, reflect their different geographical origins from the central, more isolated, part of the subcontinent. Unfortunately, insufficient data were available from the different regions of the subcontinent to substantiate this observation.

5.2.2.2 RFLP frequencies on β^T chromosomes

A number of significant differences in allele frequencies on the β^T chromosomes were noted between the different groups, though there were no obvious trends. These differences reflect the presence of different ß thalassaemia mutations in the major groups occurring on different chromosomal backgrounds. These are discussed further in Section 6.2.2.

Significant differences in allele frequencies between β^A and β^T chromosomes occur in a number of the groups at some of the sites. As with the -α chromosomes, chromosomes carrying ß thalassaemia mutations are likely to have undergone positive selection. The allele frequencies reflect the original subset of parent chromosomes on which the mutations arose. Differences do not occur at all sites or in all the Indian groups, partly because relatively small numbers of β^T chromosomes were studied in most of the groups and thus only large differences between the frequencies reach statistical significance. Further, similar frequencies may occur at some sites by chance.

If differences in allele frequencies occur between β^A and β^T chromosomes, the polymorphic site is likely to be useful for prenatal diagnosis using linked marker analysis, as this type of analysis relies on distinguishing the two chromosomes. Thus, one might expect the *Ava*II/ß (site 7) and the *Hin*dIII/3'ß (site 8) RFLPs to be most useful for prenatal diagnosis using linked marker analysis in the South African Asian Indians as these differ significantly in most of the groups. In the analysis of 31 families with ß

thalassaemia, the *Ava*II/ß system was most useful, but still only fully informative in four families and semi-informative in 22. The *Hin*dIII/3'ß site was fully informative in five and semi-informative in 20. The assessment of the usefulness of the individual linked markers results are presented and discussed in Section 7.2.1.1.

5.2.2.3 A new *Xmn*I variant

A new variant due to the creation of an additional *Xmn*I site 5' to the $^{A}\gamma$ gene was described. The fragment sizes produced are consistent with a mutation occurring at -158 to the $^{A}\gamma$ gene, the homologous position to the polymorphic site at position -158 to the $^{G}\gamma$ gene. Further it is proposed that a similar C→T transition must occur to produce an *Xmn*I site. A patient with low HbF $^{G}\gamma\beta^{+}$ HPFH has been described in which an *Xmn*I site was found 5' to both γ genes. However, further analysis in this patient showed that the chromosome carried the $^{G}\gamma^{G}\gamma$ rearrangement (Gilman and Huisman 1985). In the individuals in this study, the chromosomes were of the $^{G}\gamma^{A}\gamma$ type, thus suggesting separate origins for the mutations. In addition, the variant in this study appears to occur on two different chromosomal backgrounds. The chromosomes differ both 5', at the *Xmn*I site at -158 to the $^{G}\gamma$ gene, and 3', within the ß globin sequence, suggesting that the mutation has occurred at least twice in South African Indians.

The mechanism for the mutation has not been determined and may be due to a point mutation or a gene conversion event. Gene conversion as a mechanism of maintaining identity or generating diversity between the two γ genes has been implicated in a number of studies (Jeffreys 1979, Slightom *et al.* 1980, Starck *et al.* 1990, Lanclos *et al.* 1991).

5.2.3 ß globin haplotypes

The ß globin haplotypes in this study were characterised using eight dimorphic RFLP systems. Unfortunately, not all studies characterising the ß globin haplotypes have used the same sites, thus making comparisons difficult in some cases.

As with the α globin haplotypes, marked linkage disequilibrium exists between the polymorphic sites (Antonarakis *et al.* 1982a), so that only a limited number of the

possible haplotypes have been observed. This non-random association of the sites limits their combined usefulness for prenatal diagnosis, so that little additional information is obtained by increasing the number of DNA polymorphisms studied (Antonarakis *et al.* 1982a).

5.2.3.1 ß^A haplotypes

In the present study, 29 haplotypes were observed on 320 ß^A chromosomes, with four common haplotypes accounting for between 57 and 61.5% of the chromosomes in each major group. The heterozygosity rates are, however, still high, $\geq$0.875 in all groups. At least nine haplotypes not previously described in Asian Indians are reported. All occur at relatively low frequency, with the exception of haplotype 4, which occurs at frequencies of 4.7% and 5.9% in the Moslem Gujerati and Hindu Hindi respectively. As this is the largest sample of Indian ß^A chromosomes studied, one might expect to find additional rare haplotypes. Varawalla *et al.* (1992) studied 196 normal Asian Indian chromosomes from north-west Pakistan, Gujerat, Punjab and Sindh and described 19 different haplotypes, five of which were not observed in the South African Indian sample. All occur at low frequency and thus their absence may be due to sampling error or the different geographical origins of the two samples. Thus, at least 34 different ß^A haplotypes have been described in Asian Indians.

In the South African Asian Indians, haplotypes 1 (+---- --+) and 16 (+---- +++) were commonest in all groups, though the next two most common haplotypes differ. None of the groups differs significantly, however, in overall haplotype distribution. Haplotype 16 was the commonest in the Moslem Gujerati and Hindu Tamil, while haplotype 1 was commonest in the Hindu Gujerati and Hindu Hindi. In a study of Asian Indians and Pakistanis in the United States 16 followed by 1 were also the two most common haplotypes (Kazazian *et al.* 1984a). Similarly, a study of Asian Indians in the United Kingdom was consistent with 1 and 16 being the commonest two haplotypes, accounting for between 40 and 55% of chromosomes (Old *et al.* 1984, Varawalla *et al.* 1992), although the *Hin*dIII/3'ß site was not studied and thus haplotypes 7 and 16 could not be distinguished. In individuals from north-western Pakistan and Sindh, haplotype 16 was more common, while haplotype 1 was more common in Gujerat and the Punjab. In

individuals from Sindh, however, haplotype 1 occurs in only 9% of individuals (Varawalla *et al.* 1992). Thus, haplotypes 1 and 16 appear to be the two commonest on virtually the entire subcontinent, together accounting for at least 35%, and sometimes up to 55%, of chromosomes.

Haplotype 1 was also the commonest in a number of other Asian populations studied, including the Japanese (Shimizu [illegible], Chinese (Chan *et al.* 1986b) and Cambodians (Antonarakis *et al.* 1987b). In Thailand, however, haplotype 22 occurs most commonly (Hundrieser *et al.* 1988a). Haplotype 16 was also the commonest haplotype in the Mediterranean (Cypriots) (Old *et al.* 1984). Thus there appears to be an increase in frequency of haplotype 1 from west to east, with a corresponding decrease in haplotype 16, though individual groups may vary because of factors such as genetic drift.

In general, in all the groups studied in the present and previous studies (Kazazian *et al.* 1984a, Old *et al.* 1984, Varawalla *et al.* 1992), another two or three haplotypes of haplotypes 2, 17 (or 8), 18 (or 9) and 22 make up another 20-30% of chromosomes. The interesting exception was again the Hindu Hindi who had haplotype 24 on 10% of chromosomes, though the latter occurs at frequencies of 2-6% in the other groups. No geographical trends in the haplotype distributions were obvious from the available data, but perhaps further studies on the Indian subcontinent would reveal these. Other factors apart from geographical location, including genetic drift, may be important in determining the haplotype frequencies in individual groups. Many groups have remained isolated and factors such as inbreeding and strict endogamy of the caste system may have altered the frequencies and perpetuated local genetic diversity. For example, a group of individuals from Gujerat were studied in the United Kingdom (Varawalla *et al.* 1992) and in the present study two major groups originate from the province of Gujerat. The United Kingdom Gujerati have haplotypes 17 (or 8), 18 (or 9) and 2 as the next most common after haplotypes 1 and 16, while in the South African Moslem Gujerati haplotypes 7 and 22 were the next most common, and in the Hindu Gujerati it was haplotypes 17 and 22. The groups thus had some shared features, but also some differences, though all originate from roughly similar geographical areas.

The ß globin cluster consists of 5′ and 3′ regions of non-random association, separated

by 9.1kb of DNA 5′ to ß which is a hotspot for meiotic recombination and in which there is a 3-30 times increased recombination rate (Antonarakis *et al.* 1982a, Chakravarti *et al.* 1984). The 5′ and 3′ haplotypes have thus been analysed separately.

Only 11 of the 32 possible 5′ haplotypes were observed on ß^ chromosomes, thus confirming the non-random association of the RFLP sites observed previously (Antonarakis *et al.* 1982a, Wainscoat *et al.* 1986a). This is thought to be due to selective pressure for DNA sequences within these regions and different frequencies of recombination within and between these regions (Antonarakis *et al.* 1982a). The (+----) 5′ haplotype was the commonest in all the major groups of South African Asian Indians, as it was in all other Asian Indians (Kazazian *et al.* 1984a, Old *et al.* 1984, Varawalla *et al.* 1992), accounting for at least 44%, and up to 66% of chromosomes. This is not unexpected since the two commonest haplotypes, 1 and 16, share this 5′ haplotype. Three 5′ haplotypes, (+----), (-+-++) and (-++-+) account for approximately 90% of all chromosomes in the groups of South African Asian Indians studied here as well as in previous Indian samples (Antonarakis *et al.* 1982a, Kazazian *et al.* 1984a, Old *et al.* 1984, Varawalla *et al.* 1992).

The relative contributions of the three haplotypes differ, however. The (-+-++) was the second most common 5′ haplotype in the Moslem Gujerati as it was in samples from Gujerat, Punjab and Sindh (Varawalla *et al.* 1992). The Hindu Gujerati, however, also from the province of Gujerat, had equal frequencies of the 5′ haplotypes (-+-++) and (-++-+), as do the Hindu Hindi and the Hindu Tamil. In north-west Pakistan, the (-++-+) was the second most common (Varawalla *et al.* 1992). The differences between the groups do not appear to correlate well with their geographical origins and factors such as genetic drift in the isolated groups may be important.

The same three 5′ haplotypes account for over 90% of chromosomes in Europe, Asia and Oceania (Antonarakis *et al.* 1982a, reviewed by Flint *et al.* 1993b). Thus three 5′ haplotypes are common in most non-African populations, with (+----) being the most common globally, followed by (-+-++). In Africa, a fourth haplotype (----+) predominates, and a fifth, (-+--+), was also relatively common, the latter two dividing African and other populations (Wainscoat *et al.* 1986a). The 5′ haplotypes are thought

to be relatively stable over time, with four ancestral haplotypes identified, none of which can be derived from any other by a single point mutation or crossover event, predating racial divergence. Almost all other 5′ haplotypes can be derived from the four ancestral types by single recombination events (Wainscoat *et al.* 1986a). The reason for the stability in the region is, however, uncertain, though a low rate of intrachromosomal recombination has been implicated (Antonarakis *et al.* 1982a, Hill and Wainscoat 1986, Wainscoat *et al.* 1986a). The division between African and other Eurasian populations and the greater heterogeneity of haplotypes in Africa are consistent with a small founder population migrating out of Africa and giving rise to all non-African populations (Wainscoat *et al.* 1986a). Others see the distribution of the haplotypes as evidence for separate African and non-African lineages for mankind (Jones and Rouhani 1986).

The similarity in 5′ haplotype frequencies between all non-African populations makes this region relatively insensitive for population comparisons. This is emphasised by population affinity studies in the Asian Pacific populations which do not show consistent patterns in haplotype distributions, suggesting that sampling error, the confounding effects of gene flow, and genetic drift may be important (Hill and Wainscoat 1986, Wainscoat *et al.* 1986a, Chen *et al.* 1990).

Three 5′ haplotypes not previously described in Asian Indians (Kazazian *et al.* 1984a, Old *et al.* 1984, Thein *et al.* 1988, Varawalla *et al.* 1992) were seen in the South African sample. One was the common African haplotype (----+) and a second (+---+) has also been described in Africans, though it occurs at relatively low frequency (Ramsay and Jenkins 1987). As both occur in single individuals African admixture is a possible explanation, although such admixture appears to be relatively rare in the South African Indians. There is some evidence for a small amount of admixture, about 1-4%, based on red cell antigens, serum protein markers and mitochondrial DNA studies (Jenkins *et al.* 1970, Moores 1980, Soodyall 1993). The (----+) haplotype, however, also occurs at low frequency in Melanesia, while the (+---+) haplotype occurs at low frequency in Australian Aborigines, Papua New Guinea Highlanders, Polynesians and Micronesians (Chen *et al.* 1990, Flint *et al.* 1993b). Their presence may thus reflect a common ancient origin of the Indians with these peoples and the haplotypes may represent rare Asian haplotypes.

The 3′ haplotypes are invariably associated with the different polymorphic normal ß globin gene sequences or 'frameworks', three frameworks (Fw) occurring in each racial group (Antonarakis *et al.* 1982b, Orkin *et al.* 1982a). Five of eight possible 3′ haplotypes were observed, representing all three frameworks. The haplotypes (– +) and (+ + +) were the most common, occurring at almost equal frequencies in the Moslem Gujerati, (– –+) being more common in the Hindu Gujerati and Hindu Hindi and (+ + +) being more common in the Hindu Tamil. In a study of Indians in the United States, mainly from Bombay and New Delhi, (+ + +) was the most common (Kazazian *et al.* 1984a). As most studies have excluded the *Hin*dIII/3′ß site, it is necessary to combine haplotypes (+–+) and (+ + +) (Fw1) as well as (+ +–) and (+– –) (Fw2) for comparison.

Previous descriptions of 3′ haplotypes have noted the *Hin*dIII/3′ß site to be present or absent on Fw1, present on Fw2, and absent on Fw3 (Kazazian *et al.* 1984a,c). In the present study, a few Fw2 chromosomes with the site absent have been described on $ß^A$ and $ß^T$ chromosomes, but only in the Hindu Gujerati. This may be due to a relatively recent mutation and genetic drift in this group.

In all the major groups Fw2 was found on the fewest chromosomes, between 15% and 25%. In the Moslem Gujerati and Hindu Tamil, the majority of chromosomes were Fw1, while in the Hindu Gujerati there were approximately equal numbers of Fw1 and Fw3 chromosomes, and in the Hindu Hindi most chromosomes were Fw3. In the study of Indians in the United States, about 50% of chromosomes were Fw1, 30% Fw3 and 20% Fw2. In the study by Varawalla *et al.* (1992) of Indians in the United Kingdom, about 50% of the chromosomes were Fw1, 35% Fw3 and 15% Fw2. The common haplotypes accounting for approximately 90% of chromosomes were presented according to the geographical origins of the subjects. From these data, and assuming the rare haplotypes were not of a single framework, Fw1 was commonest in north-west Pakistan and Sindh, while Fw1 and Fw3 occur at approximately equal frequencies in Gujerat and Punjab (Varawalla *et al.* 1992). In contrast to the Hindu Tamil in this study, where the majority of chromosomes were Fw1, Fw3 was the commonest in a group from Tamil Nadu in southern India (Venkatesan *et al.* 1992). The latter sample from Tamil Nadu were from thalassaemia major families, whereas the South African Hindu Tamil studied were from a random group who do not appear to have as high a rate of ß thalassaemia. It is possible

that ß thalassaemia only occurs at high frequency in a subset of the population of Tamil Nadu. Further, the families with individuals who have thalassaemia major have a high rate of consanguinity and thus their $ß^A$ chromosomes may not represent a random sample. The South African sample may have originated from different geographical regions or may be different due to a founder effect, though no other evidence for this theory has been seen in this group. Fw1 may thus be more common in the northern and western areas of the Indian subcontinent, while Fw3 may be more common in the east. The commonest Fw in the south is difficult to determine.

In Mediterranean and Black populations Fw1 predominates and Fw3 is the least frequent (Antonarakis *et al.* 1982b), while in south-east Asian populations, including those of China, Cambodia and Thailand, Fw2 and Fw3 were more common, with Fw3 occurring at a higher frequency in China and Cambodia (Antonarakis *et al.* 1982b, Chan *et al.* 1986b, Hundrieser *et al.* 1988a). In Melanesia, Fw2 is most common in Papua New Guinea, whereas Fw1 predominates in Vanuatu (Hill *et al.* 1988). The Indians appear to have a framework distribution intermediate between that of Mediterranean and south-east Asian groups. However, Fw3 has four nucleotide substitutions in Mediterraneans, but only three in Asians and Blacks relative to Fw2 (Antonarakis *et al.* 1982b). The presence of Fw3 in the Hindu Hindi at high frequency may thus suggest a higher south-east Asian contribution to this group. The latter group has been found to have other features suggestive of a south-east Asian influence, including a high frequency of the $-\alpha^{4.2}$ chromosome (see Section 4.1.1).

5.2.3.2 $ß^T$ haplotypes

The 11 different haplotypes observed on 58 $ß^T$ chromosomes in the present study were all consistent with those previously reported in Asian Indians on $ß^T$ chromosomes, though six could differ at the *Hin*dIII/3'ß site (Kazazian *et al.* 1984a, Old *et al.* 1984, Thein *et al.* 1988, Varawalla *et al.* 1992). As has been previously reported (Old *et al.* 1984, Varawalla *et al.* 1992), the $ß^T$ haplotypes were less diverse than the $ß^A$ ones, occurring as a subset of the $ß^A$ haplotypes. All the $ß^T$ haplotypes were represented on $ß^A$ chromosomes in the South African Indians, except haplotype 30. A consistent haplotype is, however, reported by Varawalla *et al.* (1992) on $ß^A$ and $ß^T$ chromosomes.

No significant differences in the ß^T haplotype distributions were seen between the subgroups of South African Asian Indians. This may in part be due to the small numbers of chromosomes studied in each group, as there do appear to be interesting features and different haplotype frequencies in some of the groups.

Haplotype 1, one of the two common haplotypes on ß^A chromosomes, was also the most common on ß^T chromosomes in all groups except the Hindu Gujerati and the Moslem Memon. It occurs at much higher frequency, however, on the ß^T chromosomes. This haplotype was also the most common on ß^T chromosomes in Asian Indians in the United States, occurring on 63.6% of chromosomes (Kazazian *et al.* 1984a) and in Asian Indians in the United Kingdom originating from Gujerat, Punjab and Sindh, occurring on 62%, 37% and 60% of chromosomes respectively, but not in those from north-west Pakistan (Varawalla *et al.* 1992). The haplotype frequency was highest in the Hindu Hindi, where it reaches 90%, though only 10 chromosomes were studied and sampling error may have occurred. The Hindu Hindi have, however, been shown to have unique features with a number of systems, possibly due to their geographical origin or genetic drift. On the basis of these data, one might have predicted that the mutations associated with haplotype 1, namely IVS1nt5 (G→C) and the 619bp deletion would be the most common on the Indian subcontinent in all the groups, except the Hindu Gujerati, Moslem Memon and those from north-west Pakistan. In general, haplotype 1 appears to increase in frequency from west to east across the subcontinent, though there were some exceptions, possibly due to genetic drift or founder effect in some of the groups.

Haplotype 16, the second most common haplotype on ß^A chromosomes, appears to be extremely rare or absent on ß^T chromosomes in all the South African Asian Indian groups. It was only found on one chromosome in the entire study. (Haplotype 7, which differs only at the *Hin*dIII/3'ß site was also absent). This is in contrast to the groups studied in the United Kingdom, where (+---- ++) was the most common ß^T haplotype in north-west Pakistan (45%), second most common in Punjab (17%), and third most common in Gujerat (9%) and Sindh (15%) (Varawalla *et al.* 1992). One might thus predict that the mutation associated with this haplotype, namely frameshift codon 8/9, would be extremely rare in South African Asian Indians. In the study of Indians in the United States, the frameshift codon 8/9 mutation is associated with haplotype 7, which differs only at the

*Hin*dIII/3'ß site from haplotype 16 (Kazazian *et al.* 1984a), thus the haplotype in the United Kingdom Indians may be haplotype 7 rather than 16. The (+---- ++) haplotype thus appears to be more common in the north-western parts of the Indian subcontinent, decreasing in frequency toward the south and east. The clines in frequency of haplotype 1 and 16 (or 7) on $ß^T$ chromosomes thus appear to mirror those which occur on $ß^A$ chromosomes.

The Hindu Gujerati appeared to have different $ß^T$ haplotypes from all the other groups in this study, though only eight chromosomes were studied and thus sampling error may have occurred. The commonest haplotype in this group, 2, was rare in all the other groups, except perhaps the Hindu Tamil. It does, however, occur at significant frequencies in United Kingdom Indians, originating from Sindh and Gujerat (Varawalla *et al.* 1992). In addition, haplotypes 22, 28 and 30 do not occur in any of the other groups in this study, though 22 has been reported among the common haplotypes in Indians from the United Kingdom (Varawalla *et al.* 1992). One might thus predict that the Hindu Gujerati may have a different distribution of ß thalassaemia mutations compared to the other groups. There were insufficient data available to determine whether the differences were specific to the South African Hindu Gujerati or occur in this group in India also. They may be due to an unusual composition of the South African group, perhaps due to a large number of related individuals, or individuals from the same caste having emigrated together. Alternatively genetic drift may have occurred in the parent group in India, perhaps because of a high rate of intra-caste marriages.

ß thalassaemia mutations are likely to have arisen a few times. The chromosomes carrying these mutations would be expected to undergo positive selection in malarial areas, as heterozygotes for ß thalassaemia are relatively well protected against the disease, resulting in a difference in haplotype distribution on $ß^A$ and $ß^T$ chromosomes, as well as a more limited number of haplotypes. This was observed in all the groups, though the differences were only statistically significant in the total sample and the Moslem Gujerati. This is probably because of the smaller sample sizes in the other groups. Differences in haplotype distribution on $ß^A$ and $ß^T$ chromosomes, with a more limited number of $ß^T$ than $ß^A$ haplotypes, have been observed in previous studies of Asian Indians (Kazazian *et al.* 1984a, Varawalla *et al.* 1992), as well as in studies on other populations, including those

in Italy (Wainscoat *et al.* 1986b, Toffoli *et al.* 1988), Sardinia (Wainscoat *et al.* 1983b) and south China (Chan *et al.* 1986b). The similarity of β^A and β^T haplotypes on the Indian subcontinent as a whole and in the regional groups is seen as evidence for a relatively recent origin for the ß thalassaemia mutations on chromosomal backgrounds already existing in the population (Varawalla *et al.* 1992).

The 5′ and 3′ β^T haplotypes were also analysed separately. As on the β^A chromosomes, (+----) was the commonest 5′ haplotype and three 5′ haplotypes, (+---), (-+--++) and (-+-+-+), account for 100% of all β^T chromosomes in South African Asian Indians studied. In the Indians in the United Kingdom the three 5′ haplotypes account for 99% of all β^T chromosomes (Varawalla *et al.* 1992). The rarer 5′ haplotypes thus seem to be under-represented on the β^T compared to the β^A chromosomes. The frequencies of the haplotypes were similar in all the groups on β^A and β^T chromosomes, except in the Hindu Hindi where (+---) only occurs on 44% of β^A chromosomes, but on 100% of β^T chromosomes. The 5′ β^T haplotypes do not contain the ß globin gene, the region which has undergone positive selection. Thus there would be no selection against recombination at the hotspot between the δ and ß genes and one might expect the 5′ haplotype frequencies to approach those on β^A chromosomes, after many generations of recombination. The 3′ haplotype (--+), associated with Fw3 was the commonest on β^T chromosomes in all the groups, except the Moslem Memon, in the present study and occurs on at least 50% of chromosomes. Fw3 was also the commonest in United Kingdom Indians from Gujerat, Punjab and Sindh, accounting for 67%, 52% and 60% of chromosomes respectively (Varawalla *et al.* 1992), in United States Indians, predominantly from Bombay and New Delhi, accounting for 68% of chromosomes (Kazazian *et al.* 1984a) and in Indians from Tamil Nadu in southern India, accounting for 80% of chromosomes (Venkatesan *et al.* 1992). In Indians from north-west Pakistan, however, Fw1 was the commonest and reflects the haplotype occurring with the predominant mutation in this region, frameshift codon 8/9 (Varawalla *et al.* 1992). Fw3 was associated with two of the common mutations, namely the IVS1nt5 (G→C) and the 619bp deletion, and it is thus not surprising that it was the commonest framework. Further, one could predict from the high frequency of Fw3 in the South African Indians, that these two mutations are likely to be common.

The 3′ haplotype or framework is in strong linkage disequilibrium with the ß thalassaemia mutations, which have undergone positive selection. It is not unexpected that only a limited number of haplotypes were present in each group, and that their distributions were different from those on $ß^A$ chromosomes, both in the present study and those of Old *et al.* (1984) and Varawalla *et al.* (1992). These differences only reach statistical significance in the Moslem Gujerati, possibly because the sample sizes in the other groups were too small.

The associations of haplotypes and ß thalassaemia mutations are discussed further in Chapter 6.

5.2.3.3 $ß^S$ haplotypes

In all the South African Indians studied, originating from the four major groups and two of the minor groups, the $ß^S$ gene was associated with haplotype 25, which is extremely rare on $ß^A$ chromosomes in the South African Indians and in Asian Indians studied in the United Kingdom (Varawalla *et al.* 1992). It is, not unexpectedly, identical to the Arab-Indian $ß^S$ haplotype previously described (Bakioglu *et al.* 1985, Wainscoat *et al.* 1985, Kulozik *et al.* 1986, Miller *et al.* 1986,1987, Labie *et al.* 1989) and has reached its high frequency on $ß^S$ chromosomes because of positive selection acting on the heterozygous carriers of the $ß^S$ gene. The haplotype differs from the four African $ß^S$ haplotypes both 5′ and 3′ to the recombination hotspot and is therefore likely to be of independent origin (Kulozik *et al.* 1986).

A form of sickle cell anaemia, associated with mild clinical and haematological manifestations, and characterised by high HbF levels in the homozygotes, had been observed in individuals from eastern Saudi Arabia (Weatherall *et al.* 1969, Perrine *et al.* 1972, Pembrey *et al.* 1978, Perrine *et al.* 1978), Kuwait (Ali 1970), Iran (Haghshenass *et al.* 1977) and India (Brittenham *et al.* 1979, Kar *et al.* 1986). The distribution of the Arab-Indian $ß^S$ haplotype corresponds with the distribution of this mild form of the disease (Kulozik *et al.*1986).

On the Indian subcontinent, the Arab-India haplotype accounts for over 90% of the $ß^S$

haplotypes in India. Two atypical haplotypes described could have arisen by recombination at the hotspot in the ß globin cluster (Labie *et al.* 1989). The virtual homogeneity of the $ß^S$ haplotype in both the tribal and non-tribal populations of India, as well as in a number of the Arab populations, has some important anthropological implications.

It is suggested that the Arab-India $ß^S$ mutation arose once at a time when the tribal populations were in direct contact and subjected to either panmixis or gene flow, a different situation from that of today where these populations are isolated, endogamous groups, with different languages, cultures and religious customs, scattered all over the Indian subcontinent and surrounded by the mainstream or 'non-tribal' Indian populations. It is hypothesised that the tribal populations originated from the Harappan culture that flourished in the Indus river region, in what is now Sindh and Pakistan, about 4 000-5 000 years ago. It has been suggested that the tribal populations of India were part of the original inhabitants of the subcontinent and they may have preceded the Aryan and Mongol invasions of India. They migrated east after the collapse of that civilisation, around 5 000-2 000BC, and constituted the original inhabitants of the Indian subcontinent. They may have been further dispersed and isolated by the Aryan and Mughal invasions about 1 600-2 000BC. Figure 5.4 illustrates the proposed origin and spread of the Arab-Indian $ß^S$ haplotype. The Harappan culture was one of active advanced agriculture, based on the cultivation of barley, wheat and other crops and animal husbandry, in an area subjected to floods. Further, there was a large concentrated population estimated at about five million people, at the height of the civilisation, and thus ideal conditions existed for the development of malaria as a pandemic, a necessary requirement for the selection and expansion of the $ß^S$ gene. The mutation would have arisen 4 000-7 000 years ago, making it older than the African mutations that arose 2 000-3 000 years ago, at the time of the development of agriculture, and malaria pandemics, on the African subcontinent (Labie *et al.* 1989, Nagel and Labie 1989, Nagel and Ranney 1990, Nagel and Fleming 1992).

The non-tribal populations of India would have acquired the $ß^S$ gene from the tribal populations over a long period of time during which they co-existed (Labie *et al.* 1989). Further, under the influence of the Aryan invasion, groups could have migrated from the Indus valley west, east and south carrying the $ß^S$ gene to areas including what is presently

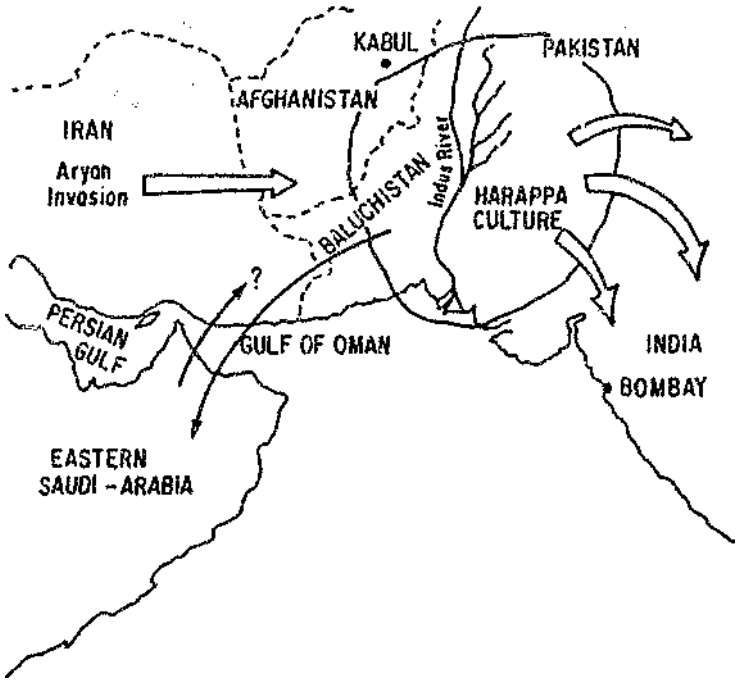


FIGURE 5.4 The proposed origin and spread of the Arab-Indian β^S haplotype (from Labie *et al.* 1989, Nagel and Fleming 1992).

The open arrows on the right represent the dispersion of the Harappan culture members after the collapse of the civilisation and the Aryan invasion (open arrows - left).
The solid arrows represent the possible pathways of gene flow of the β^S gene between the Indian subcontinent and the Arabian peninsula.

eastern Saudi Arabia, Afghanistan and Iran (Nagel and Labie 1989).

The Senegal and Arab-India haplotypes, which both have a mild phenotype, have the *Xmn*I site, 5′ to $^{G}\gamma$, which is associated with high $^{G}\gamma$ levels and relatively high HbF levels under conditions of erythropoietic stress (Thein *et al.* 1987a). The *Xmn*I site is, not unexpectedly, present on all the β^S chromosomes in this study

It is uncertain whether this mutation alters the binding of a regulatory protein or whether it is in linkage disequilibrium with an important mutation. This site is discussed further in Section 2.3.7. The $(AT)_xT_y$ motif 5′ to the ß globin gene, which differs on each of the haplotypes, appears to act as a silencer for transcription of the ß globin gene, by binding the repressor protein, Bp1. Binding is strongest to the Indian motif and weakest to the Bantu motif. Thus the Indian haplotype produces less β^S protein, resulting in a lower intracellular concentration of HbS, less polymerisation and a reduced severity. It is suggested that the silencing of the ß globin gene may lead to increased HbF levels by altering the balance between γ and ß expression (Elion *et al.* 1992).

A study of the 5′ hypersensitive site-2 of the locus control region from Benin and Senegal haplotypes found nucleotide variation in areas binding Sp1, Bp1 and other erythroid-specific and ubiquitous proteins. It is suggested that factors produced under conditions of haemopoietic stress, together with the genetic determinants of the different haplotypes, may allow for high level expression of the γ genes (Öner *et al.* 1992). Further, five different sequence configurations of a repeat motif $(AT)_xN_{12}(AT)_y$ are in strong linkage disequilibrium with each of the five major β^S haplotypes, which may be of functional relevance (Périchon *et al.* 1993).

The mechanisms by which the different haplotypes influence the severity of the disease are, however, still incompletely defined, though it does appear as if individuals have a preset rate of progression of disease, which is, at least in part, haplotype determined (Powars *et al.* 1990). In addition, it seems as if other factors not related to the haplotype, which may or may not be linked to the ß cluster, have a major effect on HbF levels (Rieder *et al.* 1991). Other genetic factors, such as co-inherited α thalassaemia also alter the clinical manifestations of the disease (reviewed in Powars *et al.* 1990).

5.2.3.4 β^E haplotypes

The two haplotypes found associated with the β^E gene in the South African Indians, namely 22 (+---- ++-) and 23 (-+-++ ++-), are both associated with framework 2. Haplotype 23 was the commonest β^E associated haplotype in most south-east Asian populations, followed by haplotype 22 (Hundrieser *et al.* 1988a). As both occur in south-east Asia and can be derived from one another by a single meiotic crossover at the recombination hotspot, they are thought to represent a single origin of the β^E gene (Antonarakis *et al.* 1982b). In the Kachari people of Assam in north-eastern India, who have an extremely high frequency of the β^E gene, the haplotypes were also types 23 and 22 (Hundrieser *et al.* 1988b).

A third haplotype associated with framework 3, also occurs in south-east Asia and is thought to represent a second origin of the gene (Antonarakis *et al.* 1982b). The framework 3 haplotype is found at high frequency only in the Khmer population of Cambodia (Hundrieser *et al.* 1988a, Nagel and Ranney 1990) and was not found in this study.

In view of the limited geographical distribution of HbE, it has been suggested recently that the mutation arose only once in south-east Asia and has spread to two different frameworks by gene conversion events (Flint *et al.* 1993a,b), a similar explanation to that put forward for HbS (this is discussed in detail in Section 6.2.2).

The data on the β^E haplotypes on the Indian subcontinent are consistent with a common origin of the gene in the populations of south-east Asia and India. Although there are no historical links between the Indian subcontinent and south-east Asia, a number of south-east Asian genetic markers were found on the Indian subcontinent, including the β^E gene, the $-\alpha^{4.2}$ chromosome (see Section 4.2.1) and the codon 41/42 ß thalassaemia mutation (see Section 6.2.1), suggesting either that gene flow has occurred between the populations or that they had a common ancestry.

5.2.3.5 β^D haplotypes

The β^D gene was associated with the same haplotype in the two individuals studied, providing evidence for a common origin for the mutation. The haplotypes associated with the gene have not been studied elsewhere and thus no comparisons were possible.

5.2.3.6 Haplotypes associated with γ globin rearrangements

The γγγ haplotypes observed in the two individuals studied cannot be unequivocally determined at the two $^{G}\gamma$ sites. They were thus (+[±±±]++ +++) and (+[±±±]++ --+). The haplotypes associated with this rearrangement have not been studied in other populations, except the Melanesians, where the 5′ haplotypes were either (+[---]--) or (-[+++-]++) (Hill *et al.* 1986), and thus of different origins from those observed in this study.

The γγγ rearrangement is more likely to be due to intrachromosomal sister chromatid exchange, rather than interchromosomal recombination (Hill *et al.* 1986). The two types found in Melanesia are derived from the two commonest 5′ haplotypes, (+----) and (-+--++). The rearrangements in the Asian Indians cannot be derived from any of the 5′ haplotypes observed in this study or other studies of Asian Indians, either by intra- or interchromosomal recombination. Their origin is thus difficult to determine. They do, however, have the same 5′ haplotype. The 3′ ends differ but this could be due to recombination at the hotspot between the δ and ß genes. They are therefore likely to have a single common origin.

The -γ chromosome has a (-[-]±± --+) haplotype. This may be the same as the more frequently observed -γ haplotype in Melanesia, (-[-]++), the second commonest being (+[-]--). If shared, the haplotype may suggest a common ancestral origin. In studies of the α globin haplotypes (see Sections 4.2.5.1 and 4.2.6.1), Indians and Melanesians have been shown to share some features. This is discussed in Section 8.2.2. The -γ rearrangement may be due to unequal intrachromosomal cross-over or intrachromatid deletion (Hill *et al.* 1986) and recurrent mutation on the common (-+--++) haplotype could have occurred.

5.2.3.7 Haplotypes associated with the *Xmn*I variant

The haplotypes which were associated with the new *Xmn*I variant in the Moslem Gujerati and Hindu Gujerati differ both 5′ and 3′ to the recombination hotspot, strengthening the evidence for two origins for the mutation, one on a ß^A chromosome, with an *Xmn*I site 5′ to $^{G}\gamma$ and a second on a ß^T chromosome, lacking an *Xmn*I site 5′ to $^{G}\gamma$. The *Xmn*I site 5′ to $^{G}\gamma$ occurs on a number of different haplotypes as shown in Table 5.8. It is thus possible that the sequence at -158 5′ to either of the γ genes represents a hotspot for mutation. Alternatively, as the site is situated in an area of homology between the two γ genes, gene conversion may be implicated. The mutation 5′ to $^{G}\gamma$ has been shown to increase gene expression, particularly under conditions of erythropoietic stress (Thein *et al.* 1987a). Positive selection may have acted on the mutation to increase its frequency.

5.2.4 Correlation of ß globin haplotypes and the *Xmn*I polymorphic site

The HPFH syndromes may be of the deletion or non-deletion type. In the former group, both γ genes are typically over-expressed, resulting in raised HbF levels in adult life. In the latter group, however, either the expression of the $^{G}\gamma$ or $^{A}\gamma$ gene is increased, resulting in raised HbF, predominantly of either $^{G}\gamma$ or $^{A}\gamma$ type. They are due to point mutations clustered upstream of the genes, which are thought to alter the binding of *trans*-acting factors, with increased affinity for proteins which activate the γ genes, or decreased affinity for repressor molecules (Superti-Furga *et al.* 1988, Martin *et al.* 1989, Ronchi *et al.* 1989, Wood 1993).

Mutations 5′ to $^{G}\gamma$ at -202 (C→G) (Collins *et al.* 1984) and -175 (T→C) (Ottolenghi *et al.* 1988b) result in HbF levels of 15-25% in heterozygotes, predominantly of $^{G}\gamma$ type. In contrast, mutations 5′ to $^{A}\gamma$ at -196 (C→T) (Giglioni *et al.* 1984, Gelinas *et al.* 1986), -117 (G→A) (Collins *et al.* 1985, Gelinas *et al.* 1985) and -198 (-A) (Tate *et al.* 1986) result in HbF levels in heterozygotes of 5-20%, predominantly of $^{A}\gamma$ type. A 4bp deletion in the $^{A}\gamma$promoter reduces binding of a *trans*-acting factor, thus reducing $^{A}\gamma$ expression (Beldjord *et al.* 1992).

The C→T mutation at -158 to $^{G}\gamma$, which creates an *Xmn*I site, is somewhat different in that it appears to alter the $^{G}\gamma$:$^{A}\gamma$ ratio from the normal adult 40:60 to 60:40, typical of a neonate, but produces minor or no elevations in the HbF levels in normal individuals. HbF levels, predominantly of $^{G}\gamma$ type, are raised under conditions of erythropoietic stress, such as sickle cell anaemia or ß thalassaemia, though the site itself is poorly correlated with absolute HbF levels. Lack of the site is, in general, correlated with low $^{G}\gamma$ expression (Gilman and Huisman 1985, Labie *et al.* 1985b, Motum *et al.* 1989). In neonates, the site is not correlated with HbF levels or composition (Rochette *et al.* 1990).

It is uncertain whether the *Xmn*I mutation is the determinant responsible for increased $^{G}\gamma$ expression or whether it is linked to the determinant, though DNA sequencing revealed no other mutations in the region (Miller *et al.* 1987). Additional circumstantial evidence suggests that the base substitution at -158 is itself responsible for increased HbF production (Thein *et al.* 1987a). Chromosomes with the $^{G}\gamma^{G}\gamma^{A}\gamma$ can be classified into four types based on the presence or absence of the *Xmn*I site 5′ to the two $^{G}\gamma$ genes and the corresponding $^{G}\gamma$ levels. $^{G}\gamma$ levels are raised where both *Xmn*I sites are present, and low when they are absent (Liu *et al.* 1988). In the South African Indians, the γγγ chromosomes lack the *Xmn*I site 5′ to both $^{G}\gamma$ genes and would thus be expected to have low $^{G}\gamma$ levels. As the site is near a DNase I hypersensitive site, it was proposed that the substitution increases the probability of an open chromatin configuration, making it more accessible to components of the transcription apparatus of the adult erythroid cell *in vivo* (Gilman and Huisman 1985). High HbF ($^{G}\gamma$ type) expression under erythropoietic stress and the *Xmn*I site, appears to be in strong linkage disequilibrium with the 3′ haplotype (±+-++) (Gilman and Huisman 1984, Harano *et al.* 1985c, Labie *et al.* 1985b, Hattori *et al.* 1986b, Thein *et al.* 1987a), though the mechanism is not defined. Recent studies have shown that interaction between the *Xmn*I site and a rare $(AT)_5T_9$ motif 5′ to the ß gene may be important in generating high HbF levels (Ragusa *et al.* 1992). This motif may be linked to a particular haplotype.

Recent research has shown that the mutation occurs in the region of a binding site for the ubiquitous octamer binding protein, OCT-1, flanked by binding sites for the erythroid protein GATA-1. Recent evidence suggests that the GATA-1 erythroid transcription factor binds slightly more efficiently to the mutated promoter (-175) than to the normal one

(reviewed in Ottolenghi *et al.* 1992), while the mutation abolishes binding of OCT-1 *in vitro* (reviewed in Wood 1993).

In contrast, the haplotype (-+ + -+ + +) has no associated *Xmn*I site, but high $^{G}\gamma$ expression, suggesting that the site, although highly correlated with $^{G}\gamma$ expression, does not perfectly predict the presence of high $^{G}\gamma$ expression and that the site is not solely involved in determination of high $^{G}\gamma$ expression. Other site substitutions in the same area or control sequence elsewhere in linkage disequilibrium with the haplotype may have similar effects (Labie *et al.* 1985a). Some heterozygotes for ß thalassaemia have raised HbF levels, independent of the mutation at position -158 (Kutlar *et al.* 1990).

In addition, other factors, not linked to the ß globin cluster may also alter HbF levels. For example, an X-linked determinant appears to affect the level of HbF response to anaemia (Miyoshi *et al.* 1988, Dover *et al.* 1992). Evidence for a major gene for heterocellular HPFH unlinked to the ß globin cluster has been found in a large Asian Indian pedigree (Thein *et al.* 1994). Analysis of Mediterranean patients with thalassaemia major suggests a major role for unknown loci unlinked to the ß globin cluster controlling γ chain production (Efremov *et al.* 1994).

The association of the *Xmn*I site and haplotypes was studied in South African Indians. As previously reported, the 5′ haplotypes (-+-+ +) and (+ + -+ + +) were strongly correlated with the presence of the site, and (+----) and (-+ + --) with the absence of the site on $ß^{A}$, $ß^{T}$ and $ß^{S}$ chromosomes. Unfortunately we were not able to correlate this with $^{G}\gamma$ or HbF levels.

The majority of $ß^{T}$ chromosomes studied originate from individuals with thalassaemia major and it is thus not surprising that most chromosomes studied lack the *Xmn*I site and most individuals were (-/-), since the inheritance of the *Xmn*I site is associated with a milder clinical phenotype due to higher HbF levels (Thein *et al.* 1988). Further, only one of the two 5′ haplotypes found to be strongly linked to the *Xmn*I site, (-+-+ +), was found on $ß^{T}$ chromosomes and it was found on only a small percentage of the chromosomes (24.6%). Although nine thalassaemia major individuals were (-/+), the evidence suggests that a single copy of this determinant may not be enough to effect a

sufficient increase in HbF to modify β^0 thalassaemia (Thein *et al.* 1987a). As most of the common Asian Indian ß thalassaemia mutations cause β^0 or severe β^+ thalassaemias, it would thus be expected that the (-/+) individuals would still have thalassaemia major. Three individuals with a severe thalassaemia major phenotype were (+/+), suggesting that even two copies of the *Xmn*I site may, in certain cases, be insufficient to modify the phenotype, perhaps because these chromosomes lack the other necessary cofactors or unidentified motifs for raised HbF, like the $(AT)_9T_5$ motif.

The β^S chromosomes in this study have the *Xmn*I site, consistent with previous findings. Further an $(AT)_9T_5$ motif is found on this haplotype and is associated with high HbF expression (Trabuchet *et al.* 1991a).

The β^E chromosome and the two β^D chromosomes associated with the (+----) and (-+--++) haplotypes represent interesting exceptions. The β^E chromosome has the *Xmn*I site, a pattern found only on two of 176 β^A chromosomes with the (+----) haplotype. Similarly, the two β^D chromosomes lack an *Xmn*I site, like only four of 61 β^A chromosomes with a (-+--++) haplotype. It is possible that the β^E and β^D mutations arose, by chance, on such a chromosome. Alternatively, the altered site could be due to a point mutation or gene conversion event on these chromosomes. In the case of β^E, a chromosome with the *Xmn*I site present may have been at a selective advantage, because of the increased $^G\gamma$ expression, with erythropoietic stress.

Two variant chromosomes observed in this study had an *Xmn*I site at the homologous position 5′ to the $^A\gamma$ gene. It would be interesting to study the effects of these mutations on the differential expression of the two genes. Unfortunately this was not possible, as no facilities for determination of globin chain ratios are available in South Africa. One could speculate that the chromosome with sites 5′ to both $^G\gamma$ and $^A\gamma$ would have normal ratios of the two chains, but may have slightly increased γ expression and HbF levels. HbF levels of 1.9% and 1.2%, only slightly above the normal range (<1%), were noted. In the individuals with a site 5′ to $^A\gamma$, but no $^G\gamma$ site, a relative increase in $^A\gamma$ chains would be expected. One of the two individuals studied has slightly raised HbF levels of 1.6%. However, this variant was found on a β^T chromosome and the individuals are ß thalassaemia heterozygotes and thus may have slightly raised HbF levels due to the effects

of the thalassaemia gene. The variant previously described with *Xmn*I sites 5′ to the two $^{G}\gamma$ genes on a $^{G}\gamma^{G}\gamma$ chromosome is associated with high $^{G}\gamma$ levels and modestly elevated HbF (Gilman and Huisman 1985).

It appears as if the $(AT)_9T_5$ motif may be necessary, but not sufficient alone, for high HbF expression. A combination of the haplotype (--+-++), the *Xmn*I site and the motif may elicit high HbF levels even in patients with very mild anaemia (Ragusa *et al.* 1992). A study of two Sicilian thalassaemia major patients, one with a mild and one with a severe phenotype, supports this hypothesis. Both had the same haplotypes and the *Xmn*I site, but the mild phenotype was associated with the presence of homozygosity for an $(AT)_9T_5$ motif, whereas the severe one was homozygous for an $(AT)_7T_7$ motif at the $(AT)_xT_y$ site 5′ to the β gene. It is suggested that the $(AT)_9T_5$ motif is a negative regulatory binding sequence, whereas the more common $(AT)_7T_7$ has less affinity for a regulatory protein (Ragusa *et al.* 1992). The $(AT)_9T_5$ motif has been shown to act as a silencer in K562 cells (Berg *et al.* 1989, 1991). However, more recently it has been proposed that variation in this motif is a common polymorphism, with no role in regulation of γ expression (Wong *et al.* 1989, Galanello *et al.* 1993, Dimovski *et al.* 1994).

CHAPTER 6 - ß THALASSAEMIA MUTATIONS AND THEIR ASSOCIATED HAPLOTYPES IN SOUTH AFRICAN INDIANS

This chapter describes the mutations causing ß thalassaemia in the South African Indians. The associated ß globin haplotypes have been characterised in an attempt to provide some insight into the origins of the mutations.

6.1 Results

6.1.1 ß thalassaemia mutations

Altogether 68 ß thalassaemia chromosomes from unrelated individuals in 44 Asian Indian families were studied. The religious and linguistic affiliations were not known for three of the families (six chromosomes).

The ß thalassaemia mutations identified are shown in Table 6.1, together with the religious and linguistic affiliations of the individuals studied. The IVS1nt5 (G→C), IVS1nt1 (G→T), codon 41/42 (-CTTT), codon 30 (G→C), codon 15 (G→A), cap site (+1) (A→C) mutations were identified using the ARMS technique. The 619 bp deletion was detected with the control primers A and B used for the ARMS technique. The remaining mutations, codon 30 (G→A), codon 44 (-C) and Poly-A (T→C) were identified by sequencing.

Four of the five mutations, previously described as common in Asian Indians (Kazazian *et al.* 1984a, Thein *et al.* 1988, Varawalla *et al.* 1991b, 1992), namely IVS1nt5 (G→C), the 619bp deletion, IVS1nt1 (G→T) and codon 41/42 (-CTTT), were found in this cohort, and account for the majority (84%) of the ß thalassaemia chromosomes. The fifth mutation, codon 8/9 (+G), was not found.

The IVS1nt5 mutation was the commonest mutation, occurring in all but one group (the Moslem Memon) with frequencies between 40% and 83%. The other three common mutations each account for between 8% and 60% of chromosomes in the different groups, in which they were found. In the Hindu Gujerati, the IVS1nt5 and codon 41/42 mutations

TABLE 6.1 Frequencies of β thalassaemia mutations in South African Indians

MUTATION[1]	MOSLEM GUJERATI		HINDU GUJERATI		HINDU HINDI		HINDU TAMIL		MOSLEM URDU		MOSLEM MEMON		HINDU TELEGU		TOTAL[3]	
	N[2]	Freq	N[2]	Freq	N[2]	Freq	N[2]	Freq	N[2]	Freq	N[2]	Freq	N[2]	Freq	N[2]	Freq
IVS1nt5 (G→C)*	13	0.54	4	0.40	7	0.70	4	0.67	5	0.83	-	-	1	1.00	39	0.57
619bp deletion*	2	0.08	-	-	1	0.10	-	-	-	-	2	0.40	-	-	5	0.07
IVS1nt1 (G→T)*	3	0.13	-	-	-	-	-	-	1	0.17	3	0.60	-	-	7	0.10
Codon 41/42 (-CTTT)*	1	0.04	4	0.40	1	0.10	-	-	-	-	-	-	-	-	6	0.09
Codon 30 (G→C)*	1	0.04	1	0.10	-	-	-	-	-	-	-	-	-	-	2	0.03
Codon 15 (G→A)*	1	0.04	1	0.10	-	-	1	0.17	-	-	-	-	-	-	3	0.04
Cap site (+1)(A→C)*	1	0.04	-	-	-	-	-	-	-	-	-	-	-	-	2	0.03
Codon 30 (G→A)#	1	0.04	-	-	-	-	-	-	-	-	-	-	-	-	1	0.01
Codon 44 (-C)#	1	0.04	-	-	-	-	-	-	-	-	-	-	-	-	1	0.01
Poly A (T→C)#	-	-	-	-	-	-	1	0.17	-	-	-	-	-	-	1	0.01
Unknown	-	-	-	-	1	0.10	-	-	-	-	-	-	-	-	1	0.01
Total	24		10		10		6		6		5		1		68	

[1] The following mutations were not found in this cohort: codon 8/9 (+G)*, codon 5 (-CT)#, codon 16 (-C)*, IVS2nt837 (T→G)*, IVS2nt1 (G→A)*, -88 (C→T)*, codon 88 (+T)#, IVS1 -25bp deletion*, codon 55 (+A)#, codon 47-48 (+ATCT)#.

* Identified by the ARMS technique.

Identified by direct sequencing from PCR products.

[2] N = number of chromosomes.

[3] The religious and linguistic affiliations were unknown for individuals with 6 of the β-thalassaemia mutations, 5 IVS1nt5 and 1 Cap site (+1).

account for equal numbers of chromosomes. The Moslem Gujerati appeared to have the largest number of different mutations.

Seven mutations (detected by the ARMS system) accounted for 64 out of 68 (94.1%) ß thalassaemia chromosomes. Four ß thalassaemia chromosomes with unknown mutations remained and DNA sequencing was carried out in an attempt to identify these. Three mutations were defined, two of which have not been previously described in Asian Indians:

(i) Codon 30 (G→A) - reported previously as a rare Indian mutation (Varawalla *et al.* 1991c) (Figure 6.1a).

(ii) Frameshift mutation codon 44 (-C) (Figure 6.1b).

(iii) Polyadenylation site mutation (AATAAA→AACAAA) (Figure 6.1c).

The fourth mutation remains undefined after sequencing the regions of the ß globin gene where the majority of ß thalassaemia mutations have been found. The regions of the gene which remain to be sequenced include IVS2, from nt50 onwards, exon 3, as well as the region 5' to -60 from the cap site. All 17 previously described Indian mutations (Kazazian *et al.* 1984a, Thein *et al.* 1988, Varawalla *et al.* 1991b,c) have been excluded on this chromosome by ARMS screening or DNA sequencing. The two mutations most recently described in individuals from Maharashtra and Punjab, codon 55 (+A) and codon 47-48 (+ATCT) respectively (Garewal *et al.* 1994) have also been excluded by sequencing. A large deletion or rearrangement is also unlikely as normal-sized fragments were observed on Southern blots after digestion with the restriction enzymes *Ava*II, *Bam*HI, *Hin*fI and *Pst*I and hybridisation with the ß globin probe.

Eight of the 19 previously described Indian ß thalassaemia-causing mutations (Orkin *et al.* 1983, Kazazian *et al.* 1984a, Thein *et al.* 1988, Varawalla *et al.* 1991b,c, Garewal *et al.* 1994) were identified in this cohort. Two mutations not previously described in Indians were also found, while one mutation still remains unknown. Thus, in total, 10 mutations have been identified in South African Asian Indians, on 67 of 68 (98.5%) ß thalassaemia chromosomes studied.

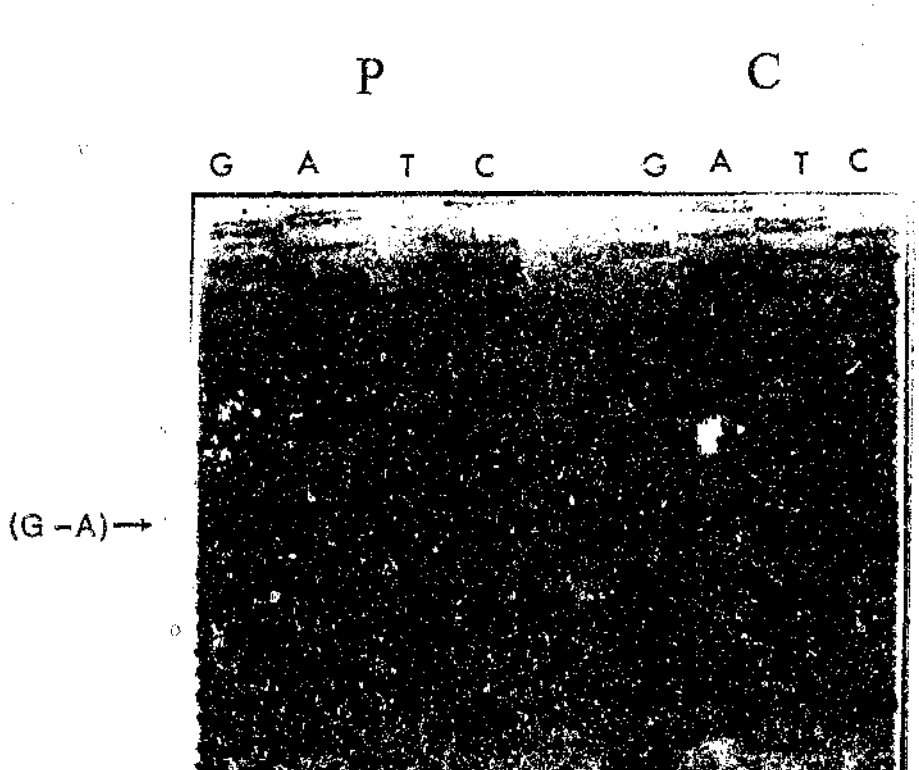
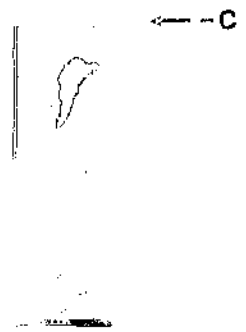


FIGURE 6.1a Sequence analysis of the amplified β globin gene DNA from a thalassaemia major patient (P) heterozygous for the rare Indian mutation, codon 30 (G→A), and a normal control (C)



gene DNA from a
terozygous for the

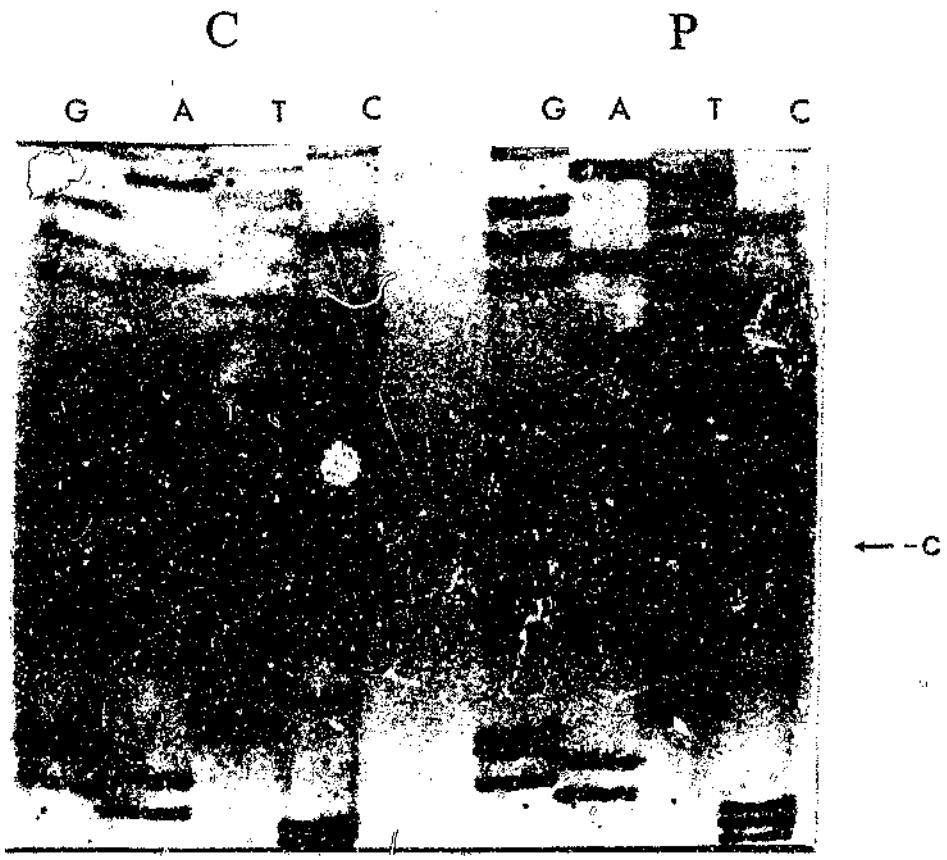


FIGURE 6.1b Sequence analysis of the amplified β globin gene DNA from a normal control (C) and a thalassaemia major patient (P) heterozygous for the frameshift mutation, codon 44 (-C)

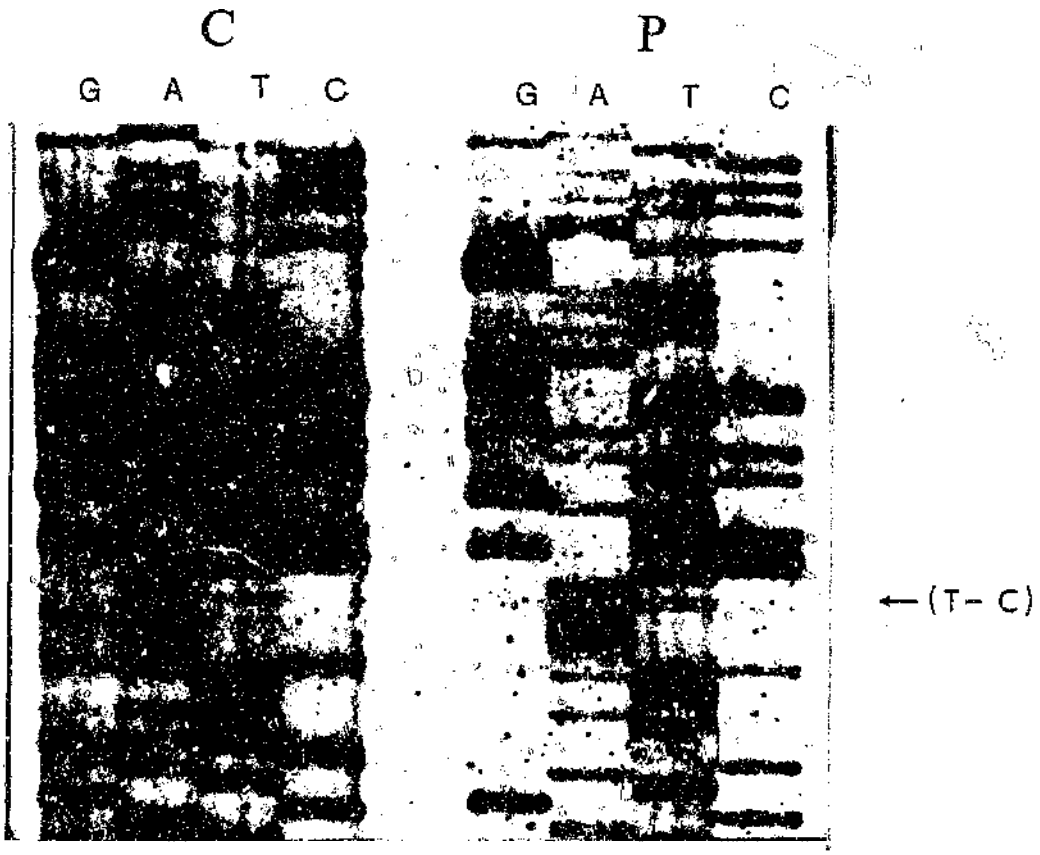


FIGURE 6.1c Sequence analysis of the amplified β globin gene from a normal control (C) and a thalassaemia major patient (P) heterozygous for the polyadenylation site mutation, AATAAA→AACAAA

6.1.2 Haplotypes associated with ß thalassaemia mutations

The mutations and their associated haplotypes are shown in Table 6.2. There were 11 haplotypes associated with the ß thalassaemia mutations. Four haplotypes were associated with more than one mutation and three mutations occur on more than one haplotype in this cohort. The association between the common mutations and their major haplotypes varies from 50-100%, while the mutation can be predicted from the haplotype with an accuracy of 66-100%.

The IVS1nt5 mutation occurred in this study with three different haplotypes, though all were Fw3. Haplotype 1 (+---- --+) was predominant in all the subgroups, in which the mutation was found, except the Hindu Gujerati, as shown in Figure 6.2. The codon 41/42 mutation occurs with four haplotypes, with four 3′ haplotypes and two frameworks. The haplotype distributions in the subgroups in which the mutation occurs are shown in Figure 6.3. Haplotype 22 (+---- ++-) appears to be the most common, though the numbers in the groups were small. All the haplotypes occur on the same framework (Fw2), except the one chromosome in a Moslem Gujerati individual (Fw1). Codon 15 (G→A) was associated with haplotype 2 (-+--++ --+) in a Hindu Tamil individual and haplotype 3 (-++--+ --+) in a Moslem Gujerati, though both have the same framework (Fw3).

6.2 Discussion

6.2.1 ß thalassaemia mutations

Of the 19 ß thalassaemia mutations described in Asian Indians to date (Orkin *et al.* 1979a, Orkin *et al.* 1983, Kazazian *et al.* 1984a, Thein *et al.* 1988, Old *et al.* 1990, Varawalla *et al.* 1991b,c, Garewal *et al.* 1994), eight occur in the South African cohort. Two mutations not previously described in Indians were found and a third mutation remains to be characterised. Thus a total of 21 Asian Indian ß thalassaemia mutations are now known, 10 of which have been identified in the South African Asian group.

An initial survey of Indians in the United Kingdom (Thein *et al.* 1988) showed that no one mutation predominated, but that 88% of the ß thalassaemia alleles were due to five

TABLE 6.2 β thalassaemia mutations and their associated haplotypes in South African Indians

HAPLOTYPE[1]									MUTATION										
Type	Sites								IVSInt5 (G→C)	619bp deletion	IVSInt1 (G→T)	Codon 41/42 (-CTTT)	Codon 30 (G→C)	Codon 15 (G→A)	Cap site (+1) (A→C)	Codon 38 (G→A)	Codon 44 (-C)	Poly A (T→C)	?	Total
	1	3	4	5	6	7	8	9												
1	+	–	–	–	–	–	–	+	20	5	-	-	-	-	1	-	-	-	1	33
2	–	+	–	+	+	–	–	+	5	-	-	-	-	1	-	-	-	-	-	6
3	–	+	+	–	-	–	–	+	2	-	-	-	-	1	-	-	-	-	-	3
8		+	–	+	+	+	–	+	-	-	7	-	-	-	-	-	-	-	-	7
9	–	+		–	+	+	–	+	-	-	-	-	-	-	-	-	-	1	-	1
16	+	–	–	–	–	+	+	+	-	-	-	-	-	-	-	1	1	-	-	2
17	–	+	–	+	+	+	+	+	-	-	-	-	1	-	-	-	-	-	-	1
18	–	+	+	–	+	+	+	+	-	-	-	1	-	-	-	-	-	-	-	1
22	+	–	–	–	–	+	+	–	-	-	-	3	-	-	-	-	-	-	-	3
28	+	–	–	–	–	+	–	–	-	-	-	1	-	-	-	-	-	-	-	1
30	–	+	+	–	+	+	–	–	-	-	-	1	-	-	-	-	-	-	-	1
	Not determined								6	-	-	-	1	1	1	-	-	-	-	9
	Total								39	5	7	6	2	3	2	1	1	1	1	68

[1] Sites 1, 3-9 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes.

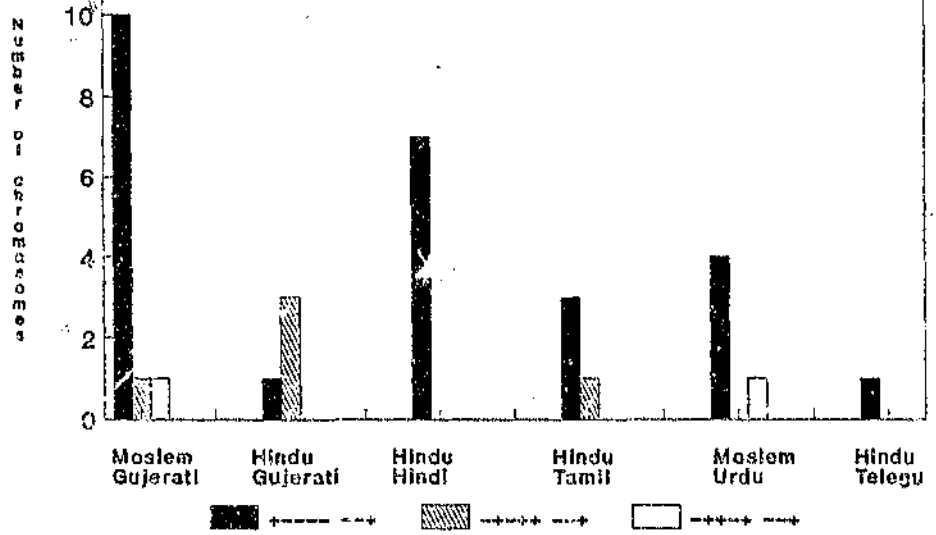


FIGURE 6.2 β globin haplotypes associated with the IVS1nt5 mutation in the South African Asian Indian groups

Sites 1, 3-9 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes

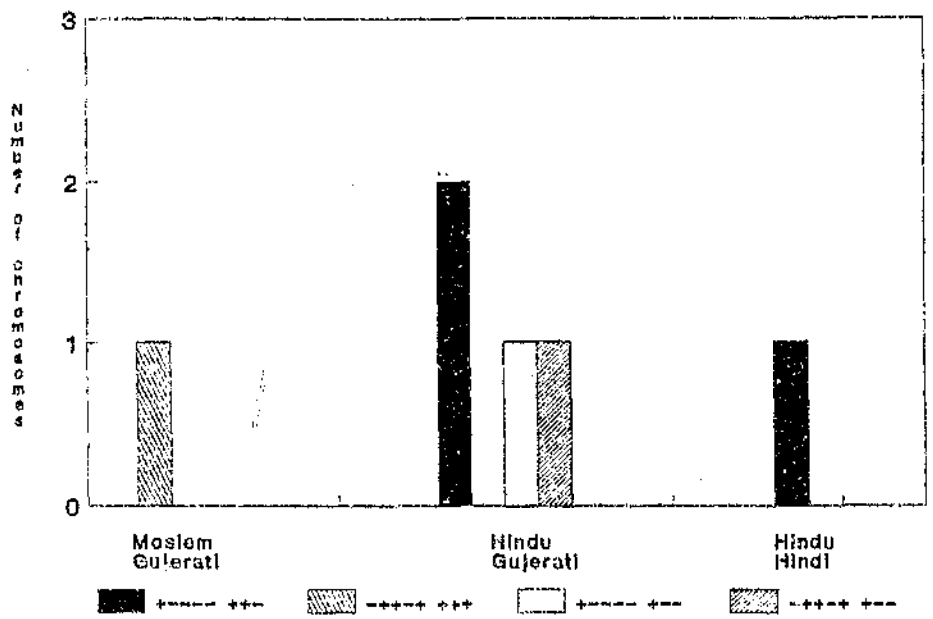


FIGURE 6.3 β globin haplotypes associated with the codon 41/42 mutation in the South African Asian Indian groups

Sites 1, 3–9 shown in Figure 2.2 and described in Table 2.5 were used to construct haplotypes

mutations, with the IVS1nt5 mutation, the 619bp deletion and the codon 8/9 mutation each accounting for about 20% and the IVS1nt1 and codon 41/42 mutations accounting for 14% and 12% of mutations, respectively. Further studies of Indians in the United Kingdom have shown similar overall frequencies, the five mutations accounting for 89-98.5% of mutations, though the IVS1nt5 mutation appears commoner than originally estimated, up to 43% in some groups (Parikh *et al.* 1990, Jain *et al.* 1991, Varawalla *et al.* 1991b, 1992). The distribution of ß thalassaemia mutations in different regions of the Indian subcontinent is shown in Figure 6.4.

Similarly, in the South African cohort, four of the common mutations account for the majority (67-100%) of the chromosomes in the subgroups studied. However, the fifth mutation, the codon 8/9 mutation, has not been found in South African Indians. The differences between the studies can be explained partly by the different geographical origins of individuals but also by factors such as genetic drift or founder effect, which may have operated both in India and in the small groups that have emigrated to different countries. For example, individuals in the study of Thein *et al.* (1988), came from Punjab, Rajasthan, Kashmir, Pakistan and Gujerat. The only area common with this study is Gujerat, from which three different groups were studied, all of which have different mutation distributions. Regional differences have also been shown by Varawalla *et al.* (1991b, 1992).

The IVS1nt5 (G→C) mutation has been shown to be the commonest on the Indian subcontinent (Varawalla *et al.* 1991b, 1992). It appears to be even more common in the Asian Indians in South Africa, occurring at higher frequencies (40-83%) than reported in previous studies. As predicted from the haplotype and framework data (Section 5.2.3.2), IVS1nt5 was the commonest mutation in all but the Moslem Memon, where sampling error cannot be excluded because of the small sample size. The mutation accounts for 79% of the haplotype 1 chromosomes. In general, there appears to be a rise in the frequency of this mutation from the western and northern regions of India towards the eastern and southern regions, the Hindu Hindi, Moslem Urdu and Hindu Tamil having the highest frequencies in this study. High frequencies were also observed in Bangladesh and Bengal in the east and Tamil Nadu in the south (Varawalla *et al.* 1991b, 1992, Venkatesan *et al.* 1992) and the lowest frequencies were observed in Pakistan, Punjab and particularly Sindh (Varawalla *et al.* 1991b, 1992). A high frequency of the mutation was

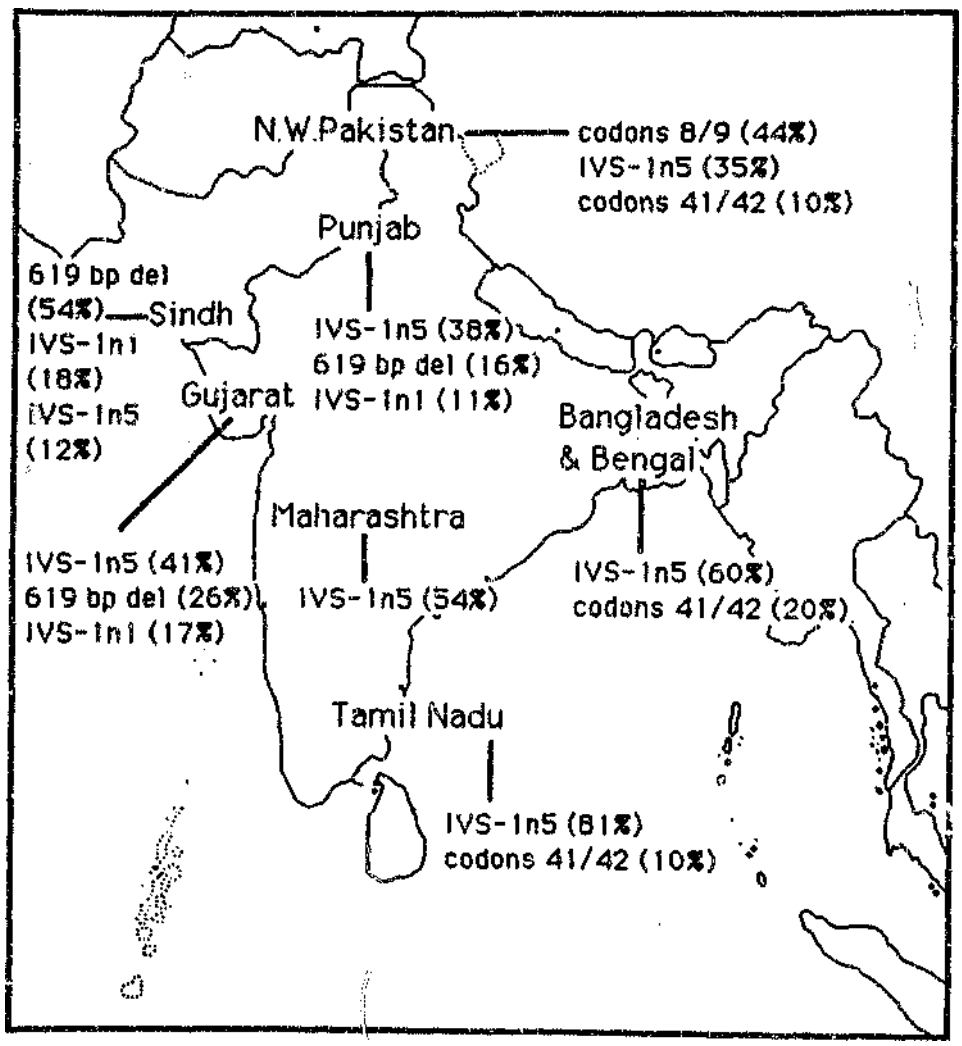


FIGURE 6.4 Distribution of β thalassaemia mutations in different regions of the Indian subcontinent (after Varawalla *et al.* 1991b).

also found in the United Arab Emirates (El-Kalla and Mathews 1993, Quaife *et al.* 1994). Further relatively high frequencies (>5%) of the mutation were also found in south-east Asia, including Indonesia (Lie-Injo *et al.* 1989), Thailand (Fucharoen *et al.* 1989, Laig *et al.* 1989, Thein *et al.* 1990, Laosombat *et al.* 1992), Burma (Brown *et al.* 1992) and Malaysia (Yang *et al.* 1989, Fucharoen *et al.* 1990b). Low frequencies of the mutation were found in China (Huang *et al.* 1990), Turkey (Altay and Gurgey 1992), Italy (Rosatelli *et al.* 1992a), Algeria (Bennani *et al.* 1993), Albania (Boletini *et al.* 1994), Azerbaijan (Kuliev *et al.* 1994) and among the Kurdish Jews (Rund *et al.* 1991). Groups in the northern and western regions of India may have had more contact with traders and invading groups and thus have a higher percentage of admixture with other groups. This is supported by finding that the Hindu Tamil from the southern regions were the most homogeneous, both in the current cohort and in the studies of Varawalla *et al.* (1991b) and Venkatesan *et al.* (1992). In addition, groups from the Gujerat region, Punjab and north-west Pakistan appear to have more diverse subsets of mutations in this study and others (Parikh *et al.* 1990, Varawalla *et al.* 1991b).

The 619bp deletion is unique to Indians and reaches its highest frequencies in subgroups from Sindh, where it is more common than the IVS1nt5 mutation (Parikh *et al.* 1990, Varawalla *et al.* 1991b, 1992). It was originally thought to be confined to Gujerati- and Sindhi-speaking groups of Pakistan and adjacent Gujerat (Thein *et al.* 1984b), but has also been found in Tamil Nadu and Punjab (Parikh *et al.* 1990, Varawalla *et al.* 1991b, 1992) and in the central subcontinent, in the Hindu Hindi in this study, albeit at lower frequency. It is thought to have arisen in Sindh and spread to the other regions at the time of the Harappan culture and subsequent Aryan invasion by gene flow with population migration (Thein *et al.* 1984b, Varawalla *et al.* 1992). In the South African Indians the mutation appears to be more common in the Moslem groups, though its frequency (0-10%) was generally lower than that predicted from the haplotype 1 distribution (Section 5.2.3.2) and lower than in previous studies, where it accounts for 13-36% of mutations (Kazazian *et al.* 1984a, Thein *et al.* 1988, Parikh *et al.* 1990, Varawalla *et al.* 1991b, 1992). Even in groups that appear to have come from the same region, namely Gujerat province, the South African groups had lower frequencies, possibly suggesting that individuals who emigrated to South Africa were not representative of the parent population. The South African groups also had a lower incidence of ß thalassaemia trait

than the 10-15% found in Gujerat (Sukumaran and Master 1974). In addition, the South African immigrants from Gujerat originated from the more southern parts of Gujerat (see Figure 1.4b), regions which may have had less contact with Sindh.

The IVS1nt1 (G→T) mutation occurs exclusively in the Moslems in this study, at similar frequencies to those previously reported in individuals from Gujerat and Sindh (Thein *et al.* 1988, Parikh *et al.* 1990, Varawalla *et al.* 1991b, 1992), around 15%, except for the Moslem Memon, where it reaches 60%. Its frequency was lower in Punjab and Pakistan, <10% (Varawalla *et al.* 1991b, 1992), again suggesting a cline in frequency from west to east on the Indian subcontinent. It also occurs at frequencies >5% in Burma (Brown *et al.* 1992), Indonesia (Lie-Injo *et al.* 1989), and Malaysia (Yang *et al.* 1989), and at lower frequency in Thailand (Fucharoen *et al.* 1989, Thein *et al.* 1990, Laosombat *et al.* 1992) and China (Kazazian *et al.* 1986b, Zhang *et al.* 1988, Huang *et al.* 1990, Liang *et al.* 1994).

The small Moslem Memon group studied in South Africa has only the 619bp deletion and the IVS1nt1 mutation. These individuals came from a few towns and villages in the Kathiawar region of Gujerat, and settled in South Africa in two relatively isolated groups in Potchefstroom and Rustenburg, in the western Transvaal (Bhana and Brain 1990), where they have maintained a high rate of consanguineous and endogamous marriages. A founder effect may thus have operated in this group to produce the unusual mutation distribution. The numbers of chromosomes studied are, however, small and the significance of the differences is difficult to determine. This group is disproportionately represented in the thalassaemia major families, which may be due in part to a high rate of consanguineous marriages, but also to a higher carrier frequency for ß thalassaemia in this subgroup, although this still requires verification.

While the 619bp deletion and IVS1nt1 mutations appear to occur predominantly in Moslems in South Africa, the codon 41/42 mutation occurs predominantly in Hindus. In the Hindu Gujerati, it occurs at the same frequency as the IVS1nt5 mutation (40%). The presence of the codon 41/42 mutation is responsible for the different haplotype distribution noted in the Hindu Gujerati (Section 5.2.3.2), particularly as it occurs on three different haplotypes in this group. The samples are, however, very small. The

mutation was also absent in two Moslem groups studied by Parikh *et al.* (1990), but appears more widespread in the groups studied by Varawalla *et al.* (1991b, 1992). It also occurs at high frequency in Burma (Brown *et al.* 1992), Thailand (Fucharoen *et al.* 1989, Laig *et al.* 1989, Thein *et al.* 1990, Laosombat *et al.* 1989), Taiwan (Chiou *et al.* 1993) and China (Kazazian *et al.* 1986b, Chan *et al.* 1987, Zhang *et al.* 1988, Huang *et al.* 1990, Liang *et al.* 1994), and at lower frequency in Malaysia (Yang *et al.* 1989) and Indonesia (Lie-Injo *et al.* 1989).

The codon 8/9 (+G) mutation has not been found in South African Indians, which may have been predicted from the rarity of haplotypes 7 and 16 on β^T chromosomes (Section 5.2.3.2). The mutation was found predominantly in Pakistan, Punjab and Sindh (Varawalla *et al.* 1991b, 1992), regions not represented in the South African sample, though it has been reported in Gujerat at relatively low frequency (Parikh *et al.* 1990, Varawalla *et al.* 1992). As the sample sizes were relatively small, it is possible that it does occur at low frequency in South Africa. The mutation also occurs in Malaysia (Fucharoen *et al.* 1990b), Bulgaria (Efremov *et al.* 1992), the United Arab Emirates (El-Kalla and Mathews 1993, Quaife *et al.* 1994), Azerbaijan (Kuliev *et al.* 1994) and in Turkey (reviewed in Flint *et al.* 1993b).

The rest of the Indian mutations are relatively rare, and it is thus difficult to comment on their distributions. The codon 15 (G→A) mutation has also been reported in Turkey (Aulehla-Scholz *et al.* 1990), the United Arab Emirates (El-Kalla and Mathews 1993, Quaife *et al.* 1994), Azerbaijan (Kuliev *et al.* 1994) and Indonesia (Lie-Injo *et al.* 1989), while the IVS2nt1 (G→A) mutation occurs in Turkey (Diaz-Chico *et al.* 1988, Altay and Gurgey 1992), Egypt (Novelletto *et al.* 1990), the United Arab Emirates (El-Kalla and Mathews 1993, Quaife *et al.* 1994), Azerbaijan (Kuliev *et al.* 1994) and Iran (Quaife *et al.* 1994), Italy (Rosatelli *et al.* 1992a), Yugoslavia (Efremov *et al.* 1992), Bulgaria (Efremov *et al.* 1992) and in the Kurdish Jews (Rund *et al.* 1991). The codon 30 (G→A) mutation has been found in Algeria (Bennani *et al.* 1993) and the codon 30 (G→C) mutation has been found in the Kurdish Jews (Rund *et al.* 1991), Italians (Rosatelli *et al.* 1992a), the United Arab Emirates (Quaife *et al.* 1994) and Indonesians (Lie-Injo *et al.* 1989).

Although only eight of the 19 previously described Asian Indian mutations were found in this cohort, others may be found in due course, as the sample size is increased. Two 'new' Asian Indian mutations are described in this study. The frameshift mutation, codon 44 (-C), was first described in a Kurdish Jew (Kinniburgh *et al.* 1982) and was thought to be exclusive to Kurdish Jews (Oppenheim *et al.* 1988, Rund *et al.* 1991). It has also now been described in Italy (Rosatelli *et al.* 1992a), Albania (Boletini *et al.* 1994), Azerbaijan (Curuk *et al.* 1992), and Tunisia (Fattoum *et al.* 1991). The polyadenylation mutation (AATAAA→AACAAA) has been described in an American Black (Orkin *et al.* 1985), an individual of Malay origin (Fucharoen *et al.* 1990b) and in a Turkish family (Altay *et al.* 1991).

One unidentified ß thalassaemia mutation remains in this survey, even after sequencing the regions of the ß globin gene where the majority of mutations have been found. The proband has a typical transfusion-dependent thalassaemia major. He was a compound heterozygote for the IVS1nt5 mutation and the unknown mutation, determined by both the ARMS technique and sequencing. He had normal sized fragments with the probe/enzyme combinations described in Table 2.5 and his mother, who carried the unknown mutation, was heterozygous at the intragenic *Ava*II/ß site. She appeared haematologically to be typical of thalassaemia minor, though her HbA_2 value of 7.2% was relatively high. There have been a few reports of ß thalassaemia not linked to the ß globin cluster (Semenza *et al.* 1984b, Thein *et al.* 1993), however the limited data in this family were consistent with linkage to the ß globin cluster. The mutation may still be in the unsequenced regions of the gene or in the 5′ promoter region, which includes the TATA box and the proximal and distal transcriptional elements at -90 and -105. Deletions of the 5′ regions of the ß globin gene have been found to be associated with unusually high HbA_2 in the heterozygous state (>6.5%) (Craig *et al.* 1992, Motum *et al.* 1992, Dimovski *et al.* 1993, Motum *et al.* 1993, Thein 1993). Two mechanism have been proposed, one implicating removal of competition for limiting transcription factors and the second proposing that loss of the 5' ß gene promoter removes competition for the locus control region, leading to increased interaction between the enhancer elements in the LCR and the δ genes in *cis*, thus enhancing their expression (Hanscombe *et al.* 1991, Waye *et al.* 1991, Craig *et al.* 1992). Some additional mutations may lie in the upstream locus control region. Some mutations still remain unknown after sequencing this region and the entire ß globin gene (Kazazian 1990, Kazazian *et al.* 1990, Garewal *et al.* 1994).

The 10 ß thalassaemia mutations identified in South African Asian Indians account for 98% of the $ß^T$ chromosomes studied. As a large proportion of these mutations were identified in thalassaemia major patients, it is possible that the mutations which cause milder disease may have been missed or their frequencies underestimated. As an important aim of the mutation analysis was to refine the prenatal diagnosis service, those which cause thalassaemia major are of most concern.

6.2.2 Mutation/haplotype associations

In general, ß thalassaemia mutations are strongly associated with specific haplotypes within an ethnic group, usually on the same framework (Orkin *et al.* 1982a, Kazazian *et al.* 1984b). The Asian Indians were no exception and strong linkage disequilibrium between the ß thalassaemia mutations and the major haplotypes was found in this cohort and in previous studies. The associations were not absolute and vary from 50-100% with the common mutations in this and other studies (Kazazian *et al.* 1984a, Thein *et al.* 1988. Varawalla *et al.* 1992). Further, the mutation could be predicted from the haplotype with an accuracy of 66-100% both in this study and that of Varawalla *et al.* (1992). Generally, there were no differences in mutation-haplotype associations in the different groups studied, irrespective of their regional origins, though the haplotype distributions may vary, as shown in Figures 6.2 and 6.3.

6.2.2.1 IVS1nt5 haplotypes

All three haplotypes associated with the IVS1nt5 mutation, in this study, had the same 3′ haplotype or framework (Fw3). A further six haplotypes occur with this mutation in Asian Indians, three of which were associated with Fw1 (Kazazian *et al.* 1984a, Thein *et al.* 1988, Varawalla *et al.* 1992, Venkatesan *et al.* 1992). Haplotype 1 was the most common in all the studies. None of the haplotypes appeared to be specific to any group, suggesting that they existed prior to the extant religious and linguistic divisions of the Indians.

It appears as if the commonest haplotype in the Hindu Gujerati, haplotype 3 (-+-++ --+), differs from that in the other groups, haplotype 1 (+---- --+), although only 4 IVS1nt5 bearing chromosomes were studied. In previous studies, this haplotype

was only rarely associated with the mutation, 0/23 chromosomes (Thein *et al.* 1988) and 3/148 chromosomes (Varawalla *et al.* 1992). In addition to the unusual Fw2 configuration and general ß thalassaemia mutation and haplotype distribution in the Hindu Gujerati (see Sections 5.2.3.1, 5.2.3.2 and 6.2.1), this provides further evidence for the presence of unusual frequencies of genetic markers in this group and the possibility of a founder effect or genetic drift having occurred.

The IVS1nt5 mutation has also been found in Indonesia on four haplotypes, with two frameworks. Two haplotypes were shared with the Indians, while a third, the common Indonesian haplotype differs from the common Indian haplotype, at the *Hin*dIII/3′ß site only (Lie-Injo *et al.* 1989). This may suggest a common origin, with a subsequent point mutation at the *Hin*dIII site, or recombination 3′ to the gene, or a gene conversion event. The mutation was also found in China (Kazazian *et al.* 1986b, Zhang *et al.* 1988), Malaysia (Yang *et al.* 1989), Kurdistan (Rund *et al.* 1991), Lebanon (Chehab *et al.* 1987b), Turkey (reviewed in Flint *et al.* 1993a), Burma (Brown *et al.* 1992), Thailand (Laig *et al.* 1989) and Algeria (Bennani *et al.* 1993) on haplotypes which were found in India, though the frequencies of the different haplotypes vary. An additional four rare haplotypes were associated with the mutation in Burma (Brown *et al.* 1992). The mutation also occurs at high frequency on a single haplotype on the island of Maewo in Vanuatu (Hill *et al.* 1988), a haplotype shared with the Indonesians, the Chinese and the Indians. It was, however, thought unlikely that the Melanesian and Asian mutations had a common origin as there is lack of evidence for early migration from India to the south-west Pacific and it was thought to have arisen locally (Hill *et al.* 1988), though the origin of Melanesians is uncertain.

The IVS1nt5 (G→C) mutation was thus found associated with at least 15 haplotypes and two frameworks over a geographical area, spanning the Asian continent, from the middle east to south-east Asia, and extending to the Pacific islands and north Africa. It is possible that the IVS1nt5 position is a hotspot for mutation, as G→A and G→T mutations have also been reported, and that recurrent mutation has occurred. However, the geographical distribution, together with shared haplotypes in different groups, argue for one or two common ancient origin(s). The different haplotypes which share a common framework are relatively easily explained by recombination events at the recombinational hot spot and

are likely to have a single origin. The haplotypes associated with the se . framework, not found in this study, may represent the results of a second mutation or a gene conversion event followed by recombination.

6.2.2.2 Codon 41/42 haplotypes

The codon 41/42 mutation/haplotype associations present an even more complex picture. The codon 41/42 mutation was associated with two frameworks, Fw2 and Fw3, and four haplotypes in South African Asian Indians. Two of the haplotypes, 22 (+---- ++-) and 28 (+---- +--), differ only at the *Hin*dIII/3'ß site. Another five haplotypes have also been observed in Asian Indians (Thein *et al.* 1988, Varawalla *et al.* 1992, Venkatesan *et al.* 1992), making a total of nine haplotypes on at least two frameworks. One haplotype has a (--) 3' haplotype and the associated framework is not defined (Venkatesan *et al.* 1992). The haplotype (+---- +-) was commonest in all the studies. It is interesting to note that the single Moslem Gujerati haplotype represents one of the rarer ones. It is difficult to interpret the significance of this finding, as it may represent sampling error. However, it may be the result of genetic drift in this group.

West Malaysia had the same common haplotype as the Indians (Yang *et al.* 1989), while the commonest haplotype in Burma (Brown *et al.* 1992), Thailand (Laig *et al.* 1989), China (Kazazian *et al.* 1986b, Chan *et al.* 1987, Zhang *et al.* 1988) and Indonesia (+---- ++) (Lie-Injo *et al.* 1989) is one of the rarer Indian haplotypes. A further 12 haplotypes were associated with the mutation in south-east Asia, including five on . and two with the undefined Fw, (--) 3' haplotype (Zhang *et al.* 1988, Laig *et al.* 1989, Lie-Injo *et al.* 1989, Brown *et al.* 1992). The codon 41/42 mutation thus appears to be associated with a greater number of haplotypes than the IVS1nt5 haplotype, but appears to have a more limited geographical distribution, including India and south-east Asia, but excluding the Middle East and Pacific islands. The shared haplotypes are again suggestive of a common origin at a time when these populations were in close contact.

The IVS1nt1 (G→T) mutation was found only in association with haplotype 8 (Fw1) in this study and those of Kazazian *et al.* (1984a) and Thein *et al.* (1988). Although this was the predominant haplotype, two other minor haplotypes have been found in Asian Indians,

one of which occurs with Fw3 (Varawalla *et al.* 1992). In Burma, the mutation also occurs predominantly on haplotype 8, with no fewer than eight minor haplotypes (on all three frameworks) (Brown *et al.* 1992), while in Malaysia it occurs on the rarer Fw1 haplotype reported in Indians (+---- ++) (Yang *et al.* 1989). In Indonesia the mutation occurs on a single haplotype that differs from the Indian one at the 3′ *Hin*dIII site only (Lie-Injo *et al.* 1989).

6.2.2.3 619bp deletion haplotypes

The 619bp deletion, found only in Asian Indians, was associated almost exclusively with one haplotype in this and other studies (Kazazian *et al.* 1984a, Thein *et al.* 1988, Varawalla *et al.* 1992). One individual studied by Varawalla *et al.* (1992) has a different haplotype on the same framework. These data are consistent with a single origin for this mutation and a crossover in one individual at the recombination hotspot. This mutation is thus likely to be of more recent origin than the other mutations, having arisen on the Indian subcontinent, probably in the region of Sindh (Varawalla *et al.* 1992).

6.2.2.4 Codon 8/9 haplotypes

The codon 8/9 mutation was found on four haplotypes and two frameworks in Asian Indians, the haplotype (+---- ++) being predominant (Kazazian *et al.* 1984a, Thein *et al.* 1988, Varawalla *et al.* 1992). The mutation occurs in Turkey (reviewed in Flint *et al.* 1993a) and in an Iranian (Wong *et al.* 1986) on the same haplotype.

6.2.2.5 Haplotypes associated with the rare Indian mutations

Some of the rarer mutations found in this study were also described on multiple haplotypes and in different ethnic groups. The codon 30 (G→C) mutation occurs on three haplotypes, all Fw1, in Asian Indians (this study; Varawalla *et al.* 1992), one of which was shared with Indonesians (Lie-Injo *et al.* 1989) and Kurdish Jews (Rund *et al.* 1991). The codon 15 (G→A) mutation was found on five haplotypes, with two frameworks (this study; Varawalla *et al.* 1992), two being shared with Indonesians (Lie-Injo *et al.* 1989). The cap site (+1) mutation occurs on two haplotypes with the same framework, but

appears to be confined to Indians. It accounts for 6% of ß thalassaemia genes in Punjab and may have originated in this region (Garewal *et al.* 1994). Further, the newly described codon 44 (-C) mutation was found on the same haplotype as in Kurdish Jews (Rund *et al.* 1991).

The codon 5 (-CT) mutation, not found in this study, was associated with three haplotypes with two frameworks (Varawalla *et al.* 1992), one of which was shared with Greeks (Kollia *et al.* 1989), while the IVS2nt1 (G→A) occurs on the same haplotype in India (Varawalla *et al.* 1992), Turkey (Diaz-Chico *et al.* 1988), Egypt (Novelletto *et al.* 1990), Lebanon and Italy (reviewed in Flint *et al.* 1993a), and on a second haplotype with the same framework in Kurdistan (Rund *et al.* 1991), Bulgaria and Turkey (reviewed in Flint *et al.* 1993a).

6.2.2.6 The significance of the ß thalassaemia mutation-haplotype associations

The mutation distribution data together with the haplotype data suggest that most of the mutations predate the religious and linguistic subdivisions of the Indian people, as they are not confined to any particular subgroup. Furthermore, as described above, many of the mutations have been found in other Asian, Middle Eastern and even southern and eastern European populations, on different haplotypes with different frameworks, in some cases, but also on shared haplotypes.

While initial studies suggested that each haplotype might indicate the presence of a specific ß thalassaemia allele, it was soon realised that in many cases there was an association of a single mutation with more than one haplotype and also the presence of more than one mutation on a single haplotype (Orkin *et al.* 1982a). It was proposed that different frameworks (polymorphisms within the more limited region of the ß globin gene itself) are clear evidence for multiple origins of a mutation, perhaps in a mutation sensitive region. Multiple haplotypes but with the same framework suggest a single origin for the mutation, with subsequent recombination events at the hotspot between the δ and ß globin genes (Orkin *et al.* 1982a, Wong *et al.* 1986).

The evidence for multiple recombination events at the hotspot is strengthened by finding

that most of the mutation-associated haplotypes have one of the three common 5′ haplotypes found on β^A chromosomes. Further, with the exception of the IVS1nt1 mutation, the predominant haplotypes associated with all the common mutations, and many of the rarer ones, have a (+----) 5′ haplotype, the commonest 5′ haplotype. It is possible that they have arisen, by chance, on these haplotypes as they are the commonest on β^A chromosomes, with each framework (see Table 5.2a), but they may also have resulted from recombination events, so the frequency of 5′ haplotypes approaches that on β^A chromosomes.

The geographical clustering of a number of the mutations and the association of many of the common mutations with at least two frameworks in a limited geographical region seem to favour a limited number of origins prior to divergence of the populations or spread of the mutations by gene flow between the populations, rather than a mutational hotspot theory.

The observation of a mutation on at least two frameworks in a population is more likely to be due to intraallelic gene conversion. A Chi sequence is located in the ß globin gene between codon 2 and IVS2nt16, the region in which most of the mutations discussed are located. Chi sequences are known to promote or initiate recombination in *E. Coli* or λ phages (Matsuno *et al.* 1992).

A number of mutations including β^S in Africa, β^E in Asia and many of the ß thalassaemia alleles have been associated with more than one haplotype in a defined geographical region. It has been argued that this is more suggestive of gene conversion than recurrent mutation, since it would be extremely unlikely that the same β^S mutation arose four times in Africa at more or less the same time, but not in south-east Asia, or that the β^E allele arose twice in south-east Asia, but not elsewhere (Flint *et al.* 1993a,b). Based on the haplotype distribution of the β^S gene in Africa, it has been argued that it is more likely that a single mutation arose in a small genetically isolated population in a region with a number of such populations. The mutation would have reached high frequency on a single haplotype in this population, providing a source for occasional migration to carry the gene into other populations, which might have had different haplotype compositions as a result of genetic drift. The presence of different haplotypes would make it possible for

recombination and gene conversion to spread the mutation onto new haplotypes, which in turn could reach high frequency by natural selection. The proposed gene conversions would be small, not affecting flanking sequences. The β^S gene would thus, initially, be associated with a large number of haplotypes, all equally selectively advantageous. Genetic drift and selection would then operate so that finally a single, probably different, haplotype dominated in each group (Flint *et al.* 1993a,b). A similar mechanism, invoking recombination and gene conversion, could explain the associations of the ß thalassaemia mutations with different haplotypes and frameworks in Asia. However, there is not a single mutation in each group as there is with β^S, perhaps suggesting more prolonged gene flow between the small populations so that the new haplotypes were widely spread. Alternatively, the population may have expanded rapidly so that a number of haplotypes increased in frequency together, although, in general, one haplotype is still dominant in each group.

The high frequency and wide distribution of the IVS1nt5 mutation on the Indian subcontinent, together with its multiple haplotype associations has been seen as evidence for its being a very old mutation (Varawalla *et al.* 1992), particularly as five of the haplotypes occur with one framework. This argument is reinforced by its very wide distribution in Asia, from the Middle East, to India and south-east Asia. It may be older than the β^S mutation in India, estimated to be 4 000-7 000 years old, which has one predominant haplotype and only two other haplotypes associated with it, all with one framework. It also occurs in south-east Asia, where β^S is not found, perhaps suggesting a more ancient origin, when the peoples of south-east Asia and India were in direct contact. Further, its high frequency, in some areas, may suggest a particular selective advantage or an older origin than other ß thalassaemia mutations.

The other mutations appear to cluster into four groups, those only described in Indians, including the 619bp deletion and the cap site (+1) mutation; those shared by Indians and south-east Asian populations, including the IVS1nt1 and codon 41/42 mutations; those shared by Indians and Middle Eastern populations, including the codon 8/9 mutation, IVS2nt1 (G→A) and codon 44 (-C) mutations; and those, in addition to the IVS1nt5 mutation, that appear to be spread across Asia from the middle east to south-east Asia, like the codon 15 and codon 30 (G→C) mutations.

The mutations shared by the Indians and Middle Eastern populations have a distribution similar to that of the Arab-India β^S mutation and it is thus proposed that these mutations, similarly, originated and spread from the Harappan culture of the Indus river valley. It is likely, however, that some of the spread, particularly to areas like Turkey, may have occurred through later interactions, like the Turkish invasion of north India in the eleventh century, which is thought to have spread the IVS2nt1 mutation to Turkey (Varawalla *et al.* 1992).

The mutations which are common to India and south-east Asia are generally associated with a larger number of haplotypes. This may suggest that they are older, so that enough time has elapsed for multiple recombination, and perhaps gene conversion, events to occur. However, their place of origin is more difficult to determine. Political, military and commercial interactions with central, western and south-east Asia, and later Europe, are part of the history of the Indian subcontinent, and are likely to have resulted in gene flow, and thus spread of the alleles (Varawalla *et al.* 1992), perhaps from south-east Asia to India or from India to south-east Asia. Indian sea-traders exploring the coasts of Burma, the Malay peninsular and western Indonesia, as early as the 6th century BC, may have been a possible route of spread of the alleles (Varawalla *et al.* 1992). However, the Thailand and Chinese alleles cannot be similarly explained. Independent origins may be responsible (Varawalla *et al.* 1992), though the number of shared haplotypes and mutations make this less likely. In addition, the codon 41/42 was associated with a rare 3′ haplotype (--) in India (Venkatesan *et al.* 1992) and in China, and this is highly suggestive of a common origin.

The IVS1nt5 mutation may have originated with the other Indian/south-east Asian mutations, and only spread to the Middle East later, perhaps as a result of migrations associated with the spread of the Harappan culture or later interactions, as its frequency in the Middle East is much lower than in south-east Asia. However, high frequencies in the United Arab Emirates suggests that this and other Asian Indian mutations were introduced by population migration across the straits of Hormus via Oman from the region previously known as Baluchistan, now in Iran, Afghanistan and Pakistan (Quaife *et al.* 1994).

Although it is difficult to date the mutations accurately, it has been suggested, based on studies in Mediterranean populations, that the ß thalassaemia mutations are of relatively recent origin, because they occur on a subset of haplotypes and ß globin frameworks (Orkin *et al.* 1982a). This appears to be true also for the Asian Indian mutations. Further, the different ß thalassaemia mutations in different parts of the world suggest that they arose after the Asian and European populations split. If the mutations arose when the founder population left Africa, one would expect shared mutations, as is found with other genetic markers, including the ß globin haplotypes (Flint *et al.* 1993a,b). It is suggested that ß thalassaemia mutations arose at similar times throughout the world and reached high frequencies by natural selection, though genetic drift and population migration were also important, particularly for bringing the mutations into contact with new pools of haplotypes and promoting further haplotype diversity by gene conversion and recombination (Flint *et al.* 1993a,b).

Further evidence for the theory of a limited number of ancient common mutations comes from the codon 44 (-C) mutation, found in an Indian in the present study, on the same haplotype as in Kurdish Jews (Rund *et al.* 1991). The mutation has also been found in other Middle Eastern and Mediterranean countries (see Section 6.2.1), although no haplotype data are available. Kurdish Jews constitute a small highly inbred group, but still have mutational diversity, with 13 ß thalassaemia mutations having been found. Some of these mutations occur in other populations, including those from the Mediterranean and India. Four, possibly five, of the other mutation have been considered unique to the Kurdish Jews, including three of the five most prevalent and including the codon 44 mutation. Furthermore, the codon 44 mutation occurs predominantly in central Khurdistan, the area from which the unique Kurdish mutations are thought to have arisen (Rund *et al.* 1991). Although a second origin for the mutation cannot be excluded, the finding of the mutation in the present study and in other Middle Eastern countries is more consistent with a common ancient origin with subsequent founder effect and natural selection having operated in the geographically isolated Kurdish Jewish population to increase the local frequency. Furthermore, the IVS1nt5 mutation, as well as the codon 30 (G→C) and IVS2nt1 (G→A) mutations, occur on the same haplotypes in Kurdish Jews (particularly those from eastern Turkey) as the common Indian ones (Rund *et al.* 1991), further suggesting an ancient common gene pool, although the frequencies differ

considerably. A recent study on the Mestizo Mexicans has shown a number of typically Asian Indian mutations on identical haplotypes to those in their presumed parent population. In addition, they carry another of the mutations considered unique to the Jews of central Khurdistan [-28 (A→C)] on the identical haplotype (Economou *et al.* 1991).

Despite their ß thalassaemia mutational diversity, the G6PD deficiency in Kurdish Jews is almost entirely the Mediterranean variant from all three geographical regions, suggesting that this variant may predate their exile. It has been suggested that the G6PD variant may be due to an ancient mutation and founder effect, while the thalassaemia mutations are largely unique and may have arisen through local mutation events before or after migration to Khurdistan (Oppenheim *et al.* 1993). It is more likely that the ß thalassaemia mutations also originated from an ancient common gene pool, and have increased in frequency in the Kurdish Jews due to founder effect and natural selection as discussed above.

Further, the poly-A mutation, the second new Indian mutation, was found on the same haplotype in the Indian in this study as in an American Black (Orkin *et al.* 1985), again suggesting a possible common origin, particularly as there is considerable genetic admixture in the American Black population. The mutation has also been described in an individual of Malay origin, but on a different haplotype and framework (Fucharoen *et al.* 1990b) and in a Turkish family, where no haplotype was determined (Altay *et al.* 1991).

While the haplotypes provide valuable information for some mutations and reinforce the idea that some mutations have arisen once, probably relatively recently, they do little to resolve the extent to which recurrent mutation, interallelic gene conversion, recombination, mutant gene migration and other factors have contributed to the more widely distributed ß thalassaemia mutations with multiple haplotype associations. In general, it was thought that the same framework suggested recombination at the recombinational hotspot (Antonarakis *et al.* 1982a), while different frameworks in different ethnic groups suggested multiple, independent origins (Wong *et al.* 1986).

In the populations of India and south-east Asia, the evidence seems to point towards common ancient origins of the mutations with subsequent recombination and limited gene

conversion to explain the multiple associated haplotypes. However, differences may have occurred between more closely related groups in more recent history due to founder effect and genetic drift as the groups have become geographically, religiously and linguistically isolated, thus compounding the problem of determining the origins of the mutations.

CHAPTER 7 - MEDICAL IMPORTANCE OF THE STUDY OF THE HAEMOGLOBINOPATHIES IN SOUTH AFRICAN INDIANS

ß thalassaemia is the commonest clinically significant haemoglobinopathy in South African Indians, and it appears to represent a greater disease load than expected. Sickle cell anaemia also occurs, though patients appear to be under-represented in haemoglobinopathy clinics (see Section 3.2.1.3), possibly because of the mildness of the disease.

While the severe forms of α thalassaemia, namely HbH disease and Hb Bart's hydrops fetalis, are extremely rare in South African Indians, the mild forms, -α/αα and -α/-α are common. The clinical importance lies in distinguishing individuals with these genotypes from those who are ß thalassaemia heterozygotes and those with iron deficiency anaemia.

At present, no formal haemoglobinopathy screening programme to detect 'at-risk' couples exists, though a local prenatal diagnostic service has been available since 1984, and before that a few patients were referred to overseas centres. Thus, it is necessary to assess the need for and value of a haemoglobinopathy screening programme in South African Indians. Further, the data on the families who have used the prenatal diagnostic service need to be analysed and the service itself evaluated.

7.1 Screening for the haemoglobinopathies in South African Indians

In order to assess the viability of a haemoglobinopathy screening programme, the principles of screening need to be reviewed, well-established programmes need to be analysed, a technical approach suitable for the local population has to be defined, a cost/benefit analysis must be carried out, and other social and cultural problems specific to the local population need to be identified.

The term 'haemoglobinopathy' will, unless otherwise stated, be used in the rest of this chapter to refer to the main clinically-significant haemoglobinopathies in South African Indians, namely those involving mutations of the ß globin gene.

7.1.1 Principles of screening

A disease for which a screening programme is established must represent a substantial health problem, the prevention of which would be a worthwhile goal. The screening programme should accurately identify the majority of 'at-risk' individuals, with a minimum of false positive results, and the detection of these individuals should result in an improved outcome for the affected families and society as a whole, so that people with a genetic disease can live and reproduce as normally as possible. The technical procedures used should be safe, acceptable to the patient, simple, accurate and relatively inexpensive. The benefits of the programme should outweigh the disadvantages. The effectiveness of the programme requires a comprehensive service, equally accessible, in terms of information and facilities, to the entire 'at-risk' community, and constant monitoring, with a network of collaboration for information and laboratory control. In general, the programme should be established as an integral part of the primary health care facilities; an association of affected families or individuals should be established; and the patients identified should be educated, counselled, treated and offered support, where appropriate. According to the WHO a control programme for a genetic disorder should be "an integrated approach combining the best possible patient care, with prevention through community information, carrier screening and counselling, and the availability of prenatal diagnosis" (reviewed in Westring and Grand 1973, Kaback 1982, Modell and Kuliev 1993).

In addition there are other ethical and legal considerations. Prior to testing, individuals should be appropriately informed, participation should be voluntary, and individual privacy must be protected. This is a point of potential conflict as there are implications to the family of a positive result. Possible legal issues relate to the disclosure of the risks of the procedure, failure to screen, mislabelling of samples, laboratory error, poor quality control and informed consent (Westring and Grand 1973, Kaback 1982).

7.1.2 Screening for the haemoglobinopathies

In general, the haemoglobinopathies have been shown to be good candidate diseases for the establishment of screening programmes. For genetic diseases, in general, primary

prevention is not possible and treatment is burdensome, expensive and unsatisfactory. For the haemoglobinopathies, however, carriers can be detected by a procedure which requires a single blood sample and is safe and generally acceptable. Carriers can be counselled on their risks and prenatal diagnosis can be offered to 'high-risk' couples. A prenatal diagnosis programme can reduce indiscriminate abortion and the associated guilt, decrease anxiety and allow for family planning. In view of the high costs of thalassaemia management, community control programmes by prospective heterozygote detection, education and fetal diagnosis are seen to be vital even in developing countries. The low costs of setting up and running such programmes are trivial compared to the financial and social implications for the community of supporting individuals with the disease. As treatment for the disease improves, so the costs increase, because of prolonged survival of the patients, thus a preventive programme becomes essential to contain the costs (WHO Working Group 1982, WHO 1983, Modell 1985).

The WHO working group for the community control of hereditary anaemia proposed that the strategy for prospective control of thalassaemia should be planned on three levels: sensitisation and involvement of the population, identification of 'at-risk' couples, and counselling. Such control programmes should also include optimal care of existing patients and prenatal diagnosis (WHO working group 1982, WHO 1983, Modell 1985).

An educated population is one of the most vital components of a successful screening programme. Many problems arise from inadequate information and there is therefore a strong case for improving the community's and the medical profession's knowledge about the disease, the genetic risks and the role of prenatal diagnosis. In Sardinia and Cyprus, where the two most successful programmes have been introduced, national programmes were initiated to educate the population. The high incidence of disease, together with the increasing survival of homozygotes and the increasing costs of therapy, had virtually made preventive programmes imperative. The success of these programmes is evidenced by a dramatic decline (of over 90%) in homozygote births (Angastiniotis *et al.* 1986, Cao *et al.* 1989b).

The programmes in Cyprus and Sardinia were based on the general principles outlined by the WHO working group for the community control of hereditary anaemias (WHO 1983).

The entire population was involved. The local community, voluntary associations and parent associations participated in community education and counselling. Meetings were set up with community leaders to encourage co-operation and obtain advice. Medical teams were consulted to discuss the approaches to the programmes, formal education was introduced into the school curricula and mass media education and information leaflets were written and widely circulated. Provision was made for follow-up and assessment of individuals. Specially trained counsellors were involved in individual heterozygote counselling, which included reassurance of carriers, information on the risks to their children of having the disease, the availability of prenatal diagnosis and discussion of treatment, where relevant.

These programmes were facilitated by having a small concentrated population, a high literacy rate, a relatively high standard of living, a low birth rate and high standards of health care. In addition, because of the high incidence of the disease, most individuals had some awareness of it (Angastiniotis *et al.* 1986, Cao *et al.* 1989b, 1990). It has been shown that uptake of screening and prenatal diagnosis requires informed couples and experience of the disease (Anionwu *et al.* 1988).

7.1.3 Technical considerations

There are a number of methods which can be used to screen for ß thalassaemia heterozygotes (reviewed in Rowley 1976). Abnormal red cell morphology may be studied, though this is time-consuming, variable and requires expert evaluation. Reduced osmotic fragility is a cheap and simple screening test but it is non-specific for all hypochromic anaemias (Kattamis *et al.* 1981). The rate of globin chain synthesis appears to be the best measurement to diagnose ß thalassaemia trait, the latter being characterised by a reduced ß:α synthetic ratio. However, the method is too elaborate and time-consuming for routine screening. Microcytosis and/or hypochromia appear to be the best criteria for identifying subjects requiring determination of HbA_2 levels (Tammis-Hadjopoulos *et al.* 1977). HbA_2 levels, raised in individuals with ß thalassaemia trait, require electrophoresis, elution and absorption spectrophotometry for accurate quantitation. Column chromatography may be used, but quantitation by densitometry is unsatisfactory. This technique is relatively time-consuming, expensive and requires some experience (Schmidt *et al.* 1975).

Thus, an ideal method for ß thalassaemia trait screening is still not available, though in countries where ß thalassaemia is common, screening programmes have generally relied on using a low MCV as an indicator to quantitate the HbA_2.

In South African Indians microcytosis and hypochromia have been shown to be extremely common, mainly due to iron deficiency and α thalassaemia. Only a minority of these cases, a maximum of 5%, are due to ß thalassaemia heterozygosity. If one were to use a low MCV or MCH for screening, one would have to quantitate HbA_2 on about 30% of those screened. In addition, while homozygous sickle cell anaemia is a relatively mild disease in Indians, the compound heterozygotes for ß thalassaemia and $ß^S$ have a relatively severe disease and it would be unacceptable to miss them in a screening programme. Carriers of the $ß^S$ gene generally have normal haematology and are thus only detectable on Hb electrophoresis. The $ß^E$ allele occurs at a low frequency in Indians, but may also cause relatively severe disease in homozygous form or, possibly, together with ß thalassaemia. Carriers of the $ß^E$ allele have microcytic indices (Fairbanks *et al.* 1979) and would thus be detected with electrophoresis or MCV screening.

As there are many populations in which α and/or ß thalassaemia as well as iron deficiency occur together, there have been numerous attempts to design formulae which distinguish between these conditions on the basis of red blood cell indices (England and Fraser 1973, Mentzer 1973, Srivastava 1973, Shine and Lal 1977, Bessman and Feinstein 1979, Green and King 1989). Most are based on the principle that in thalassaemia the MCV and MCH are disproportionately low with a high RCC, whereas in iron deficiency the RCC, MCV and MCH are all proportionally low. Quantitative anisocytosis or the measurement of red cell volume dispersion or red cell distribution width (RDW) is also used. The RDW is narrow in thalassaemia and wide in iron deficiency (Bessman and Feinstein 1979, Green and King 1989). Free erythrocyte protoporphyrin (FEP) (Stockman *et al.* 1975), zinc protoporphyrin (Han *et al.* 1990) and three dimensional plots of red blood cell indices (Han and Fung 1991) have also been used in an attempt to distinguish thalassaemia from iron deficiency. These discriminant functions have diagnostic efficiencies between 50 and 82% in uncomplicated cases, but perform worse in combination cases (Bentley *et al.* 1989). Further, in mixed cases with α or ß thalassaemia, iron deficiency predominates, a major problem for screening, especially as in communities where iron deficiency is

common, thalassaemia carriers are not exempt (Economidou *et al.* 1980).

Two alternative screening protocols are possible in South Africa using currently established techniques. Microcytosis and/or hypochromia could be screened for using an electronic cell counter and then HbA_2 quantitated on those with an MCV<80fℓ and/or an MCH<27pg. In addition, qualitative electrophoresis would have to be done on all samples for ß globin variants. With this protocol, no diagnosis would be made in the majority of microcytic hypochromic individuals, and it may thus be necessary to assess the costs of establishing the causes of microcytosis and hypochromia in these individuals.

Alternatively, quantitative electrophoresis could be done on all samples to determine the HbA_2 levels. Those with raised HbA_2 could have a follow-up electronic cell count to confirm that the indices are consistent, in concordance with WHO (1983) recommendations that the diagnosis in heterozygotes be confirmed by at least two methods because of the clinical significance of the results. The other ß globin variants would be detected simultaneously on the electrophoretic strips. In both protocols, citrate agar electrophoresis would be necessary to confirm the identity of the electrophoretic variants detected on cellulose acetate.

Theoretically, some cases may be missed by both screening methods. In Mediterranean countries, individuals with raised HbA_2 and normal RBC indices were shown to have coinherited ß thalassaemia trait and either $-\alpha/-\alpha$ or $-\alpha/\alpha\alpha$, which normalises the chain imbalance and improves the haemoglobinisation of the red cells (Kanavakis *et al.* 1982, Melis *et al.* 1983, Rosatelli *et al.* 1984). On this basis, it was suggested that in areas where both α and ß thalassaemia were prevalent, MCV/MCH screening may miss a significant proportion of those with ß thalassaemia trait, whereas HbA_2 would still be raised and appears to be more sensitive for carrier detection (Globin Gene Disorder Working Party of the BCSH General Haematology Task Force 1994). HbA_2 levels do not appear to be normalised (Kanavakis *et al.* 1982, Melis *et al.* 1983, Cao *et al.* 1984), even though MCV's up to 90fℓ have been found in $-\alpha/-\alpha$ heterozygotes (Kanavakis *et al.* 1982). In the present study, seven obligate ß thalassaemia carriers who had coinherited a $-\alpha/\alpha\alpha$ genotype and two with a $-\alpha/-\alpha$ genotype all had microcytic hypochromic indices, although the MCV and MCH were significantly higher in obligate ß thalassaemia carriers

with -$\alpha/\alpha\alpha$ than with $\alpha\alpha/\alpha\alpha$ ($p<0.01$). No significant differences in the HbA_2 levels were however observed between the two groups. An MCV less than 75fℓ or an MCH less than 25pg are recommended as the cut-offs for screening for ß thalassaemia (WHO 1983). These figures were based on studies such as that of Walford and Deacon (1976), where no ß thalassaemia heterozygote had an MCV above 70fℓ, whereas in α thalassaemia 64% had an MCV between 70 and 80fℓ. In the present study these figures are true for ß thalassaemia heterozygotes who are $\alpha\alpha/\alpha\alpha$, but not for those with coinherited -$\alpha/\alpha\alpha$. Thus, in a population such as the South African Indians, with a high incidence of α thalassaemia, it seems as if cut-offs of an MCV<80fℓ or and MCH<27pg would be more sensitive. Such screening would be unlikely to miss many heterozygotes, especially if one considers that in the Moslem Gujerati, in whom the α and ß thalassaemia allele frequencies are highest, only 1/500 individuals would have ß thalassaemia trait together with -α/-α and only a small percentage of these would be likely to have normal red blood cell indices.

Conversely, there may be individuals with ß thalassaemia variants who have abnormal red blood cell indices and normal HbA_2 levels, including δß thalassaemia carriers or 'silent carriers' with coinherited δ thalassaemia, who may constitute up to 10% of ß thalassaemia carriers in Greece (Kattamis *et al.* 1979, Marsh and Koenig 1982, Cao *et al.* 1989b). These individuals tend to have HbA_2 levels in the high normal range, in contrast to α thalassaemia carriers who may also have abnormal indices, but tend to have relatively low HbA_2 levels (Paglietti *et al.* 1985). In addition, HbA_2 levels in normal iron deficient individuals are reduced compared with those who are non iron deficient, though there is no correlation between the decrease and the degree of anaemia or hypochromia (Steiner *et al.* 1971, Kattamis *et al.* 1972). However, iron deficient ß thalassaemia heterozygotes still have levels which are always within the carrier range (Galanello *et al.* 1981, Mehta and Pandya 1987). In 52 obligate ß thalassaemia heterozygotes in the present study all had HbA_2 levels above the normal range. Thus, the incidence of normal HbA_2 ß thalassaemia carriers is probably low in South African Asian Indians, consistent with a study in Punjab in north India, where only 3% of ß thalassaemia heterozygotes had normal HbA_2 (Dash 1988). It is likely that the normal HbA_2 ß thalassaemia heterozygotes may have a particular ß thalassaemia mutation or a coinherited δ thalassaemia mutation which, once identified, could be tested for in 'high-risk' couples where one is a ß thalassaemia

heterozygote and the other has equivocal haematological indices. The mean HbA_2 of the $-\alpha/\alpha\alpha$ heterozygotes with normal ß globin genes was 2.9 ± 0.4 and of the $-\alpha/-\alpha$ homozygotes 2.8 ± 0.3, neither of these being significantly different from the normal.

It appears that both approaches to screening would detect the majority of individuals, if not all, 'at-risk' of producing offspring with a severe clinical phenotype. Recent studies have shown a correlation between the haematological phenotype, particularly the MCV and MCH, in ß thalassaemia heterozygotes and the ß thalassaemia mutation present (Rund *et al.* 1990, Rosatelli *et al.* 1992b, Rund *et al.* 1992b). The HbA_2 level does not appear to be related to the ß thalassaemia mutation, in general, though a number of variants associated with high HbA_2 levels have been described (Craig *et al.* 1992, Motum *et al.* 1992, Dimovski *et al.* 1993, Motum *et al.* 1993). Thus, the wide variation of haematological characteristics in heterozygous ß thalassaemia is related not only to co-existent α thalassaemia and iron deficiency, but also to the type of mutation present (Rund *et al.* 1990, Rosatelli *et al.* 1992b, Rund *et al.* 1992b). The assessment of the phenotype associated with the different mutations present in the South African Indians may further assist in evaluating the most reliable method for screening.

7.1.4 Cost/benefit analysis

The costs of both screening protocols for the detection of carriers of ß thalassaemia and ß globin variants have been estimated using techniques currently operational in the reference laboratories in South Africa. The details are shown in Appendix V(A). The laboratory costs to identify one 'at-risk' couple are virtually identical using the two protocols (± R17 000). If one were to follow up all the microcytic hypochromic individuals identified, do iron studies and DNA analysis to determine the causes of their microcytosis and hypochromia, and counsel them, this would raise the costs of the first protocol enormously. In contrast, in the second protocol, only those individuals who were identified as having raised HbA_2 levels, the ß thalassaemia heterozygotes, would require counselling.

Counselling and follow-up of microcytic and hypochromic individuals, once detected, would be important, particularly where a screening programme is established as part of

a primary health care programme and one is attempting to offer a comprehensive service. Firstly, iron deficiency anaemia may be clinically important and require treatment; secondly, individuals with α thalassaemia should be made aware of the cause of their microcytosis and hypochromia so they are not subjected to unnecessary investigation or iron replacement therapy in the future. Since the -- chromosome is rare in South African Indians, individuals with a $-\alpha/\alpha\alpha$ or $-\alpha/-\alpha$ genotype need to be reassured of their low risks of a child with a major haemoglobinopathy such as HbH disease or Hb Bart's hydrops fetalis. The diagnosis of α thalassaemia, however, requires DNA studies which are highly specialised and expensive. In addition, the anxiety associated with recalling microcytic hypochromic individuals for follow-up and counselling requires consideration. Although, in theory, such anxiety and fears could be allayed by counselling, there would still be a period of anxiety while the investigations were proceeding.

It has been suggested that in populations with a high rate of microcytosis, such as the South African Indians, couples should be screened and only those recalled where both spouses are microcytic or hypochromic (Modell and Modell 1990). However, in the local population, it would still mean that 10% of couples would be recalled for further investigation.

It appears that it would be more cost-effective to follow the second protocol, based on HbA_2 determinations, in the South African Indian population. The costs for detection of an 'at-risk' couple [Appendix V(A)] have been estimated assuming random mating. The frequencies for the Moslem Gujerati have been used as these were most complete. A higher incidence of homozygotes has been noted due to the high rate of consanguineous marriage and intra-caste marriage. Thus, one may find 'at-risk' couples in the Moslem Gujerati subgroup more often than the 1/918 expected from the allele frequencies, reducing the cost of finding each couple. In the other subgroups, however, the allele frequencies may be slightly lower and the costs of finding 'at-risk' couples may be slightly increased. Again consanguinity and intra-caste marriages may alter this somewhat. Thus, a number of assumptions, which may not always be valid, have been made in the calculations, but they are unlikely to affect the figures to any large degree.

The testing procedure would identify those 'at-risk' of having children with thalassaemia

major, sickle cell anaemia, HbE disease and compound heterozygosity for ß thalassaemia and $ß^S$ or $ß^E$ or other electrophoretic variants. Not all 'at-risk' couples would necessarily elect to have prenatal diagnosis, though they would be fully informed of their risks. Those 'at-risk' of having children with sickle cell anaemia would be faced with a difficult choice considering the benign nature of the disease in most patients. The costs for each 'at-risk' couple to have an average-size family, without an affected child, have been estimated to be about R9 650. The calculations appear in Appendix V(B).

The major haemoglobinopathies in South Africa occur in a minority ethnic group and at relatively low frequency compared to some areas of the world, thus the total number of patients with the disease is relatively small. The disease still represents a substantial health problem in terms of the costs of therapy, the minimum annual cost of treatment for one child with ß thalassaemia having been estimated at R78 500. This excludes the cost of specialist and other staff who look after these patients as well as the cost of making the initial diagnosis. The details of the cost estimation are shown in Appendix V(C).

The costs of the screening and prenatal diagnosis programme have been compared to the cost of treatment for thalassaemia major. The details are shown in Appendix V(D). For the calculations it has been assumed that all children with haemoglobinopathies are similarly treated. This is not necessarily true, particularly with regard to sickle cell anaemia where hospitalisation and medication are rarely required. The detection of couples 'at-risk' for children with sickle cell disease may not be entirely justified in a cost benefit analysis, because of the mild nature of the condition. The main aim of a screening programme would be prevention of the major haemoglobinopathies, which have significant associated morbidity. The increased cost of detection of some couples 'at-risk' of a less severe disease can be justified in this context. It could be argued that couples 'at-risk' for children with sickle cell anaemia, albeit mild, have the right to full information and free choice with regard to prenatal diagnosis for affected pregnancies. Further, the possibility of consanguineous marriages, which may raise the number of 'at-risk' couples has been excluded.

Despite the relatively low frequency of the disease, the costs for treatment of the existing patients are extremely high compared with the costs for prevention of the disease. As

treatment improves so the costs escalate, as more patients survive and require treatment for longer periods and the treatment regimens become more intensive. There are other additional costs for families with affected children which are difficult to evaluate. Parents are often required to take time off work to take children to the hospital for their transfusion therapy and regular monitoring. The burdens of a chronic disease on a family are heavy. The costs for detection of an 'at-risk' couple and the prevention of the birth of an affected child in their family are estimated at 34% of that of one year of treatment for a child with thalassaemia major or 1.4% of the lifetime cost of treatment of one child with thalassaemia major [Appendix V(D)]. At the same time, the child and family are saved much emotional turmoil and psychological and functional morbidity.

At a population level, the cost of screening the entire South African Indian population between the ages of 20 and 50, with the prevention of the birth of children affected with thalassaemia major in all of them, togetner with the achievement of a family size of 2.4 children for each, would be equivalent to the lifetime cost of treating 3.5 thalassaemia major children or the costs of treatment for the existing patients for one and a half years [Appendix V(D)].

In Cyprus where the ß thalassaemia carrier rate is 1/7, the cost of prevention of the disease for 1984 was the same as the cost of treatment for the existing patients for an 8-week period (Angastiniotis *et al.* 1986). It was estimated that if no preventive programme was instituted, the total medical budget of Cyprus would be doubled within 20 years and 40% of blood donations would be for thalassaemia major (WHO 1983). Similarly in a mixed population in Quebec the cost per case prevented was slightly less than the cost per patient treatment year or about 4% of the treatment cost incurred in 25 years for an affected individual (Scriver *et al.* 1984).

Although the cost-benefits of the screening and prevention programmes have been demonstrated in other countries, particularly where thalassaemia is common, the current study emphasises that such programmes are beneficial even where the disease has a relatively low frequency.

7.1.5 Targeting a screening programme

In the calculations above, the target population has been assumed to be Asian Indian adults of childbearing age (20-50 years), with the aims being to identify 'at-risk' couples (where both are heterozygotes for a major haemoglobinopathy). There has been much debate on the appropriate time to screen individuals for recessive disease.

Senior school pupils represent an ideal target for screening in that they are accessible and generally close to their families. If a heterozygote is detected, it is relatively easy to test other members of the family 'at-risk'; testing of adolescents may present other problems, however. Carrier status may lead to stigmatisation and loss of self-esteem in individuals at a vulnerable age. In addition, as the information is not of immediate relevance, it may not be remembered and individuals may have to be tested again at a later stage (Westring and Grand 1973, Kaback 1982). It appears that screening should be focused on individuals for whom a positive result has high significance.

Young individuals or couples immediately before marriage or engaged in reproduction represent the best target group (Motulsky 1973). Factors shown to predict screening success (as judged by higher knowledge after counselling or transmitting information to other family members) include plans to have children in the near future, younger age, higher education level, knowledge of having a trait personally or in the family, and the spouse's carrier status (Lipkin *et al.* 1986, Loader *et al.* 1991a). Screening should be appropriately targeted to individuals or couples who are fully informed and know all the available options open to them. Couples offer an advantage over individuals in that if an equivocal result is obtained in one spouse and the other is normal, the couple can be reassured and thus undue anxiety is alleviated and further detailed investigation is avoided. In rare cases where the second spouse is a heterozygote, further testing may be necessary, perhaps involving globin chain synthesis or detailed DNA studies (Akerman *et al.* 1990, Kazazian 1990).

'At-risk' couples should be identified at least before they reproduce, so that the birth of affected children can be prevented. Premarital couples would be an ideal target, as they would have the widest choices available to them, including the choice not to marry,

though genetic information rarely affects the choice of a marriage partner (Modell 1990). They are a difficult group to access, though it has been achieved in Cyprus where there is virtually a single marriage system, through the church. After marriage, screening should optimally be offered prior to conception as part of a primary health care programme, although access to couples may be difficult, unless the population is highly informed.

'At-risk' couples may choose not to reproduce, to 'take a chance', to reproduce only with prenatal diagnosis and selective abortion, to reproduce with a substitute partner (e.g. artificial insemination by donor) or to divorce and find another marriage partner, the latter two being unpopular choices. Prenatal diagnosis allows couples to ensure healthy children and to achieve the family size they want. First trimester prenatal diagnosis has increased the uptake rate considerably and thus couples need to be informed early enough to take advantage of this technique. Prior to the availability of prenatal diagnosis, many couples avoided conception and 70% of accidental pregnancies were terminated indiscriminately. Some couples opted to 'take a chance', at least once, but stopped if they had an affected child, thus frustrating their ambitions for a normal family (Modell *et al.* 1980, Modell 1990).

As married couples are difficult to access, screening through antenatal clinics has been proposed, especially as a full blood count is often done as a routine investigation. Most women would present too late for first trimester prenatal diagnosis, however, and some too late for prenatal diagnosis at all. In addition, the laboratory has little time to confirm results and often hurried decisions are made because of deadlines. The emotional involvement of a couple with a pregnancy may increase the difficulties of making an objective decision (Modell 1990, Petrou *et al.* 1990).

In the South African Indian population, young couples would be the preferred target group. Although the population is a heterogeneous one, religion and culture continue to play a significant part in the lives of Indians (Mistry 1965). As almost all couples are married with some type of religious ceremony, premarital screening with the co-operation of these religious institutions may be possible in this group. Such co-operation would also increase the acceptability of the screening programme. Screening through primary health

care facilities may also be possible, though it would have to be combined with screening through general practitioners, who often represent a primary health contact. Screening of couples would be particularly important in the local context because of the high incidence of iron deficiency and α thalassaemia which may interfere with the diagnosis.

7.1.6 Social considerations

Part of the reason for the success of screening programmes in countries such as Sardinia and Cyprus was the involvement of the entire country. Programmes were organised at the national level and targeted the entire population. A similar screening programme introduced in the United Kingdom and aimed at Cypriots and Asians, mainly Pakistani Moslems, had rather less success initially (Modell *et al.* 1984). A number of reasons were proposed, including the groups being small and scattered through a large population not 'at-risk', thus making access and a nationwide level of awareness among 'at-risk' communities difficult. Communication problems in first-generation Pakistani Moslems were identified related to culture and religion. It was shown that adequate delivery of genetic counselling had to be provided at home by a woman of the same ethnic group who also shared the Moslem religion (Modell *et al.* 1984, Modell 1985). Where minority groups are involved, it is particularly important to involve the whole population in education programmes so that screening is accepted by the minority community and not seen as a method of stigmatising the group by the larger population (Motulsky 1973).

The South African Indians constitute such a minority group. Although the total population is relatively small (947 000 individuals), over 95% reside in Natal and the southern Transvaal (Department of National Health and Population Development 1990). The majority lives in large 'Indian' townships near to the major city centres of Durban or Pietermaritzburg in Natal; or Johannesburg or Pretoria in the Transvaal, or in the cities themselves. Access to screening would not be a major problem for the majority of the population. The smaller groups cannot be ignored, however, as in outlying areas kinship networks often operated most strongly in promoting immigration (Bhana and Brain 1990). Such groups often originated from one village in India and have maintained highly endogamous marriage customs and may represent pockets with a higher disease frequency. The Moslem Memon community of Potchefstroom in the south-western Transvaal may

represent such a group.

The majority of local Indians are third or fourth generation South Africans, but despite considerable environmental changes, novel economic circumstances and wide contacts with peoples of other cultures and traditions, traditional cultures and religions have been widely maintained (Mistry 1965). Involvement of the community would therefore be vital in designing an acceptable screening and prenatal diagnosis programme. Further, counsellors from the different subgroups would have to be trained in order to deliver effective genetic counselling.

The importance of ethnocultural sensitivity in addressing the health care needs of various populations has been emphasised by Wang and Marsh (1992). Some of the differences between Western thought and the Asian ethnocultural perspective may compromise the delivery of effective genetic services, the main differences being the idea of collective autonomy in Asian cultures in contrast to individual autonomy in Western culture, which may influence the decision-making process. The Western medical model is complementary to Asian expectations of authority and may result in readily agreed upon consent to recommendations due to the respect of authority valued by Asians, rather than personal feelings. Direct, structured and concrete recommendations for a specific course of action are expected from an authority figure and the perceived ambiguous nature of non-directive counselling may result in inappropriate decisions due to incongruent expectations of a presumed authoritarian model (Wang and Marsh 1992).

In the Asian family, the extended family unit is the fundamental social unit, and thus, contrary to Western perspective, the individual's behaviour has different importance and consequence. Personal actions reflect not only on the individual, but also on preceding and future generations and personal autonomy is not a fundamental factor in decision-making. Individuals are encouraged to submerge behaviour and feelings that may disrupt family harmony to further the welfare and reputation of the family. Acts of individuality and independence are dissuaded, and harmonious familial relationships are maintained by fostering dependence and conformity in the family. The family is seen as a team, with role behaviour determined by generation, age and sex. The father's position is one of undisputable leadership and unquestionable authority. Further, in Asian culture,

dishonourable situations such as infertility and ill health diminish not only the individual, but the whole family. Such perceived failures are viewed as a reflection of the family's collective conscience and assistance outside the family is considered a public admission of shame and is not condoned (Wang and Marsh 1992). These concepts can be a major impediment to Asian patients accepting screening or counselling, and these have to be introduced in a culturally acceptable fashion, so that problems of guilt, loss of self-esteem and blame are avoided (Naveed *et al.* 1992).

7.1.7 Overall assessment of viability of a screening programme in South Africa

Though not as common as in some of the Mediterranean countries, the haemoglobinopathies still constitute a major health load on the South African Indian community, particularly when the costs of maintenance therapy are considered. The technical expertise required for establishing a screening programme is present in the main centres, though laboratories may have to increase their staff to cope with an increased workload initially. Some consideration as to how to screen small outlying communities would be necessary. The establishment of further specialist laboratories may not be warranted. As long as an efficient method for transportation of samples is available, these samples could be tested in the major centres, though provision for counselling would have to be made. Screening by determination of HbA_2 levels and haemoglobin electrophoresis appears to be the most suitable protocol, with couples of childbearing age appearing to be the most appropriate target group.

A laboratory offering prenatal diagnosis by DNA analysis is available in Johannesburg. At present testing is offered to couples mainly identified retrospectively through the birth of an affected child. The current service is discussed in detail in Section 7.2. Again samples from other parts of the country could be transported to this laboratory, as is being done at present, rather than establishing other prenatal diagnosis centres at great cost. Centralised services which accumulate relevant experience are also better in small countries.

Some consideration would have to be given to the methods for delivery of effective community education, so that the awareness of the population is increased. At present,

the majority of the population is seemingly unaware of the existence of the haemoglobinopathies in the community. The presence of the disease or even heterozygosity for the disease in a family is considered a stigma and often hidden from other 'at-risk' family members. Small parent associations do exist, their main functions being support of new families and fund-raising to help families to buy pumps for desferrioxamine infusion. Parent support groups should be capable of mobilising substantial reserves, producing and distributing information, raising funds, exerting political pressure and acting as an information and informal counselling network (WHO 1983), and should thus be actively involved in a screening programme.

The entire community should play an active role in the planning, organisation, delivery and targeting of services. Community leaders must be included, so that the programme is accepted by the community. Ideally counsellors from the different subgroups would have to be trained for effective delivery of the service, so that information regarding inheritance and prenatal diagnosis is provided in a culturally-acceptable fashion. The introduction of the relevant facts into the school curriculum needs to be considered, as well as the use of the mass media. Radio and television programmes directed at the Indian community exist and could be used effectively in community education.

A screening programme for haemoglobinopathies in South African Indians is therefore a viable proposition and it should be planned and initiated as soon as possible. The initial costs in mass education and training will be minimal compared to the ongoing escalating costs of maintenance therapy for an ever-increasing number of affected homozygotes.

7.2 An assessment of the prenatal diagnostic service for haemoglobinopathies in South African Indians 1984-1993

The diagnostic service providing prenatal diagnosis by DNA analysis for haemoglobinopathies in South African was initiated in 1983, for both the Mediterranean and Asian Indian communities, the first prenatal diagnosis being done in 1984 for an Indian couple. Prior to this time, the only option available to 'at-risk' couples was fetal blood sampling and globin chain synthesis studies, or DNA studies, carried out in overseas centres like London.

Linked marker analysis was used initially, followed by mutation analysis, when this became available. The use of these two techniques in the South African situation is evaluated and compared to the experience elsewhere. A practical approach for the South African service is proposed in this section. The use of prenatal diagnosis in the South African Indians is assessed and the factors affecting uptake of prenatal diagnosis are reviewed. Fetal tissue sampling has been achieved using amniocentesis and chorionic villus sampling (CVS) and the impact of these procedures is also discussed. Only the service for the South African Asian Indian patients will be covered. The patients are those described in Section 2.1.3.

Most of the family workups and prenatal diagnoses were carried out by the author from 1986-1993. The earlier ones were performed in the same laboratory by Dr Michele Ramsay, using linked marker analysis. The techniques for mutation detection using ARMS were set up and all the mutation analysis in the families was done by the author.

7.2.1 Linked marker analysis

Linked marker analysis, using eight RFLPs, was performed in 31 families, who had requested studies in order to assess the feasibility of prenatal diagnosis. All had at least one child with a major haemoglobinopathy, 26 with ß thalassaemia, four with $ß^S$/ß thalassaemia and one with $ß^S$/ß thalassaemia.

7.2.1.1 Usefulness of linked marker systems

In 15 or 48% of families, at least one system was fully informative, while in a further five, each of the parents was informative with at least one system, making prenatal diagnosis possible using linked marker analysis in 20 or 64.5% of the families. Further, in one of the $ß^S$/ß thalassaemia families, the uninformative parent was the carrier of the $ß^S$ gene, which could be identified directly, initially with Southern blotting and later with PCR. In the other 10 families or 32.2%, however, only one parent was informative. In these families it was only possible to offer a modification of the risks for prenatal diagnosis, either excluding the informative parent's ß thalassaemia chromosome (in which case the fetus could only be normal or a heterozygote) or including the informative

parent's chromosome (in which case a fetus could either be affected or a heterozygote), though these two possibilities would not be distinguished. For many of the latter group of families, prenatal diagnosis was unacceptable because of the possibility of terminating a pregnancy where the fetus was at only 50% risk of being affected.

In a study of Asian Indian families in the United Kingdom, using only seven of the eight RFLP systems (excluding *Hin*dIII/3'ß), the number of informative families was slightly higher than those obtained here. In 76% of families a full diagnosis was possible, while in the remaining families, there was a 50% chance of successful diagnosis in 19% and no diagnosis possible in 5%. However, a diagnosis could be made using one enzyme in 63% of families, compared to the 48% in this study (Old *et al.* 1984).

In general, linked marker studies rely on the differentiation of $ß^A$ and $ß^T$ chromosomes, their usefulness depending on the frequency of alleles in the target population, and not on linkage disequilibrium with either the $ß^A$ or $ß^T$ chromosomes (Kazazian *et al.* 1980). Further, the usefulness of individual markers is affected by linkage disequilibrium between them, so that a second marker may not provide much more information than a first, if it is nearby (Chakravarti and Buetow 1985).

One of the two common haplotypes on $ß^A$ chromosomes, haplotype 1, is the commonest on ß thalassaemia chromosomes. This has important practical implications as parents often have identical haplotypes associated with their normal and mutant genes, making linked marker analysis for prenatal diagnosis impossible. On the other hand, the second common $ß^A$ haplotype, haplotype 16 is rare on $ß^T$ chromosomes, a useful finding for linked marker analysis in those parents who have this haplotype. The two haplotypes vary in frequency in the different South African Indian subgroups and in other Indian groups studied, as discussed in Section 5.1.3.1. The differences in the numbers of informative families and in the usefulness of markers may reflect the different haplotype distributions of these two, and to a lesser extent, the other haplotypes.

The eight most commonly used RFLPs in the ß globin cluster (sites 1, and 3-9 in Figure 2.2) were assessed in the 31 families in order to determine which were likely to be most informative for prenatal diagnosis. These results are shown in Table 7.1. Two other

TABLE 7.1 Analysis of linked marker systems used for workups in South African Indian families

SYSTEM	RFLP SITE[1]	FAMILIES					
		INFORMATIVE		SEMI-INFORMATIVE		NON-INFORMATIVE	
		n[2]	%	n[2]	%	n[2]	%
*Hinc*II/ε	1	3	9.7	15	48.4	13	41.9
*Hin*dIII/$^{G}\gamma$	3	2	6.3	19	61.3	10	32.3
*Hin*dIII/$^{A}\gamma$	4	0	0.0	14	45.2	17	54.8
*Hinc*II/$\psi\beta$[3]	5,6	6	19.4	18	58.1	7	22.6
*Hinc*II/5'$\psi\beta$[3]	5	3	9.7	14	45.2	11	35.5
*Hinc*II/[illegible]β[3]	6	5	16.1	16	51.6	10	32.3
*Ava*II/β	7	4	12.9	22	71.0	5	16.1
*Hin*dIII/3'β	8	5	16.1	20	64.5	6	19.4
*Bam*HI/β	9	0	0.0	13	41.9	18	58.1

[1] The numbers correspond to the RFLP sites labelled in Figure 2.2 and described in Table 2.5.

[2] n = number of families. More than one system may have been informative in a family.

[3] The two sites of the *Hinc*II/$\psi\beta$ system were analysed together with Southern blotting, but are analysed separately with PCR.

polymorphic sites, *Hgi*AI/ß and *Pvu*II/ψß, shown to have different frequencies on $ß^A$ and $ß^T$ chromosomes, and thus to be potentially useful (Old and Wainscoat 1983, Orkin and Kazazian 1984), were not studied in the South African cohort and have also been used rarely in other studies.

As predicted from the differences between allele frequencies on $ß^A$ and $ß^T$ chromosomes, the *Ava*II/ß and *Hin*dIII/3′ß sites are probably the most useful, as they are fully informative in 13% and 16%, and semi-informative in 71% and 65% of families, thus providing some information in 84% and 81% of families, respectively. The most informative system, when used alone, however, is the *Hin*cII/ψß system analysed with Southern blotting, as two RFLP sites are studied together. Even this system was only fully informative in 19% of families and is semi-informative in another 58% of families, providing at least some information in 77% of families. When the two *Hin*cII/ψß sites are analysed separately, as they are using PCR, the 3′ site provides more information than the 5′ site, providing full information in 16%, and partial information in 52%, of families respectively.

The *Hin*dIII/^{A}γ and *Bam*HI/ß systems were the least informative systems and were not fully informative in any of the families. This is not unexpected, since the rarer allele in both these systems occurs at relatively low frequency in most of the groups studied, on both $ß^A$ and $ß^T$ chromosomes.

In previous studies on Asian Indians in the United Kingdom and the USA, the *Ava*II/ß (site 7) polymorphism was found to be useful (Old *et al.* 1984, Orkin and Kazazian 1984), but the *Hin*dIII/3′ß site was not used. However, the *Hin*cII/ε (site 1), the two *Hin*dIII/γ (sites 3 and 4) and the two *Hin*cII/ψß sites (sites 5 and 6) were found to be most useful and the *Bam*HI/ß site to be least useful. Further, because the five 5′ sites are non-randomly associated, it was found that if the *Hin*cII/ε could be used for prenatal diagnosis, so could the other four in most cases (Old *et al.* 1984). This does not seem to occur in the South African Indians, and may again reflect some differences in haplotype distribution between the two groups. The differences in linked markers, and mutation distribution, between groups emphasise the importance of surveying the local target population prior to embarking on a service.

7.2.1.2 Practical approach to linked marker analysis

PCR is currently the method of choice for analysis as it is more rapid and less expensive than Southern blotting. The two systems likely to provide the most information are the *Ava*II/ß and *Hinc*II/3′ψß systems. In view of the low information rate provided, the other systems may well have to be tested in many of the families, to obtain a full diagnosis. In general, the 3′ haplotype (sites 7 and 9) should be studied first, as there is always a chance of recombination at the recombination hotspot. No PCR primers are available yet for the *Hind*III/3′ß system and, in addition, the probe, pRK29, used with Southern blotting, tends to be unreliable, making it relatively impractical to use this system for prenatal diagnosis.

At least two generations are required for linkage analysis, and studies are generally not feasible for first pregnancies, unless detailed family studies are undertaken. Family studies are often tedious requiring blood from large numbers of individuals. For example, analysis of 281 Cypriot, Indian and Pakistani families required blood from over 1800 individuals (Old *et al.* 1986). In addition to DNA from the parents, DNA is required from a normal or affected child, or from one set of grandparents if a heterozygous child is available, or both sets of grandparents if no children are available, with one of the grandparents on each side normal with respect to ß thalassaemia, so that a linkage phase can be established. Linkage analysis is dependent on correct phenotyping and true paternity for correct assignment as well as finding an informative marker (reviewed in Old and Ludlam 1991). Further, recombination resulting in error is estimated to occur in 1/350 meioses and is potentially a problem for linkage analysis (Chakravarti *et al.* 1984).

7.2.2 Mutation analysis

Mutation analysis was performed in 35 families, to assess the feasibility of prenatal diagnosis. This included the 31 families with a child with a major haemoglobinopathy, analysed with linked markers (in Section 7.2.1), as well as four in which both members of a couple were diagnosed as ß thalassaemia heterozygotes prior to any affected children being born. An additional family, where both parents were shown to be $ß^S$ heterozygotes, prior to the birth of any affected children, also requested prenatal diagnosis.

7.2.2.1 Usefulness of mutation analysis

Table 7.2 shows the distribution of mutations in the 35 families (excluding the family 'at-risk' for sickle cell anaemia). As the mutation distribution in the religious and language groups is virtually identical to that in Table 6.1, the mutation distribution of the families according to their subgroups is not shown. Of 64 ß thalassaemia chromosomes analysed, 59 or 92% mutations were identified using five sets of ARMS primers and the internal control primers A and B, which also identify the 619bp deletion. The four common mutations account for 84% of chromosomes, and the codon 15 (G→A) and codon 30 (G→C) mutations, a further 8%. A further three mutations were identified using DNA sequencing, but ARMS primers have not been synthesised for these as they have been found in single families and would thus not add considerably to our diagnostic service, at present. In two of the families the parent with the mutation was informative with a linked marker while in the third family, the parent was uninformative with linked markers, but he had a $ß^D$ mutation on his second chromosome 11 which could be identified, as described in Section 2.3.5.2.3, and thus a full prenatal diagnosis could be offered to these families. One mutation remains unknown, though the parent was informative with a linked marker.

In four of the haemoglobinopathy families, one parent was a $ß^S$ heterozygote and this mutation can be identified directly using PCR amplification of the ß globin gene and *Dde*I digestion, as described in Section 2.3.5.2.2. In one family, one parent was a $ß^E$ heterozygote. This mutation can also be identified directly as described in Section 2.3.5.2.3.

Using mutation analysis alone, with the ARMS primers we have available at present, as well as techniques for direct detection of the $ß^S$, $ß^E$ and $ß^D$ mutations, 33/36 or 92% of families would be fully informative. The ARMS technique thus makes a comprehensive prenatal diagnosis programme for ß thalassaemia possible in South African Indians as it does in the United Kingdom (Old *et al.* 1990). Combined with linked marker analysis, 100% of the families studied are informative and can be offered a full prenatal diagnosis.

In 13 or 45% of the thalassaemia major families, both parents share the same mutation,

TABLE 7.2 Mutation analysis in South African Indian families requesting prenatal diagnosis

MUTATION	N[1]	%
IVS1nt5 (G→C)*	37	52.9
Codon 41/42 (-CTTT)*	6	8.6
IVS1nt1 (G→T)*	6	8.6
619bp deletion*	5	7.1
Codon 30 (G→C)*	2	2.9
Codon 15 (G→A)*	3	4.3
Cap site (+1) (A→C)*	1	1.4
Codon 30 (G→A)#	1	1.4
Codon 44 (-C)#	1	1.4
Poly A (T→C)#	1	1.4
β^S	4	5.7
β^E	2	2.9
Unknown	1	1.4
Total	70	100.0

[1] N = number of chromosomes.
* Identified by the ARMS technique.
# Identified by direct sequencing from PCR products.

the IVS1nt5 in 11 families, and the IVS1nt1 and codon 41/42 each in one family. Based on the frequencies of the mutations calculated (Table 6.1), one would expect parents in 35% of families to share the same mutations by chance. As four of these families are known consanguineous marriages, the observed number of parents sharing mutations by chance is nine or 31%, close to that expected.

The number of parents who share mutations due to consanguinity is lower than expected, as all the Moslem groups and the Hindu Tamil encourage consanguineous marriages. It is estimated that up to 30% of marriages may be consanguineous in some groups. Consanguinity is thought to have important social functions in other Indian groups, and to result in greater compatibility of the bride with the husband's family, the maintenance of family property, reduced dowry payments and reassurance in marrying into a known family background (Modell 1991, Bittles *et al.* 1992). It is possible that those individuals who maintain the tradition of consanguineous marriage have the strongest religious beliefs and thus the strongest objections to prenatal diagnosis and termination of pregnancy. They may therefore not be presenting for genetic counselling and testing.

If about 30% of marriages are assumed to be consanguineous, and prenatal diagnosis becomes acceptable to the majority of the population, it can be estimated that in over 50% of prenatal diagnoses carried out, it would only be necessary to test for a single mutation, either a common mutation shared by chance or a mutation shared because of consanguinity. This is consistent with a study of Indians in the United Kingdom, where the majority of patients were homozygous for a single mutation, a finding thought to be related to customs and marital practices, with consanguinity favoured in some communities and mutations thought to be confined to individual endogamous groups (Thein *et al.* 1988). Further, in a study of 32 thalassaemia major patients from Tamil Nadu in India, 75% were the result of consanguineous marriages and none was a compound heterozygote (Venkatesan *et al.* 1992), probably due to the high frequency of consanguinity and of the IVS1nt5 mutation.

7.2.2.2 Practical approach to mutation analysis

As 38% of families were found to be homozygous for the IVS1nt5 mutation, it would

seem most cost-effective to screen for it first in all new families. The 619bp deletion would be detected simultaneously, as a smaller fragment would be amplified by the internal control primers, as described previously. If the IVS1nt5 mutation is absent, the codon 41/42 mutation should be tested for in Hindu families, and the IVS1nt1 and then the codon 41/42 mutation in Moslem families. The IVS1nt1 mutation appears to be absent, or at least occurs at low frequency, in the South African Hindu groups. In all groups, the codon 15 (G→A) followed by the codon 30 (G→C) mutations should then be sought. If these are absent, the other known Indian mutations could be tested for, using the available primers (shown in Table 2.8). If a mutation is still not found and linked markers are uninformative, the ß globin gene can be sequenced if time permits. Using this approach, very few uninformative or semi-informative families should be found.

This is a slightly different approach to that of Old *et al.* (1990), who screen initially for their five common mutations (including the deletion detected by the control primers) in one PCR run, and then screen for the next five rarer mutations, if required in the remaining 10% of samples, later the same day. In the local context, this approach would seem to be more expensive than that outlined above, but may be used for urgent workups to save time.

The ARMS technique is far less time-consuming and expensive than other methods of mutation analysis using ^{32}P-labelled oligonucleotides and dot blot analysis, and is non-radioactive (Old *et al.* 1990). Mutational analysis may be informative in cases where linked markers are not. Further, the main advantage of mutation analysis is that it eliminates the need for extended family studies, particularly when an affected individual is unavailable. Only the parents are needed, to identify the mutations they carry. Linkage analysis is, however, a useful backup if the relevant family members are available, or where a mutation cannot be identified. The ARMS technique was assessed using 32 first trimester Indian prenatal diagnoses in the United Kingdom. All were confirmed with RFLP studies or DNA sequencing, where no informative linked markers were found, and the technique was found to be at least as reliable as linked markers and even more so in one case where samples had been mislabelled (Old *et al.* 1990).

Although, there is still a need for prior identification of mutations in a particular family,

this can be done rapidly with the ARMS technique, so that the parents' and fetus' mutations can be identified in one day. This is particularly important in ß thalassaemia, as due to the wide range of mutations, fetuses are often compound heterozygotes. The technique thus has wide application to carrier detection and prenatal diagnosis of monogenic disease with heterogeneous molecular defects (Old *et al.* 1990).

Rund *et al.* (1992a) have shown that almost all carriers of $ß^0$ mutations had an MCV<67fℓ, while $ß^+$ carriers were above this level. The different $ß^+$ mutations have different MCV and MCH ranges, while virtually all $ß^0$ mutations had a similar range. Mutations within the intervening sequences resulting in activation of cryptic splice sites cause lower MCVs than mutations in consensus sequences or splice signals, while transcriptional and polyadenylation signal mutations are milder with higher MCVs (Rund *et al.* 1992a). In theory this may be useful to predict which mutations to screen for in known heterozygotes, but there are a number of problems in using this approach in Indians. The majority of common mutations are of the $ß^0$ or severe $ß^+$ (in the case of IVS1nt5) types and thus are likely to have similar MCV and MCH ranges. Further, while most of the MCV and MCH values are consistent with the values suggested by Rund *et al.* (1992b) in individuals who are $\alpha\alpha/\alpha\alpha$, α thalassaemia and iron deficiency are common and both affect the MCV and MCH, increasing the MCV and MCH range and making prediction of a mutation based on haematological indices extremely difficult.

7.2.3 Prenatal diagnosis

The family distress caused by the birth of a child affected with thalassaemia major, the medical effort involved in treating it and the cost-benefit advantage of prenatal detection and termination are all significant (Modell *et al.* 1984). Primary prevention is not possible and treatment is burdensome, expensive and unsatisfactory, making prenatal diagnosis, with its aim of enabling couples 'at-risk' to have thalassaemia-free families, a necessary service (Kuliev 1986, Modell 1990). Prenatal diagnosis is the major factor in the reduction of new cases, with very few being prevented by avoidance of carrier marriages or avoidance of pregnancy in 'at-risk' couples (Angastiniotis *et al.* 1986).

The aims of prenatal diagnosis are "to allow the widest range of informed choice to

women and couples 'at-risk' of having a child with an abnormality, to provide reassurance and reduce the level of anxiety associated with reproduction, to allow couples 'at-risk' to embark on having a family knowing that they may avoid the birth of seriously affected children through selective abortion, to ensure optimal treatment of affected individuals through early diagnosis" (Royal College of Physicians of London Working Party on Prenatal Diagnosis and Genetic Screening 1989).

7.2.3.1 Prenatal diagnosis in South African Indians

The development of techniques for DNA analysis has revolutionised the prenatal diagnosis of genetic disease, including the haemoglobin disorders. The techniques for prenatal diagnosis of haemoglobinopathies have evolved from globin chain analysis through DNA analysis by Southern blotting to PCR-based diagnostic methods.

In South Africa, fetal blood sampling and globin chain analysis have not been available. three Indian couples, two 'at-risk' for ß thalassaemia and one 'at-risk' for sickle cell anaemia, had prenatal diagnoses done overseas, using these techniques prior to the availability of DNA diagnosis in South Africa.

Of the 36 families studied using DNA analysis, 17 have had 22 prenatal diagnoses, in which linked markers or mutation analysis have been used for fetal analysis after either amniocentesis or CVS. The details are shown in Table 7.3. A diagnosis was obtained in all cases. Figure 7.1 shows one of the families who have used the prenatal diagnosis service to achieve their desired family size. In general, prenatal diagnosis in South Africa is following the trends elsewhere by increasingly using CVS rather than amniocentesis for collection of fetal tissue. The first CVS for ß thalassaemia in South Africa was done in 1984 (Ramsay *et al.* 1985). PCR analysis is used rather than Southern blotting and direct mutation detection rather than linkage analysis. Where a family study is possible, linked markers should always be used in addition to direct detection, to confirm the result, avoiding errors due to wrongly labelled samples, non-paternity, contamination with previously amplified target DNA sequence or maternal DNA contamination (Old and Ludlam 1991, Globin Gene Disorder Working party of the BCSH General Haematology Task Force 1994).

TABLE 7.3 Analysis of techniques used in South African Indian families for prenatal diagnoses of haemoglobinopathies (1984-1993)

METHOD OF ANALYSIS	AMNIOCENTESIS	CHORIONIC VILLUS SAMPLING (CVS)
Linked Markers	5	5
Mutation analysis	4	8
Total	9	13

FIGURE 7.1 A South African Asian Indian family who have used the prenatal diagnosis service to complete their family

The couple's oldest child (left) is a β thalassaemia heterozygote, while his younger brother (middle) has thalassaemia major. The couple's third pregnancy was terminated after an affected fetus was diagnosed on CVS. CVS in the subsequent pregnancy showed the fetus to be a β thalassaemia heterozygote, and the couple now have a healthy daughter (right).

Sixteen of the families who have had prenatal diagnosis have been 'at-risk' for thalassaemia major. Three of the fetuses were shown to be homozygous β^A, 11 heterozygous for ß thalassaemia and three to be affected with thalassaemia major. Two of the affected pregnancies were terminated. A further two fetuses were shown to be at 50% risk of being affected, and one of these pregnancies was terminated. The second pregnancy went to term and the baby was later shown to be heterozygous for thalassaemia. The diagnosis was confirmed postnatally by haematological techniques in all pregnancies that went to term. In two of the pregnancies that were terminated, DNA studies on the fetus showed the same results as on the amniocentesis. In the third pregnancy, terminated after CVS, no fetal tissue was obtained for confirmation. There were no errors in diagnosis as far as could be determined.

The seventeenth family, 'at-risk' for sickle cell anaemia, has had three prenatal diagnoses with amniocentesis. Two of the fetuses were shown to be β^A homozygotes, while the third was a β^S homozygote and the pregnancy was terminated. Since sickle cell anaemia is a relatively mild disease in Indians, prenatal diagnosis is not frequently requested.

It is difficult to determine the number of pregnancies which have proceeded without prenatal diagnosis. The uptake of prenatal diagnosis is currently about 60-70% in couples who attend the Genetic Counselling Clinic. A significant number do not keep appointments made for them, however. Further, there are likely to be additional families who do not even make appointments at the clinic. At least five thalassaemia major children have been born since 1988, two in families with a previous affected child and a third in an 'at-risk' couple identified during the pregnancy, all of whom had attended a genetic counselling clinic. The remaining two couples were unaware of their thalassaemia risk status. The number of thalassaemic newborns reflects the efficacy of the whole preventive programme, rather than prenatal diagnosis alone (Loukopoulos 1985). Residual cases have been shown to be a result of lack of information more frequently than misdiagnosis or refusal of prenatal diagnosis (Cao *et al.* 1989b). Thus, in South Africa, where there is no formal screening and education programme, the occurrence of new cases is not entirely surprising, and it reinforces the need for such a programme.

Since there is no formal screening programme, the majority of 'at-risk' couples are still, unfortunately, identified by the birth of an affected child. However, there is an increasing awareness of the disorder in the Indian community and the doctors serving it, so that a number of 'at-risk' couples have now been detected prior to the birth of an affected child. Only two of the 19 prenatal diagnoses done for thalassaemia major were done in 'at-risk' couples detected prior to the birth of a child with the disorder. A third couple was determined to be 'at-risk' during their first pregnancy. Prenatal diagnosis by amniocentesis was offered to them, but they declined as they felt the pregnancy, at 20 weeks gestation, would be too far advanced to be terminated. A major factor in this couple's decision was their lack of knowledge and experience of the disease.

7.2.3.2 Factors affecting uptake of prenatal diagnosis

Many factors have been identified which influence a couple's decision to accept prenatal diagnosis, including wanting more children, awareness of their risks, more years of education and older patient age, as well as the severity of the disease in question, the inheritance pattern and degree of risk, the ethnic group and the manner in which an 'at-risk' pregnancy is identified (Modell *et al.* 1980, Rowley *et al.* 1991a). Women who decide against prenatal diagnosis often fear the pain of the procedure or fetal injury or miscarriage, or may feel unable to terminate a pregnancy on moral or religious grounds. They may have a poor obstetric history, are less likely to have experience of the disease (either having had no family members with the disease or individuals mildly affected at the time of prenatal diagnosis), or have not been informed of the disease prior to prenatal counselling (Modell *et al.* 1980, Cao *et al.* 1984, Driscoll *et al.* 1987, Anionwu *et al.* 1988, Loader *et al.* 1991b, Rowley *et al.* 1991a,b, Petrou *et al.* 1992). Further, in those who refuse prenatal diagnosis, the partner is frequently absent at the time of the counselling session, or the woman is not living with the partner or she may be more than 18 weeks pregnant (Rowley *et al.* 1991a).

It has been demonstrated that the level of understanding obtained through genetic counselling is often below expectation, with higher knowledge related to younger patients, with more years of education, no prior children and prior knowledge of their carrier status (Loader *et al.* 1991a). It was found that many individuals lacked an understanding of

probability and basic biological concepts, making genetic counselling difficult (Loader *et al.* 1991a). In addition, in areas where the disease is less common or where the disease represents gh risk for a minority group in a low risk country, and has not been identified as an important local problem, prenatal testing is less accepted (Modell *et al.* 1984, Driscoll *et al.* 1987). A systematic screening and counselling programme is required with increased awareness of haemoglobinopathies to place couples in a position to make a fully-informed decision.

Many problems related to poor acceptance of prenatal diagnosis arise from inadequate information and there is therefore a strong case for improving community and medical education about the nature and frequency of genetic risks and the role of prenatal diagnosis (Modell *et al.* 1980). The most efficient channels to inform 'at-risk' couples have been shown to be the mass media, and general practitioners or obstetricians (Cao *et al.* 1985).

It has been shown that the establishment of services for carrier testing, genetic counselling and prenatal diagnosis assists in education, altering psychosocial, cultural and religious attitudes, and resulting in increased acceptance of the procedures (Yuen *et al.* 1990). In Greece, the increasing confidence of the public in the procedure is reflected not only in the steady increase in the number of prenatal diagnoses performed, but also in the increase in prenatal diagnoses for second and subsequent pregnancies (Loukopoulos 1985). Further, when offering prenatal diagnosis to a minority population, cultural, linguistic, religious, social and economic factors must be taken into account (Anionwu *et al.* 1988).

In the local context religious beliefs play a major role, particularly in the Moslem groups, where second trimester termination of pregnancy is completely unacceptable to many couples. A number of couples, including Moslems, who had previously refused prenatal diagnosis with amniocentesis, because of their concerns with a late termination of pregnancy, have accepted CVS for prenatal diagnosis. The increased uptake of prenatal diagnosis with the availability of CVS has also been shown in Indian communities living in the United Kingdom (Modell 1985, Petrou *et al.* 1990) and in other communities, including Lebanese Moslems (Der Kaloustian *et al.* 1987) and Sardinians (Cao *et al.* 1987). For most people abortion appears to be more acceptable morally in early