DNA INSTABILITY IN
THE HUMAN \( \alpha \)-GLOBIN GENE CLUSTER

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Kwan-Wood Gabriel Lam

Department of Genetics

University of Leicester

October 2007
I know nothing except the fact of my ignorance

— Socrates (469 BC – 399BC)
Contents

Acknowledgements vii
Abbreviations vi
Abstract viii

Chapter 1 Introduction 1

1.1 Significance of meiosis 1

1.1.1 Important stages in meiosis 1
1.1.1.1 Tetrads 1
1.1.1.2 Pachynema 2
1.1.1.3 Diplotene 3
1.1.1.4 Diakinesis 4
1.1.1.5 Metaphase I 5
1.1.1.6 Anaphase I 6
1.1.1.7 Telophase I 7
1.1.1.8 Cytokinesis 8

1.1.2 Homologous pairing 8
1.1.2.1 Matching of homologues 8
1.1.2.2 Double-strand break repair model 9
1.1.2.3 Crossing over 10
1.1.2.4 Recombination nodules 11
1.1.2.5 Chiasmata 12
1.1.2.6 Recombination hot spots in yeast 13

1.1.3 Crossover and gene conversion 14
1.1.3.1 Pedigree studies 14
1.1.3.2 Linkage disequilibrium analysis 15
1.1.3.3 Coalescent analysis 16

1.2 Allelic recombination in humans 17
1.2.1 Pedigree studies 17
1.2.2 Linkage disequilibrium analysis 18
1.2.3 Coalescent analysis 19
1.2.4 Single sperm typing techniques 20
1.2.5 Allelic recombination hot spots identified by multiple sperm typing 21
1.2.5.1 Minisatellite-associated hot spot 21
1.2.5.2 Hot spots within the major histocompatibility complex region 22
1.2.5.3 SHOX hot spot in the pseudoautosomal region 22
1.2.5.4 ß-globin hot spot 22
1.2.5.5 Hidden hot spots 22
1.2.6 Properties of allelic recombination hot spots 23

1.3 Ectopic recombination 24
1.3.1 Significance of ectopic recombination 24
1.3.1.1 Relationship between recombination frequency and copy number 24
1.3.1.2 Physical distance between repeats 25
1.3.1.3 Length of homology 26
1.3.1.4 Degree of sequence divergence between repeats 26
1.3.1.5 Sequence motifs 27
1.3.1.6 Ectopic gene conversion 27
1.3.2 Other mechanisms for genome rearrangement 28
1.3.2.1 Single strand annealing model 28
1.3.2.2 Replication slippage 29
1.3.3 Evolution of gene families 30
1.3.4 Copy number variation in the human genome 31
1.3.4.1 Array-based comparative genome hybridisation 31
1.3.4.2 PCR-based approaches for targeted screening 32
1.3.4.3 Confirmation using fluorescence in situ hybridisation 32
1.3.5 Classic examples of ectopic rearrangements in the human genome 33
1.3.5.1 CMT1A and HNPP 33
1.3.5.2 ß-thalassaemia caused by ectopic deletions in the ß-globin gene cluster 34
1.3.5.3 Unequal crossover in minisatellite MS32 34
1. Human α-globin genes
   1.1 Organisation of the human α-globin gene cluster
   1.2 Evolution of genes in the α-globin gene cluster
   1.3 Gene expression in normal individuals
   1.4 α-thalassaemia caused by α-globin gene deletions
      1.4.1 α-thalassaemia caused by equal crossover
      1.4.1 Illegitimate recombination causing α-thalassaemia
   1.5 Phenotypical differences between normal individuals and α-thalassaemics
   1.6 Malaria selection and distribution of α-thalassaemia
   1.7 Allelic recombination in the α-globin gene cluster

1.5 Overview of the thesis

Chapter 2 Materials and Methods

2.1 Materials
   2.1.1 Chemical reagents
   2.1.2 Consumables and equipment
   2.1.3 Standard solutions
   2.1.4 Sources of human DNA sample
   2.1.5 Oligonucleotides
   2.1.6 Enzymes
   2.1.7 DNA ladders
   2.1.8 Computers

2.2 Methods
   2.2.1 DNA extraction
      2.2.1.1 DNA extraction from human sperm
      2.2.1.2 DNA extraction from human blood
      2.2.1.3 DNA extraction from human testis
      2.2.1.4 Plasmid DNA extraction
   2.2.2 Measuring DNA concentration
   2.2.3 PCR amplification
      2.2.3.1 Primer design
      2.2.3.2 PCR buffer
      2.2.3.3 PCR cocktail
   2.2.4 Gel electrophoresis
      2.2.4.1 Electrophoresis conditions
      2.2.4.2 Gel photography
   2.2.5 DNA purification
      2.2.5.1 Exo/Sap purification
      2.2.5.2 Electroelution from agarose gels
      2.2.5.3 Ethanol precipitation
      2.2.5.4 Estimating DNA recovery
   2.2.6 DNA size fractionation
      2.2.6.1 Restriction enzyme digestion
      2.2.6.2 Preparative gel electrophoresis
      2.2.6.3 Physical enrichment
   2.2.7 DNA transfer to membrane
      2.2.7.1 Southern blot
      2.2.7.2 Dot blot
   2.2.8 Hybridisation
      2.2.8.1 Southern blot hybridisation
      2.2.8.2 Dot blot hybridisation
      2.2.8.3 Probe removal from membrane
   2.2.9 Automated DNA sequencing
   2.2.10 Statistical tests

Chapter 3 PCR Amplification of Human α-Globin Genes

3.1 Introduction
   3.1.1 Long PCR amplification
   3.1.2 PCR amplification of GC-rich targets

3.2 Results
   3.2.1 Attempts to amplify across the α-globin gene locus
   3.2.2 Breaking up the α-globin gene cluster
   3.2.3 Alteration of PCR buffer concentration
   3.2.4 Successful amplification with PCR additives
3.3 Discussion 76
3.3.1 GC-content of human α-globin genes 76
3.3.2 Importance of salt concentration and PCR additives 76
3.4 Conclusions 78

Chapter 4 The Dynamics of α-Globin Gene Deletion 79
4.1 Introduction 79
4.2 Results 81
4.2.1 Single-molecule PCR of α-globin genes 81
4.2.2 Selection of informative donors 82
4.2.3 Physical enrichment of deletion mutants 84
4.2.3.1 Preparative gel electrophoresis for separating deletion molecules from progenitors 84
4.2.3.2 Recovery of deletion mutants and size validation 87
4.2.4 Characterisation of deletion mutants 89
4.2.4.1 Restriction mapping of deletion mutants 89
4.2.4.2 Refining the mapping of ectopic exchange breakpoints 90
4.2.4.3 Poisson correction of the frequency of deletion mutants 92
4.2.5 Deletion frequencies 94
4.2.6 Structures of deletions 94
4.2.6.1 Intra- and interchromosomal exchanges 94
4.2.6.2 Simple and complex exchanges 96
4.3 Discussion 98
4.3.1 Substantial somatic and germ-line deletion frequencies 98
4.3.2 Distinct mechanisms in soma and germ-line 100
4.3.2.1 Mitotic recombination frequency altered by mutational mosaicism in soma 100
4.3.2.2 Mechanisms in germ-line 100
4.3.3 Random distribution of exchange points 102
4.3.4 A challenge to minimum homology requirement 104
4.3.5 Sequence identity is not the sole factor controlling ectopic recombination 105
4.4 Conclusions 106

Chapter 5 Deletion Frequencies and Population Fitness 108
5.1 Introduction 108
5.2 Results 109
5.2.1 Screening of −α chromosomes 109
5.2.1.1 Samples from the Leicester Royal Infirmary panel 110
5.2.1.2 Samples from Centre d’Etude du Polymorphisme Humain (CEPH) panel 110
5.2.1.3 Samples from M. Jobling’s panel 111
5.2.2 Characterisation of −α chromosomes 111
5.2.3 Comparison of population studies and −α chromosome screening 111
5.3 Discussion 112
5.3.1 Incidence of −α chromosomes and the effect of fitness 112
5.3.2 Possible selective forces 115
5.4 Conclusions 117

Chapter 6 Processes of De Novo ααα Duplication 118
6.1 Introduction 118
6.2 Results 120
6.2.1 Amplification of triplicated α-globin gene molecules by small pool PCR 120
6.2.2 Size fractionation for duplication mutant detection 121
6.2.2.1 Preparative gel electrophoresis and DNA recovery 121
6.2.2.2 PCR recovery of duplication molecules by inverse PCR 123
6.2.2.3 Distribution of duplications (size validation) and progenitor contamination 126
6.2.3 Characterising duplication mutants 127
6.2.4 Duplication frequencies 129
6.2.5 Alternative approach of size fractionation 130
6.2.6 Structures of duplications 133
6.2.6.1 Intra- and interchromosomal recombinants 133
6.2.6.2 Simple vs. complex exchanges 136
6.2.6.3 Ectopic recombinants within Y homology blocks 136
6.2.6.4 Duplications accompanied by deletion 137
6.3 Discussion 140
6.3.1 Elimination of inverse PCR artefacts 140
6.3.2 Reciprocity of α57 deletions and ααα mutations 141
6.3.2.1 Two distinct mechanisms

6.3.2.2 Similar distribution of ectopic exchange points

6.3.3 Mechanisms of intrachromosomal \( \alpha \alpha^{\text{anti-7}} \) duplication

6.3.4 Duplication dynamics and the population incidence of \( \alpha \alpha^{\text{anti-7}} \) chromosomes

6.4 Conclusions

Chapter 7 Detection of Extrachromosomal Circular DNA

7.1 Introduction

7.2 Results

7.2.1 Selection of donor samples

7.2.2 Collecting extrachromosomal circular DNA from genomic DNA

7.2.2.1 Mobility of extrachromosomal circles

7.2.2.2 Preparative gel electrophoresis

7.2.2.3 DNA recovery and distributions of circular DNA molecules and progenitors

7.2.3 Progenitor digestion and detection of putative mutants

7.2.4 Characterisation of putative mutants

7.3 Discussion

7.3.1 Low frequencies of putative extrachromosomal circles

7.3.2 Different topoforms of putative circles

7.3.3 Mutants accompanied by deletions

7.3.4 Possible routes of transmission

7.4 Conclusions

Chapter 8 Strategies for Detecting Allelic Recombination Events

8.1 Introduction

8.2 Results

8.2.1 LD analysis of the \( \alpha \)-globin gene cluster

8.2.2 Informative SNP searching by resequencing

8.2.2.1 Screening for \( \alpha \)-globin gene rearrangements

8.2.2.2 Additional SNPs identified in \( \beta \) homology blocks

8.2.3 Additional SNPs detected in the \( \theta 1 \) gene region

8.3 Discussion

8.3.1 Strategies for high-resolution sperm analysis

8.3.2 Strong LD block and low density of heterozygous SNPs in \( \alpha \)-globin genes

8.3.3 Recombination motif analysis

8.4 Conclusions

Chapter 9 Final Discussion

9.1 Introduction

9.2 Relationship between GC-content and recombination

9.3 Reciprocity of deletion and duplication

9.4 Intra- vs. interchromosomal ectopic exchanges

9.5 Factors affecting ectopic recombination

9.6 Gene conversion in mitotic and meiotic recombination

9.7 Possible mechanisms for exceptional mutants

9.8 Dynamics for maintaining normal \( \alpha \alpha \) chromosomes in populations

9.9 Future directions

Appendix I

Appendix II

References
Acknowledgements

First and foremost, I would like to express my heartfelt thanks and sincere gratitude to Alec, for all the support, time, guidance and patience that you have given to me. Thank you very much for giving me this opportunity to work in your lab. It is my absolute and precious privilege to work with such a ‘star’ scientist. You are the most knowledgeable gentleman and the best supervisor I have ever come across. Thank you very much for setting up an excellent model not only for me but also for young scientists.

I would also like to give my special thanks to all lab members in G18 and G19 for giving me a memorable three years. Thank you all for creating a harmonious and relaxing atmosphere both inside and outside the lab. I have been very pleased to be one of the members in this family. Additionally, I would like to thank Rita for the technical support, particularly for the laborious preparation of the huge amount of PCR buffer that I always asked for; Celia and Ingrid, you have been always generous and helpful!!! Indeed, my thanks to Vicki, thank you for being my company specially when I first established my life in Leicester. Ila and Jennie Cox, you have been very nice!! Thanks for the help and stipends that you arranged for me. Tim, looked like an elder brother in the lab, thanks for taking care of me. I enjoy very much playing footy with you. Thanks for organising the 5-a-side!! Adam, thanks a lot for your technical support on Macs. Aaron, thanks for being my tennis competitor; and Matt, many thanks for organising some funny games for the lab particularly the Essential Football League and the Conkers, and more importantly for the cash and vouchers that I won from these games...........

Last but not least, my thanks to the almighty God. Thanks for creating this wonderful world. Thanks for all the blessings you have given to me and for listening to my prayers!
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARS</td>
<td>autonomously replicating sequence</td>
</tr>
<tr>
<td>ASO</td>
<td>allele-specific oligonucleotide</td>
</tr>
<tr>
<td>BAC</td>
<td>bacterial artificial chromosome</td>
</tr>
<tr>
<td>BIR</td>
<td>break-induced replication</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>CEPH</td>
<td>Centre d'Etude du Polymorphisme Humain</td>
</tr>
<tr>
<td>CGH</td>
<td>comparative genome hybridisation</td>
</tr>
<tr>
<td>CHO</td>
<td>Chinese hamster ovary</td>
</tr>
<tr>
<td>CMT1A</td>
<td>Charcot–Marie–Tooth type 1A</td>
</tr>
<tr>
<td>CNP</td>
<td>copy number polymorphism</td>
</tr>
<tr>
<td>dbSNP</td>
<td>database single nucleotide polymorphism</td>
</tr>
<tr>
<td>DEASH</td>
<td>DNA enrichment by allele-specific hybridisation</td>
</tr>
<tr>
<td>DGS</td>
<td>DiGeorge syndrome</td>
</tr>
<tr>
<td>dHJ</td>
<td>double Holliday junction</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DSB</td>
<td>double strand break</td>
</tr>
<tr>
<td>FISH</td>
<td>fluorescence <em>in situ</em> hybridisation</td>
</tr>
<tr>
<td>HNPP</td>
<td>hereditary neuropathy with liability to pressure palsies</td>
</tr>
<tr>
<td>JM</td>
<td>joint molecule</td>
</tr>
<tr>
<td>LB</td>
<td>Luria-Bertani broth</td>
</tr>
<tr>
<td>LCR</td>
<td>low copy repeat</td>
</tr>
<tr>
<td>LD</td>
<td>linkage disequilibrium</td>
</tr>
<tr>
<td>LE</td>
<td>lateral element</td>
</tr>
<tr>
<td>LR</td>
<td>likelihood ratio</td>
</tr>
<tr>
<td>LRI</td>
<td>Leicester Royal Infirmary</td>
</tr>
<tr>
<td>MAF</td>
<td>minor allele frequency</td>
</tr>
<tr>
<td>MAPH</td>
<td>multiplex amplifiable probe hybridisation</td>
</tr>
<tr>
<td>MCH</td>
<td>mean corpuscular haemoglobin</td>
</tr>
<tr>
<td>MCV</td>
<td>mean corpuscular volume</td>
</tr>
<tr>
<td>MEPS</td>
<td>minimal efficient processing segment</td>
</tr>
<tr>
<td>MHC</td>
<td>major histocompatibility complex</td>
</tr>
<tr>
<td>MITE</td>
<td><em>mariner</em> transposon-like element</td>
</tr>
<tr>
<td>MLPA</td>
<td>multiplex ligation-dependent probe amplification</td>
</tr>
<tr>
<td>MMR</td>
<td>mismatch repair</td>
</tr>
<tr>
<td>NAHR</td>
<td>non-allelic homologous recombination</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Centre for Biotechnology Information</td>
</tr>
<tr>
<td>NHEJ</td>
<td>non-homologous end joining</td>
</tr>
<tr>
<td>OD</td>
<td>optical density</td>
</tr>
<tr>
<td>PAR</td>
<td>pseudoautosomal region</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PSO</td>
<td>paralogue-specific oligonucleotide</td>
</tr>
<tr>
<td>PSV</td>
<td>paralogous sequence variant</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>ROMA</td>
<td>representational oligonucleotide microarray analysis</td>
</tr>
<tr>
<td>SDS</td>
<td>sodium dodecyl sulfate</td>
</tr>
<tr>
<td>SDSA</td>
<td>synthesis-dependent strand annealing</td>
</tr>
<tr>
<td>SEI</td>
<td>single-end invasion</td>
</tr>
<tr>
<td>SMS</td>
<td>Smith-Magenis syndrome</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SP-PCR</td>
<td>small pool polymerase chain reaction</td>
</tr>
<tr>
<td>SSC</td>
<td>sodium chloride-sodium citrate</td>
</tr>
<tr>
<td>spcDNA</td>
<td>small polydisperse circular DNA</td>
</tr>
<tr>
<td>SSA</td>
<td>single strand annealing</td>
</tr>
<tr>
<td>SSA</td>
<td>single strand annealing</td>
</tr>
<tr>
<td>SSC</td>
<td>sodium chloride-sodium citrate</td>
</tr>
<tr>
<td>ssDNA</td>
<td>single-stranded DNA</td>
</tr>
<tr>
<td>TBE</td>
<td>tris-borate-EDTA</td>
</tr>
<tr>
<td>TMAC</td>
<td>tetramethylammonium chloride</td>
</tr>
<tr>
<td>VCFS</td>
<td>velocardiofacial syndrome</td>
</tr>
<tr>
<td>VNTR</td>
<td>variable number tandem repeats</td>
</tr>
</tbody>
</table>
Abstract

DNA INSTABILITY IN THE HUMAN α-GLOBIN GENE CLUSTER

Kwan-Wood Gabriel Lam

Ectopic recombination is an essential process that creates gene families, causes copy number variation and generates DNA rearrangements, sometimes leading to genetic disorders. Despite its importance, little is known about the dynamics and processes of aberrant crossover in humans. The human α-globin gene cluster is a classic system for studying ectopic recombination. DNA rearrangements such as single α-globin gene deletions (−α chromosomes) arise from unequal crossover between localised homologous regions, and are most likely favoured by malaria, leading to α⁺-thalassaemia. The α-globin gene cluster was therefore chosen as a test system for studying how copy number variation arises in the human genome. By developing single-DNA-molecule strategies including size fractionation and single-molecule PCR, de novo −α deletions and ααα duplications were both detected for the first time. These rearrangements occur both in blood and in sperm. Analyses of these mutants show that they are generated by two distinct mechanisms. The major pathway involves ectopic sister chromatid exchange, the frequencies of which appear to be strongly influenced by mutational mosaicism. These rearrangements were common in both blood and sperm, while meiotic exchanges between homologous chromosomes were restricted to the germ-line and with lower frequencies. There was significant reciprocity of deletion and duplication processes, with respect to ectopic exchange breakpoints, haplotype symmetries and recombination frequencies. This indicates that these mutants were most likely generated by a reciprocal intermolecular recombination pathway. However, there was also evidence for extrachromosomal circles, which might suggest the existence of an additional intramolecular pathway that plays only a minor role in generating deletions. The surprisingly high frequencies of de novo deletion and duplication suggest that significant selective forces must have acted against individuals with −α or ααα chromosomes to maintain a constant and high level of normal αα chromosomes in most malaria-free populations, and with low frequencies of −α and ααα chromosomes.
Chapter 1
Introduction

1.1 Significance of meiosis
Mutation and meiotic recombination are the two major processes that create biological differences and uniqueness between individuals. These two processes, increasing genetic diversity by creating and reshuffling genetic variation, can generate neutral or beneficial characters to augment individuals’ survival fitnesses in the environment. However, these processes can sometimes generate deleterious effects such as inherited diseases. Meiotic recombination is particularly important for the proper segregation of homologous chromosomes in meiosis in which newly assorted genetic information is segregated properly from diploid to haploid cells. Therefore, studies of meiosis are not only crucial for understanding the survival fitness but are also fundamental for illuminating the history and origin of a species. Most studies of recombination have been focussed on lower eukaryotes, mainly in yeast, for example *Saccharomyces cerevisiae* because of the short life cycle and the ability to recover the full inventory of meiotic recombination products in tetrad analysis. Hence, this chapter will focus on *S. cerevisiae* as a reference to the features of meiotic recombination in humans.

1.1.1 Important stages in meiosis
Meiosis shares a certain level of similarity with mitosis; however, two consecutive divisions and the presence of meiotic crossover in meiosis, generating four genetically unique haploid progeny instead of two identical diploid daughter cells, create distinct differences from mitosis. A specialised phase of cell division (meiosis I) is only found in germ cells and can be divided into four different stages (Fig. 1.1). Since processes such as pairing of homologous chromosomes, assembly of the synaptonemal complex (SC), meiotic recombination and formation of chiasmata occur in the lengthy prophase of meiosis I, this phase has been the main focus for detailed studies.

Prophase I can be subdivided into the five stages of leptotene, zygotene, pachytene, diplotene and diakinesis. In leptotene, chromatin condenses and homologous chromosomes start to search for each other and align; double strand breaks (DSBs) appear as an initiation step for meiotic recombination. In zygotene, chromosome synapsis initiates, and the formation of SCs and recombination nodules are observed in the centre of the complex. These combined homologous chromosomes are termed bivalents at this
stage. In pachytene, chromosomes are fully synapsed and non-sister chromatids of homologous chromosomes exchange segments. During diplotene, the SC degrades and homologous chromosomes begin to repel each other while chiasmata maintain their association at this stage. Finally, in diakinesis, chromosomes recondense in preparation for metaphase I.

**Fig. 1.1.** Major stages in meiosis I. Only the four main stages of prophase, metaphase, anaphase and telophase are shown. Parental and maternal homologues are highlighted in blue and red, respectively.
1.1.2 Main events of meiosis I

1.1.2.1 Homologue pairing

Homologue pairing occurs prior to chromosome synapsis (Weiner and Kleckner 1994; Nag et al. 1995), and is an essential process for promoting initiation of recombination particularly in regions where the frequency of DSBs is reduced by DNA heterologies (Rocco and Nicolas 1996). However, other evidence suggested that homologue pairing might not be obligatory for DSB formation (De Massy et al. 1994). In most organisms but excluding *Caenorhabditis elegans* (Zetka and Rose 1995), sites of chromosome pairing appear to be numerous and are widely distributed along the lengths of chromosomes (Weiner and Kleckner 1994). Additionally, it has been reported that telomeric regions of chromosomes play a crucial role in homologue alignment in many organisms. The formation of a bouquet (attachment of telomeres to the nuclear envelope) brings homologous subtelomeric regions into parallel alignment and may facilitate the pairing of homologues (Roeder 1997; Walker and Hawley 2000). Several lines of evidence indicate that, especially in organisms with bouquet formation, meiotic telomere proteins are required for normal chromosome synapsis (Conrad et al. 1997; Trelles-Sticken et al. 2000; Wu and Burgess 2006). The loss of proteins such as NDJ1 delays chromosome synapsis and thus increases the level of failure in meiotic recombination.

1.1.2.2 Double strand break repair model

Although the DSB repair model has only been well characterised in *S. cerevisiae*, it is believed that DSBs are initiators of meiotic recombination in most organisms (Lin and Smith 1994; Dolganov et al. 1996). This DSB repair pathway can be divided into several stages, with many proteins involved (Fig. 1.2). DSB formation is known to be a non-random process in which clustered sites of initiation (known as hot spots) are located in promoters and with some in coding sequences (Lichten and Goldman 1995; Bullard et al. 1996; Baudat and Nicolas 1997). These sites are usually located in open chromatin regions that are hypersensitive to DNase I and micrococcal nuclease (MNase) (Fan and Petes 1996; Keeney and Kleckner 1996). These DSBs appear not to be controlled by any specific sequence motifs but are instead scattered across a 50–200 nucleotide tract within a nuclease-hypersensitive region (Liu et al. 1995; Xu and Kleckner 1995).
1.1.2.2i Initiation of double strand breaks

The Spo11 protein, which is homologous to a family of type II topoisomerase, initiates meiotic DSBs by breaking both strands of a DNA molecule and covalently linking the 5' ends of the breaks to a tyrosine within itself (Keeney and Kleckner 1995; Keeney et al. 1997). Although the Spo11 protein plays a central role in initiation, this process cannot be completed without the involvement of other proteins (Fig. 1.2i) (Johzuka and Ogawa 1995; Roeder 1997).

1.1.2.2ii Exonucleolytic resection

Three genes (RAD50, MRE11 and COM1/SAE2) are essential in the resection process (Alani et al. 1990; McKee and Kleckner 1997; Nairz and Klein 1997). Mutation studies indicate that Rad50 and Mre11 proteins are both required for the formation and processing of meiotic DSBs since they work with each other in a two-hybrid system (Johzuka and Ogawa 1995; Trujillo and Sung 2001). After double-strand cleavage, the 5'-attached Spo11 protein is removed, along with a short oligonucleotide (Keeney et al. 1997), by the Com1/Sae2 protein along with the Rad50/Mre11/Xrs2 complex (Rattray et al. 2001) and the exonucleolytic resection is subsequently initiated by the Rad50/Mre11/Xrs2 complex (Cao et al. 1990; Nairz and Klein 1997). This leaves single-stranded tails with 3' termini exposed, ready for invasion of the homologous chromosome (Fig. 1.2ii).

1.1.2.2iii Strand invasion and double Holliday junction formation

Following resection, one of the single-stranded tails invades a homologous duplex [single-end invasion (SEI)] and displaces one strand of the equivalent sequence in the homologue by DNA synthesis. Joint molecules (JMs) consisting of sequences from two parental duplexes are formed (Fig. 1.2iii). DNA synthesis is further extended; the invading tail then ligates with the second resected end, subsequently resulting in two four-way DNA junctions called a double Holliday junction (dHJ) (Fig. 1.2iv). Four yeast proteins (Rad51, Rad55, Rad57 and Dmc1), which are homologues of the bacterial RecA strand exchange enzyme, are involved in this process (Shinohara et al. 1992; Schwacha and Kleckner 1997). Biochemical studies revealed that, in the presence of the Rad55/Rad57 heterodimeric complex, the Rad52 protein assists the entry of the Rad51 protein as well as strand exchange by dislocating the heterotrimeric single-stranded DNA-binding factor replication protein A (RPA) from single-stranded DNA (ssDNA) (Sung
1997b; Sung 1997a; New et al. 1998). Hyper-resected DSBs seen in dmcl mutants indicated that the Dmc1 protein contributes in the formation of JMs (Schwacha and Kleckner 1997) although the exact roles of this protein have not been clearly identified.

1.1.2.2iv Holliday junction resolution

Cleavage of a Holliday junction can be performed by resolvase or topoisomerase (Gilbertson and Stahl 1996; Stahl 1996). With different cleavage orientations (vertical or horizontal in Fig. 1.2iv and v), the DSB repair system can give rise to either crossover or noncrossover products. It is generally believed that reciprocal (crossover) and non-reciprocal (gene conversion without exchange) products are generated by a common pathway of recombination via Holliday junction resolution (Storlazzi et al. 1995). However, it has been suggested that noncrossover recombinants can be formed by a pathway without the involvement of a Holliday junction intermediate (Allers and Lichten 2001). Similar to mitotic systems (Nassif et al. 1994; Paques and Haber 1999), noncrossover recombinants can arise by a pathway called ‘synthesis-dependent strand annealing’ (SDSA) (details in section 1.1.4.1) (Nassif et al. 1994).

1.1.2.2v Mismatch repair

The strong correlation between the level of heteroduplex DNA formation and the frequency of meiotic gene conversion indicates that mismatched base pairs are rapidly repaired with the genetic consequence of gene conversion after heteroduplex DNA is formed (Nag et al. 1995). This phenomenon is observed in JMs in which recombinant single strands instead of heteroduplex DNA are found, suggesting that a first round of mismatch correction occurs during the formation of JMs (Schwacha and Kleckner 1995; Stahl 1996). Furthermore, it has also been suggested that the sliding of Holliday junctions before the sealing of JM may help removing any mismatch in DNA duplexes. DNA heterologies on one strand of a chromatid in heteroduplex DNA can be replaced using the opposite strand as a template by the mismatch repair system. In S. cerevisiae, this correction results either in gene conversion or in restoration, and requires the presence of three homologues of the bacterial MutS protein (Msh2, Msh3 and Msh6) and two homologues of MutL protein (Pms1 and Mlh1) (Prolla et al. 1994; Marsischky et al. 1996). Aberrant DNA structures are recognised and repaired by a Msh2–Msh3 or Msh2–Msh6 complex in conjunction with a Pms1/Mlh1 heterodimer (Kolodner 1996).
1.1.2.2vi Choice of partner for double strand break repair

It is known that exchanges between non-sister homologues are preferred in meiosis whereas sister chromatid exchanges are favoured in mitotic recombinational repair in *S. cerevisiae* (Kadyk and Hartwell 1992). These phenomena, which have also been observed in mammalian cells, can be explained by the relative proximity of chromatids. Sister chromatids are attached by cohesion proteins and tightly adhered to each other until mitosis (Miyazaki and Orr-Weaver 1994). Homologues are not as close to each other as sister chromatids in the nucleus except during meiosis. Genetic analyses revealed that genes including *HOP1*, *RED1* and *RAD51* can significantly influence the choice of template for DSB repair (Schwacha and Kleckner 1997). For example, the *red1* and *hop1* mutants reduce DSB formation, fail to make SCs and subsequently reduce the frequency of the formation of JMs between homologues but not sister chromatids (Rockmill and Roeder 1990; Roeder 1997; Smith and Roeder 1997).
Fig. 1.2. DSB repair model of meiotic recombination. Two homologous DNA molecules are shown in blue and red. Proteins involved in each step, based mainly on the study of S. cerevisiae, are listed. (i) DSB formation initiated by Spo11 and other proteins. (ii) DNA strand resection. 3' single-stranded tails with 3' -OH termini are created by 5' to 3' exonucleolytic digestion. (iii) Strand invasion and D-loop formation. A 3' single-stranded tail invades its equivalent sequence on the homologue and forms a displacement loop (D-loop). (iv) DNA synthesis, dHJ formation and resolution. The D-loop is extended by DNA repair synthesis primed by the invading strand. The second 3' single-stranded tail is also extended after annealing to the equivalent sequence on the D-loop. Strand breaks are ligated, creating a dHJ structure. The dHJ can be resolved by horizontal or vertical cleavage of a pair of symmetrical strands at each junction, resulting in noncrossover (a + b or A + B) or crossover (A + b or A + b) heteroduplex DNA (v) Involvement of mismatch repair (MMR). Heteroduplex DNA is corrected by MMR using homologous sequences on sister chromatids, removing mismatches from noncrossover and crossover products.
1.1.2.3 Formation of the synaptonemal complex

The SC is a meiosis-specific chromatin-associated protein structure that provides a closely linked association between homologous chromosomes. It appears during pachytene after the initiation of the Spo11 protein-catalysed DSBs (Meier and Gartner 2006). The SC, a tripartite ribbon-like structure, begins to form in leptotene and disassembles during diplotene. During prophase I, a single proteinaceous axis, called the axial element, develops between two sister chromatids of a single chromosome. The axial element is named instead a lateral element (LE) after the formation of an SC. The axial elements of two homologous chromosomes are connected by transverse filaments to form the SC structure. The mature SC is then completed by the presence of a central element, which lies parallel to and equidistant between the two LEs (von Wettstein et al. 1984; Page and Hawley 2004). This whole connection process is called chromosome synapsis and is believed to be a prerequisite of meiotic recombination in most eukaryotes. However, exceptions have been seen in studies of *Schizosaccharomyces pombe* (Kohli and Bahler 1994) and *Drosophila melanogastger* (McKim et al. 1998).

Several protein components of the central element of the SC, for example, the Zip1 protein of *S. cerevisiae* (Sym et al. 1993; Tung and Roeder 1998) and the SCP1/Syn1 complex in mammals (Dobson et al. 1994; Meuwissen et al. 1997), have been identified and characterised. Zip1 and SCP1/Syn1 proteins are components of the transverse filaments which only appear in synapsed chromosomes but not in asynapsed axial elements (Sym et al. 1993). Studies using the \textit{zip1} mutant of *S. cerevisiae* showed that the Zip1 protein is an essential structural element of the SC in maintaining crossover interference and providing the framework for other protein components (Sym and Roeder 1994). Recently, the Zip1 protein was found to promote centromere coupling, favouring homologue pairing and serving as sites of synapsis initiation (Tsubouchi and Roeder 2005). These proteins seem to play a significant role in the formation of the SC and the establishment of meiotic synapsis; however, exact mechanisms are still unclear.

Another set of proteins that are involved in the formation of the LE of the SC can be classified into two groups, cohesin and non-cohesin proteins. Cohesin proteins like Rec8, Smc1, Smc3, Scc1 and Scc3 (Michaelis et al. 1997; Hirano 2000; Eijpe et al. 2003) are responsible for the formation of the LE by building up an axial chromosome core where LE proteins can bind and assemble (Klein et al. 1999), and also promoting meiotic chromosome segregation (Buonomo et al. 2000). Likewise, some non-cohesin components such as SCP3/Cor1 and Red1 proteins are present in both asynapsed axial
elements and the LE of mature SCs (Dobson et al. 1994; Smith and Roeder 1997). These proteins also play an essential role for the formation of the LE and the assembly of other proteins responsible for continuous linear elements (Smith and Roeder 1997).

The formation of the SC is a critical process in meiosis; although a positive correlation between crossovers and sites of synaptic initiation has been seen in some organisms (Henderson and Keeney 2005), it is not a compulsory prerequisite of meiotic recombination (Sym and Roeder 1994; McKim et al. 1998). Possible roles for the SC have been proposed, for instance, ensuring the proper separation of homologues by converting crossovers into stable chiasmata (Heyting 1996) and stabilising pairing associations over the entire length of chromosomes during meiosis (MacQueen et al. 2002; Colaiacovo et al. 2003). However, little is known about how SCs and SC components fulfil their functions.

1.1.2.4 Recombination nodules

Recombination nodules are tiny and electron-dense structures that are associated with meiotic recombination events. These structures are present during the early stages of prophase I, and in intimate association with the SC (Carpenter 1975; Zickler and Kleckner 1999). Two types of nodules, early and late nodules, are classified according to their time of appearance, size, number and distribution.

Early nodules appear during leptotene and zygotene and are more frequent and evenly distributed than late nodules in most organisms (Zickler and Kleckner 1999). These nodules have been suggested as locating the sites of all strand exchange reactions (Roeder 1997). Two RecA-like proteins Dmc1 and Rad51, which are essential in strand invasion and Holliday junction formation during DSBs (described in section 1.1.2.2iii), are components of early nodules (Anderson et al. 1997). In addition, the identification of physical interaction between Rad51 and Rad52 proteins suggested that the Rad52 protein is also a component of early nodules (Shinohara et al. 1992).

Late nodules emerge in pachytene and their number seen in the SC is tightly related to the number of crossovers and chiasmata (Carpenter 1979). Given that late nodules exhibit interference, the even distribution observed in early nodules (in most organisms) has never been identified in late nodules (Zickler and Kleckner 1999). Proteins including Msh4, Msh5 and Mlh1 are components of late nodules in yeast (Hollingsworth et al. 1995; Novak et al. 2001). While the Mlh1 protein is required in mismatch repair (see section 1.1.2.2v), the roles of late nodule have been linked to
mismatch repair as well as the stabilisation or resolution of Holliday junctions (Roeder 1997).

1.1.2.5 Chiasmata

After the disassembly of the SC and the resolution of recombination intermediates, chiasmata are the only sites corresponding to reciprocal breakage and rejoining between two non-sister chromatids in proper chromosome segregation. It is believed that the formation of chiasmata between homologue pairs enhances proper chromosome segregation during meiosis, since the failure of obligate chiasma (at least one reciprocal exchange on every pair of chromosomes) (Lawrie et al. 1995) increases the chance of meiotic nondisjunction and aneuploidy (Ross et al. 1996). Two models describing how a chiasma achieves the binding function were proposed (Maguire 1974); the first one is that chiasma binder proteins, known as sister chromatid cohesins, maintain linkage of sister chromatids distal to chiasmata, while another one is that binder proteins act directly at chiasmata. Furthermore, studies in some organisms suggested that centromere-proximal exchanges are more effective in ensuring proper chromosome segregation than exchanges near telomeric regions (Mason 1976; Sherman et al. 1994). These observations lead to the suggestion that the effectiveness of chiasma binders decreases with the reduced length of the telomere-to-exchange interval possibly owing to the DNA sequences or chromatid structures at telomeric regions (Ross et al. 1996).

1.1.3 Crossover and gene conversion

In classic models describing molecular events of meiotic recombination, both crossover and noncrossover (gene conversion) products have been suggested to arise from a common mechanism via DSB initiation, SEI and dHJ formation but with different orientations of Holliday junction resolution (Meselson and Radding 1975; Szostak et al. 1983). Thus, the frequent association of crossover and gene conversion is always observed in yeast meiotic recombination. However, additional evidence challenges the suggestion that the relative frequencies of crossover and gene conversion are purely governed by the orientation of Holliday junction resolution. For example, mutations in meiotic recombination promoting proteins including Zip1, Zip2, Zip3, Mer3, Msh4 and Msh5 (ZMM proteins) reduce the levels of crossover but maintain high frequencies of gene conversion (Borner et al. 2004). These observations indicate that crossover events are formed by a more intricate mechanism instead. In addition, similar to suggestions by
Allers and Lichten (2001) and Hunter and Kleckner (2001), the decision whether to generate a crossover or a gene conversion is determined much earlier than the occurrence of Holliday junction resolution, possibly prior to or during the formation of SEIs depending on any alterations of ZMM proteins to SEI stability and dHJ formation (Borner et al. 2004; Whitby 2005). These findings support an 'Early crossover decision' model of meiotic recombination (Bishop and Zickler 2004). This model proposes a similar onset to DSB repair in which a primary DSB initiation is followed by a 5' to 3' resection and subsequently a 3' single-stranded tail invasion. With a nascent interaction formed by a DSB end and an invaded homologue, the process can either undergo a crossover pathway (typically a DSB repair pathway) or skip to a noncrossover pathway such as SDSA (Nassif et al. 1994) and the newly proposed 'D-loop nicking' pathway (Cromie et al. 2006).

1.1.4 Mechanisms for noncrossovers

1.1.4.1 The synthesis-dependent strand annealing model

The SDSA model was firstly proposed in the study of P-element-induced gap repair in *D. melanogaster* (Nassif et al. 1994). The transposable P-element uses a cut-and-paste process and creates breaks on both DNA strands during an excision of the element. It was found that the break is either repaired by copying the corresponding sequences from the sister chromatid or from a template on the homologue or on a non-allelic homologue (an ectopic site); recombinants with gene conversion arise in the latter case (Gloor et al. 1991; Nassif et al. 1994).

This model shares some initial steps with the classic DSB repair system (Figs. 1.2 and 1.3). For instance, both DSB repair and SDSA models are initiated by DSB formation. Following 5' to 3' resection, a 3' single-stranded tail invades an unbroken partner causing D-loop formation and subsequently DNA synthesis. Beyond this step, the SDSA model is no longer the same as the DSB repair model. The D-loop is disassembled by the displacement of the newly synthesised strand, which subsequently ligates with the other resected end (Fig. 1.3iv). The complex is then repaired by DNA synthesis along with ligation. Gene conversions without crossover are therefore generated, and can be further modified by mismatch repair of heteroduplex DNA. Given that only one 3' single-stranded tail is involved in invasion for D-loop formation (no dHJs formed), heteroduplex DNA is only found at the recipient locus whereas the donor template remains unchanged.
(Paques and Haber 1999). Under this SDSA model, the fidelity of gene conversion is very high (Gloor et al. 1991) and the repair ability is very efficient in that deletions or insertions can be completely copied as frequently as if a single base is substituted at the corresponding position instead (Nassif et al. 1994). These indicate that this model is responsible not only for the generation of gene conversions but also for duplications when ectopic sites are used for repair.

Only gene conversions but no crossovers are obtained in the simple SDSA model. However, modified SDSA models can yield conversions accompanied by crossover via the formation of dHJs (Ferguson and Holloman 1996; Paques and Haber 1999; Allers and Lichten 2001). Despite modified SDSA models producing crossovers as well as noncrossovers, the DSB repair model is still generally accepted to be a pathway for generating both crossovers and noncrossovers from a single recombination initiation event.
Fig. 1.3. The synthesis-dependent strand annealing (SDSA) model. (i–iii) These steps are the same as described in Fig. 1.2i–iii. (iv) DNA synthesis and branch migration. A D-loop is extended by DNA synthesis primed by the invading strand. The newly synthesised strand disassembles from the D-loop without formation of dHJ and ligates with the second resected end. The second 3' single-stranded tail is extended using the equivalent sequences on the same molecule instead of on the D-loop template. DSB ends are ligated and heteroduplex DNA is formed. (v) Mismatched sequences on heteroduplex DNA are corrected by MMR, resulting in noncrossovers only.

1.1.4.2 D-loop nicking model

While most molecular events of meiosis are believed to be universal in most organisms including budding yeast, distinct features of recombination were found in fission yeast. Although two Holliday junctions are a distinct intermediate in meiosis in most systems, the formation of a single Holliday junction was found in S. pombe instead (Cromie et al. 2006). Given the single Holliday junction predominantly observed in recombination intermediates in fission yeast using electron microscopy, a ‘D-loop nicking’ model was proposed by Cromie et al. (2006) (Fig. 1.4). The crucial step of this mechanism is that the
displaced donor strand (D-loop strand) is nicked before the capture of the second DSB end of the recipient without the formation of a Holliday junction. Further evidence suggested that the Mus81 protein is essential in resolving Holliday junctions in fission yeast (Boddy et al. 2001; Cromie et al. 2006) while it only gives a modest reduction of crossover frequency in \textit{mus81/mms4} mutants in budding yeast (de los Santos et al. 2003). These differences in yeast indicate that although mechanisms like DSB repair and the SDSA model have been suggested to be universal in most organisms, details including DNA intermediates and protein involvement may actually be very different between species.

\textbf{Fig. 1.4. The D-loop nicking model.} (i–iii) These steps are the same as the DSB repair model described in Fig. 1.2i–iii. (iv) Branch migration promoted by DNA extension of the invading strand. (v) D-loop nicking and HJ resolution. An extended D-loop is nicked, releasing the single-stranded DNA from the donor. This single-stranded DNA ligates with the second resected end without forming a second Holliday junction. Noncrossovers and crossovers are formed by horizontal (a) or vertical (A) cleavages of the single Holliday junction, respectively.
1.1.5 Non-homologous end joining pathway

Besides homologous recombination by processes such as the DSB repair, which has long been suggested to be a major pathway for the repair of DSBs in yeast and bacteria, illegitimate recombination is also an alternative for DSB repair. The non-homologous end joining (NHEJ) pathway is commonly used in mammalian cells where two ends of broken DNA molecules with little or no homology are joined together directly, and without using homologues or sister chromatids as repair templates (Derbyshire et al. 1994; Sargent et al. 1997). The NHEJ pathway is not the only route for DSB repair in mammalian cells; in fact, it accounts for ~60% of the total DSB repair events in random clone analyses, with the rest generated by other mechanisms including gene conversion and ectopic recombination (Liang et al. 1998; Johnson and Jasin 2000). This frequent involvement of the NHEJ pathway in mammalian cells strongly contrasts to what is observed in S. cerevisiae where homologous repair is the predominant pathway (Sargent et al. 1997). Since illegitimate repair is always accompanied by deletions or insertions, its rarity in yeast can be explained by the compact genome structure in which a small amount of deletions or insertions would easily lead to mutagenesis and gene dysfunction. Since mammalian genomes have more repetitive elements as well as large introns and intergenic regions, non-homologous repair is less likely to create detrimental consequences (Liang et al. 1998).

1.1.6 Regulation and distribution of meiotic recombination

1.1.6.1 Recombination checkpoint

To ensure that recombination intermediates are resolved before moving from prophase to metaphase, the recombination checkpoint plays a significant role in monitoring. This checkpoint involves numerous proteins, which are crucial throughout the recombination process. In S. cerevisiae, genes like ZIP1 and DMC1 have been proposed to take part in the recombination checkpoint since cell arrest is triggered in zip1 and dmc1 mutants at pachytene prior to reductional division by unrepaired DSBs or unresolved Holliday junctions (Bishop et al. 1992; Sym et al. 1993). When zip1 or dmc1 mutants are in conjunction with mutations of DSB initiation genes like SPO11, RAD50 and MEI4, the arrest is alleviated and mutants progress to sporulation. Similar observations were obtained in double mutants with red1 or mekl/mre4 and recombination intermediate mutations (RAD51, DMC1 or ZIP1 mutation) (Xu et al. 1997). This bypass indicates that
the checkpoint has to be triggered by recombination intermediates. These findings also suggest that this recombination checkpoint helps prevent chromosome segregation occurring before the completion of recombination, regulating the timing of the first meiotic division with initiation of recombination (Galbraith et al. 1997).

1.1.6.2 Crossover interference
It is known that crossovers and resulting chiasmata are distributed non-randomly in a genome. Specifically, the presence of a crossover at a particular locus may reduce the likelihood of the occurrence of nearby events. This phenomenon is called crossover interference, and while common in most organisms, its molecular basis remains unclear. Several lines of evidence indicate that crossover interference is controlled by the SC since failure to make the SC causes elimination of interference (Kohli and Bahler 1994; Heyting 1996). Genetic analyses suggested that mutations in ZIP1, which encodes a key component in the formation of SC (section 1.1.2.3), cause elimination of interference indirectly although they do not alter the functions of chiasma, and give at best only modest effects on meiotic recombination and sister chromatid cohesion (Sym and Roeder 1994). However, arguments against this hypothesis suggested that interference still occurs despite the absence of the Zip protein, implying that crossover interference can be regulated by more than one mechanism (Storlazzi et al. 1996; Fung et al. 2004).

Other mutations including taml and msh4 also reduce or eliminate interference (Ross-Macdonald and Roeder 1994; Chua and Roeder 1997). However, the relationship between the SC and crossover interference as well as the involvement of proteins still awaits elucidation.

1.1.6.3 Recombination hot spots in yeast
As mentioned earlier, DSBs do not occur randomly in a genome. In S. cerevisiae, DSBs appear to cluster into hot spots in which meiotic recombination is initiated. Around ten recombination hot spots have been analysed in yeast, although few have been studied in great detail, for instance, ARG4, HIS4 and HIS4LEU2 in S. cerevisiae (Sun et al. 1991; Fan et al. 1995; Hunter and Kleckner 2001) and ade6-M26 in S. pombe (Steiner et al. 2002). In some high-resolution studies (de Massy et al. 1995; Xu and Kleckner 1995; Xu and Petes 1996), DSBs were found to occur within a region of 100–500 bp. Although no consensus sequences have been found among these hot spots, DSB sites usually occur at intergenic regions rather than within genes (Gerton et al. 2000) and also at regions which
are DNase I-hypersensitive (Wu and Lichten 1994). In addition to these molecular surveys, directional gene conversion has been found at recombination sites, giving a loss rather than a gain of hot spots in some heterozygous sites (Nicolas et al. 1989; Detloff et al. 1992). This observation supports the DSB repair model in that initiation sites usually act as a recipient during heteroduplex formation.

Besides these general properties shared among hot spots, special features in some hot spots essential for recombination activity have been observed. Firstly, it is known that the binding of specific transcription factors is critical for hot spot activity at some hot spots; these hot spots are referred to as α-hot spots. Studies using transcription factor (Bas1, Bas2 and Rap1) mutants or mutations at the corresponding transcription factor binding sites suggested that the HIS4 hot spot requires the binding of a transcription factor with an intact activation domain for hot spot activity (White et al. 1993; Kirkpatrick et al. 1999). Similarly, the binding of heteromeric transcription factor Atfl/Pcr1 (Mts1/Mts2) is also essential for ade6-M26 hot spot activity in S. pombe (Kon et al. 1997). Another type of hot spot, which is not associated with transcription factors but instead only nuclease-sensitive sequences, is called β-hot spots. A repeating tract like (CCGNN) is not a known binding site for any transcription factor but is hypersensitive to DNase I (Wang and Griffith 1996). Lastly, γ-hot spots refer to those hot spots associated with high G + C base composition.

1.2 Allelic recombination in humans

In yeast, meiotic recombination can be studied directly and efficiently with all possible recombinant outcomes observed by tetrad analysis. However, owing to differences in biological life cycle and genome complexity, it is impossible to achieve the same goal in humans. Based on the knowledge developed from lower eukaryotes, studies of meiotic recombination in humans can be analysed using more indirect approaches. These methods can estimate ratios of genetic to physical distance as well as predict recombination frequencies and hot spot locations in the human genome.

1.2.1 Pedigree studies

Given the low frequency of meiotic recombination event per unit of physical distance (~1 cM/Mb) (Kong et al. 2002), it is difficult to obtain detailed analyses of recombination events in the human genome by traditional pedigree analysis. Classic linkage analysis by
building up a genetic map based on pedigree data can establish a correlation between physical and genetic distances, but with a low resolution, at the megabase level. This mapping is usually done by genotyping a large number of polymorphic makers such as short tandem repeats in nuclear families with two to three generations. With extensive information obtained from individuals with different sexes and generations, crossover rates related to sex and age variations have been characterised (Dib et al. 1996; Broman et al. 1998; Kong et al. 2002; Kong et al. 2004). For instance, crossover rates in females are much higher than in males at centromeric regions, whereas higher rates at telomeres are obtained in males than in females. In addition, by relating recombination data to sequence content, recombination deserts and jungles have also been located (Yu et al. 2001; Kong et al. 2002). Although these analyses cannot give fine-scale information on meiotic events, they at least reveal general properties of meiotic recombination across the genome.

### 1.2.2 Linkage disequilibrium analysis

Linkage disequilibrium (LD) analysis is a study of non-random associations between alleles at different loci. If a particular allele at one locus is found along with another allele at a second locus on the same chromosome more often than it is expected on the basis of allele frequency, these two loci are said to be in linkage disequilibrium (Ardlie et al. 2002). The frequency of decay of the association depends mainly on the time of mutation and the subsequent level of recombination between the loci. Thus, a genomic region with strong LD indicates that it has not been actively involved in recombinational activity historically. Therefore, discrete haplotype blocks in LD analysis of the human genome reflect regional variations in recombination frequency, the existence of recombination hot spots and also human evolutionary history.

Several studies using single nucleotide polymorphism (SNP) genotyping data on different populations suggest that the general picture of the human genome is in clusters of haplotype blocks (Daly et al. 2001; Reich et al. 2001; Gabriel et al. 2002; Phillips et al. 2003). The size of these haplotype blocks varies substantially and is significantly different among populations; in general, the blocks are shorter in African and African-American samples, whereas longer blocks are found in European and Asian populations (Reich et al. 2001; Gabriel et al. 2002). These findings agree with the hypothesis of a single 'out of Africa' origin (Reich and Goldstein 1998; Ingman et al. 2000) and point out that demographic factors including genetic drift and population bottlenecks may contribute to the differences. Contrastingly, although patterns of LD can be disrupted by other factors
like natural selection, admixture and allele frequencies of markers (Wall 2001; Ardlie et al. 2002), the consistent block boundaries and specific haplotypes across populations (Gabriel et al. 2002), and the remarkable similarity of LD patterns in the MHC II region in three populations with different demographic histories (Kauppi et al. 2003) indicate that recombination is the key player in the creation of haplotype blocks.

1.2.3 Coalescent analysis
An alternative for estimating recombination rates and locating hot spots is to infer historical recombination events from contemporary population genotype data using coalescent simulation (Fearnhead et al. 2004). This study, using a full-likelihood based approach (Fearnhead and Donnelly 2001), gives results generally in good agreement with direct estimates from sperm typing (Schneider et al. 2002) and other population-based surveys (Smith et al. 1998; Wall et al. 2003) for estimating recombination rates and locating hot spots. More recently, a high-resolution genetic map of the entire human genome has been developed using coalescent-based methods (McVean et al. 2004; Myers et al. 2005; Spencer et al. 2006) on SNP marker data from population screening established by the International HapMap Project (Gibbs et al. 2003). These studies not only provide strong agreement with pedigree studies and LD analyses but also identify recombination hot spots, together with motif and sequence context potentially involved in regulating hot spot activity.

However, additional recombination hot spots in regions of strong LD within a 206-kb region of chromosome 1q42.3 were discovered using high-resolution sperm typing (see section 1.2.5.5) (Jeffreys et al. 2005). This study indicates limitations of coalescent analysis and suggests that some hot spots, which have significant differences between sperm crossover frequencies and historical recombination rates, may have evolved very recently and without causing substantial LD breakdown.

1.2.4 Single sperm typing techniques
Although LD and coalescent analyses can predict the locations of putative recombination hot spots throughout the human genome, the verification of hot spots can only be achieved using high-resolution sperm typing techniques at the molecular level. There are two main classes of sperm typing techniques: single and multiple sperm typing. Single sperm typing was first used to identify recombinants using single cell sorting strategies (Cui et al. 1989; Lien et al. 1993) followed by the genotyping of PCR-amplified
polymorphic markers (Arnheim et al. 2003). Later, this technique was more broadly used for locating recombination hot spots (Hubert et al. 1994; Cullen et al. 2002). DNA from lysed individual sperm is first subjected to whole-genome amplification by multiple rounds of random-primed PCR amplification (Zhang et al. 1992; Dean et al. 2002). The location of recombination hot spots is then narrowed down by genotyping interior markers. However, the detection of recombinants using this method is highly limited by de novo frequencies of allelic recombination. For instance, with a typical hot spot crossover frequency in the male germ-line of ~10^{-4} per sperm, 10000 sperm would have to be screened to yield just one crossover event. This is not only extremely laborious, but is also limited by significant levels of sperm mis-genotyping.

### 1.2.5 Allelic recombination hot spots identified by multiple sperm typing

In contrast to the single sperm typing strategy, recombinants are selectively amplified in total sperm DNA in multiple sperm typing (Kauppi et al. 2004). DNA aliquots are diluted to comprise no more than one recombinant molecule that can be amplified by allele-specific PCR at targeted regions (Jeffreys et al. 1994; Jeffreys et al. 1998a). Similarly, the resolution of hot spots can be trimmed to the sub-kilobase level, depending on distances between polymorphic markers in and around the hot spots in individuals analysed.

All high-resolution studies of hot spots have used multiple sperm typing. This method has widely been used in analysing not only crossover hot spots but also other recombination events including unequal crossover and gene conversion (Han et al. 2000; Guillon and de Massy 2002; Jeffreys and May 2004). To date, 27 allelic recombination hot spots have been characterised systematically in humans (Jeffreys et al. 1998a; Jeffreys et al. 2001; Jeffreys and Neumann 2002; May et al. 2002; Kauppi et al. 2004; Jeffreys et al. 2005; Holloway et al. 2006; A. Jeffreys and I. Berg, unpublished). Rather than collecting data at random across a selected DNA region, all hot spots were analysed at a specific targets showing LD breakdown and thus evidence of a putative recombination hot spot.

#### 1.2.5.1 Minisatellite-associated hot spot

Minisatellite MS32 is a classic human GC-rich VNTR with 29-bp repeats which is actively involved in conversion and crossover events preferentially at one end of the repeat array (Jeffreys et al. 1994; Jeffreys et al. 1998b). Based on local LD breakdown in the upstream region of MS32, sperm typing was performed across this region and a
meiotic recombination hot spot was identified (Jeffreys et al. 1998a). Since minisatellite MS32 is located at the boundary of the hot spot and its crossover and conversion activities are affected by a particular SNP allele in the flanking region, it has been suggested that the minisatellite instability is driven by the nearby hot spot instead of within the repeat array itself, thus creating polarised instability within the array.

1.2.5.2 Hot spots within the major histocompatibility complex region

In the class II major histocompatibility complex (MHC) region (~200 kb), six hot spots including DNA1, DNA2, DNA3, DMB1, DMB2 and TAP2 were detected (Jeffreys et al. 2000; Jeffreys et al. 2001). These hot spots are not distributed randomly but with 1–7 kb of DNA separating each hot spot within a cluster and with clusters separated by 60–90 kb. As with the minisatellite MS32 hot spot (Jeffreys et al. 1998a), almost all sperm crossovers are simple exchanges with patchy gene conversion rarely seen. The lengths of the extended domains of LD (distances between hot spot clusters) are similar to estimates of north-European LD block length elsewhere in the human genome (Reich et al. 2001) and also to the patterns of 10–100-kb long LD blocks in a 500-kb interval on chromosome 5 (Daly et al. 2001). The patterning of LD blocks and associated hot spots observed in the MHC may therefore reflect more generally haplotype patterns elsewhere in the human genome (Jeffreys et al. 2001).

1.2.5.3 SHOX hot spot in the pseudoautosomal region

It is known that crossover between the human sex chromosomes during male meiosis is restricted to the terminal pseudoautosomal regions PAR1 and PAR2. PAR1 is a 2.6-Mb Xp/Yp pseudoautosomal region where an obligatory meiotic exchange occurs, creating a male-specific recombination hot region with a recombination rate ~20-fold higher than the genome average (Lien et al. 2000). With the availability of genomic sequence, sperm analyses were performed on the SHOX gene in the PAR1 region, ~500 kb from the Xp/Yp telomeres. As with the MHC region, crossovers were not randomly distributed but instead clustered into a hot spot with substantial activity (190–370 cM Mb⁻¹) in the SHOX gene (May et al. 2002). Extreme LD breakdown across the hot spot was observed in three populations tested, providing further evidence that the LD breakdown is most likely caused by the hot spot and not by the retention of randomly-distributed historical crossovers in contemporary populations.
1.2.5.4 β-globin hot spot

A putative hot spot has long been suggested in the human β-globin gene cluster (Chakravarti et al. 1984; Smith et al. 1998; Schneider et al. 2002). Linkage disequilibrium analyses have shown that this putative hot spot is active, exceeding the genome average rate of recombination by 3–30 times (Chakravarti et al. 1984; Smith et al. 1998). Recently, detailed sperm typing analyses have shown that the β-globin hot spot is the most active autosomal hot spot yet characterised (Holloway et al. 2006). In addition, although this hot spot locates in a region between the δ and β-globin genes, the rarity of Lepore-type deletions suggests that ectopic recombination is very unlikely to be triggered by the localised crossover hot spot (see section 1.3.5.2).

1.2.5.5 Hidden hot spots

Seven hot spots, NID3, NID2a, NID2b, NID1, MST1a, MST1b and MST2 were identified in a 206-kb region of chromosome 1 (a region including the minisatellite MS32 hot spot) in sperm analyses (Jeffreys et al. 2005). Although most of these hot spots were predicted from LD and coalescent analyses, significant differences between sperm crossover frequencies and historical recombination rates have been observed. For instance, the NID3 hot spot was predicted to be weak from the coalescent analysis but proved to be the most intense in sperm. Other hot spots like NID2b and MST1a, which have intermediate recombination rates, were found to locate in regions of strong LD, indicating that regions of strong marker association are not necessarily recombinationally inactive. Likewise, not all regions of LD breakdown contain (sperm) recombination hot spots (Kauppi et al. 2005). These results suggest that the evolution of hot spots may be a rapid process that can leave no full mark on haplotype diversity in certain populations. Similar arguments have been proposed in the studies of fine-scale recombination rates at orthologous loci in humans and chimpanzees using coalescent analysis (Winckler et al. 2005). No correlation in recombination patterns between humans and chimpanzees has been observed, suggesting that fine-scale recombination rates evolve rapidly, to an extent disproportionate to the change in DNA sequence.

1.2.6 Properties of allelic recombination hot spots

These hot spots, identified by high-resolution sperm typing, share some similarities in respect to their width, crossover distributions and clustering (Table 1.1). In contrast, despite the evidence of the motif CCTCCCT identified in some hot spot THE1A/1B
elements (Myers et al. 2005), these sperm hot spots do not obviously share any sequence similarity, and cannot be grouped as in yeast (Nishant and Rao 2006). Thus, the occurrence of human hot spots cannot be purely predicted by sequence content.

Despite the fact that crossovers are normally distributed within a narrow hot spot, with spreading of crossover sites at both ends, and usually with a similar exchange frequency between both haplotypes in most of these hot spots, exceptions have been reported. Although exchanges with reciprocal frequencies were obtained in the DNA2 hot spot in the MHC region, asymmetric locations of exchange points were identified in same men (Jeffreys and Neumann 2002). It has been suggested that a heterozygous SNP at the centre of the hot spot is sufficient to cause this crossover asymmetry, with one allele preferentially initiating crossover, promoting biased gene conversion and over-transmission of the recombination-suppressing allele. Similar cases have also been seen in the minisatellite MS32 hot spot and the NID1 hot spot, in which differences in crossover distribution can be regulated by one SNP at a heterozygous locus within the hot spot (Jeffreys et al. 1998a; Jeffreys and Neumann 2005). These findings strengthen the 'hot spot paradox' whereby hot spots will eventually become extinct by recombinational meiotic drive in favour of recombination-suppressing alleles as predicted by the DSB repair model (Boulton et al. 1997; Pineda-Krch and Redfield 2005).

As in yeast (Schultes and Szostak 1990), human crossover hot spots are also hot spots for gene conversion without exchange (Jeffreys and May 2004; Jeffreys and Neumann 2005). Surveys on human crossover hot spots DNA3, DMB2, SHOX and NID1 showed that short conversion tracts spread bidirectionally from the peak of crossover hot spots. This indicates that crossovers and conversions are most likely triggered by the same recombination initiating events, despite the ratio of non-crossover conversion events to crossovers being highly variable between hot spots (Jeffreys and May 2004; Holloway et al. 2006).

In silico analyses of recombination rates and hot spots across the human genome suggested that a 7-nucleotide oligomer CCTCCCT may play an important role in regulating hot spot activity (Myers et al. 2005). This motif is more frequent in THE1A/B elements in putative hot spots than in those found elsewhere in the human genome. Surprisingly, this motif is also found within the DNA2 hot spot, with a change from T to C in its third position appearing to suppress sperm recombination activity (Jeffreys and Neumann 2002). However, variable recombination frequencies in individuals with identical local sequences have been revealed in detailed studies of the MSTM1a and
MSTM1b hot spots (Neumann and Jeffreys 2006). These findings suggest that distal regulators or epigenetic factors, as well as local DNA sequences, are important in controlling hot spot activity.

**Table 1.1. Properties of recombination hot spots analysed by high-resolution sperm typing**

<table>
<thead>
<tr>
<th>Hot spot</th>
<th>Chromosome no.</th>
<th>No. of men tested</th>
<th>No. of sperm crossovers typed</th>
<th>95% width (kb)</th>
<th>Mean peak activity (cM Mb⁻¹)</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA1</td>
<td>6</td>
<td>2</td>
<td>69</td>
<td>1.9</td>
<td>0.5</td>
<td>Promoter</td>
</tr>
<tr>
<td>DNA2</td>
<td>6</td>
<td>3</td>
<td>237</td>
<td>1.3</td>
<td>3.7</td>
<td>Intergenic, Alu</td>
</tr>
<tr>
<td>DNA3</td>
<td>6</td>
<td>6</td>
<td>661</td>
<td>1.2</td>
<td>130</td>
<td>Intergenic, Alu</td>
</tr>
<tr>
<td>DMB1</td>
<td>6</td>
<td>2</td>
<td>36</td>
<td>1.8</td>
<td>3.1</td>
<td>Intragenic, introns/exon?</td>
</tr>
<tr>
<td>DMB2</td>
<td>6</td>
<td>3</td>
<td>358</td>
<td>1.2</td>
<td>28</td>
<td>Intergenic</td>
</tr>
<tr>
<td>TAP2</td>
<td>6</td>
<td>2</td>
<td>141</td>
<td>1.0</td>
<td>5.8</td>
<td>Intronic</td>
</tr>
<tr>
<td>NID3</td>
<td>1</td>
<td>3</td>
<td>1094</td>
<td>2.0</td>
<td>70</td>
<td>Alu Jo element downstream of NID</td>
</tr>
<tr>
<td>NID2a</td>
<td>1</td>
<td>2</td>
<td>302</td>
<td>1.4</td>
<td>10</td>
<td>Single copy DNA in NID intron 12</td>
</tr>
<tr>
<td>NID2b</td>
<td>1</td>
<td>2</td>
<td>107</td>
<td>1.1</td>
<td>4</td>
<td>Single copy DNA in NID intron 12</td>
</tr>
<tr>
<td>NID1</td>
<td>1</td>
<td>7</td>
<td>1345</td>
<td>1.5</td>
<td>70</td>
<td>Alu Yc5 element in NID intron 4</td>
</tr>
<tr>
<td>MS32</td>
<td>1</td>
<td>3</td>
<td>250</td>
<td>1.5</td>
<td>40</td>
<td>Intergenic, in RTLV-LTR</td>
</tr>
<tr>
<td>MSTM1a</td>
<td>1</td>
<td>2</td>
<td>179</td>
<td>1.6</td>
<td>9</td>
<td>Intergenic, in single copy DNA</td>
</tr>
<tr>
<td>MSTM1b</td>
<td>1</td>
<td>2</td>
<td>374</td>
<td>2.1</td>
<td>16</td>
<td>Intergenic, in single copy DNA</td>
</tr>
<tr>
<td>MSTM2</td>
<td>1</td>
<td>3</td>
<td>46</td>
<td>1.3</td>
<td>16</td>
<td>Intergenic, in single copy DNA</td>
</tr>
<tr>
<td>β-globin</td>
<td>11</td>
<td>2</td>
<td>284</td>
<td>1.2</td>
<td>200</td>
<td>Promoter, origin of replication</td>
</tr>
<tr>
<td>SHOX</td>
<td>XY</td>
<td>3</td>
<td>527</td>
<td>2.2</td>
<td>370</td>
<td>Intragenic</td>
</tr>
</tbody>
</table>

Information adapted from Jeffreys et al. (2001), May et al. (2002), Jeffreys et al. (2005) and Holloway et al. (2006). 95% width is the width of each hot spot within which 95% of crossovers occur, and was determined by normal-distribution fitting. Mean peak activity was the peak value estimated from best-fit normal distributions.

### 1.3 Ectopic recombination

As mentioned in previous sections, there are basically two modes of DSB repair: homologous recombination and non-homologous end joining (NEHJ) (Paques and Haber 1999). Apart from major routes including DSB repair using allelic sequence and the SDSA models, DSBs can also be repaired by other pathways including ectopic recombination, the single-strand annealing (SSA) and the break-induced replication (BIR) mechanisms, some of which can give rise to non-reciprocal products. Some of these mechanisms, which are usually used to explain recombination between repeated sequences, share one obvious similarity in that a resected DSB end anneals to a non-allelic homologous region on either the same or a different chromosome during the
invasion process. They therefore produce chromosome rearrangements including deletions and duplications, creating copy number variation in the genome.

1.3.1 Significance of ectopic recombination

Ectopic recombination, also known as non-allelic homologous recombination (NAHR), is a DSB repair mechanism that involves DNA misalignment and interaction between non-allelic homologous regions (repetitive DNA sequences). Depending on repeat orientations, different types of aberrant rearrangements can be obtained (Fig. 1.5). For example, unequal crossover between direct repeats can generate deletions and either duplications or excised DNA circles (Fig. 1.5A). In contrast, exchanges between inverted repeats give inversions instead (Fig. 1.5B).

The formation of heteroduplex DNA in both allelic and ectopic recombination provided the first evidence that these two recombination processes do share some similar mechanisms (Nag and Petes 1990). This argument is further supported by the similar frequencies of ectopic and allelic events observed in analyses of recombination between artificial inserts in the yeast genome (Lichten et al. 1987; Haber et al. 1991). Despite these similarities shared between allelic and ectopic recombination, some unique features that are essential for efficient exchange particularly in ectopic recombination have been identified. Given that the most extensive surveys of ectopic recombination have been done on budding yeast, the properties described below are based on the studies of S. cerevisiae.
Fig. 1.5. Ectopic recombination between repeats. Repeat DNA sequences are depicted as arrows. Unique DNA sequences are marked by letters with allelic sequences highlighted in different cases. (A) Ectopic recombination between direct repeats. (i) Intermolecular exchange generating reciprocal duplications and deletions. (ii) Intramolecular recombination giving a deletion and an excised circle. (B) Ectopic recombination between inverted repeats. (i) Intermolecular recombination can generate inversions or acentric plus dicentric chromosomes, while (ii) intramolecular recombination generates inversions only.

1.3.1.1 Relationship between recombination frequency and copy number
Ectopic recombination occurs between repeated sequences, but whether the efficiency of recombination depends on the number of repeats is arguable. Successful homologous recombination is not rate-limited by the search for a homologous partner in transformation experiments in mammalian cells (Zheng and Wilson 1990). However, a positive correlation between recombination frequencies and the number of repeats has been shown in a similar study in *S. cerevisiae* (Wilson et al. 1994), despite the fact that a previous study had suggested that ectopic gene conversions do not increase in proportion to the number of donor repeats in budding yeast (Haber et al. 1991).

1.3.1.2 Physical distance between repeats
Ectopic recombination can occur as frequently as allelic recombination during meiosis in artificial repeat studies of *S. cerevisiae*. However, the analysis of position effects between DNA repeats on both meiotic allelic and ectopic recombination showed that the frequency
of recombination between heteroallelic sequences in different genomic locations can be very variable (Lichten et al. 1987). Although allelic recombination can share a similar frequency with ectopic recombination between non-homologous chromosome in mitosis, distinct features between these two processes have been observed (Lichten and Haber 1989). When _LEU2_ copies were inserted into different locations on parental homologues, no significant variations were observed in mitotic allelic recombination. This result, in contrast to the previous meiotic study (Lichten et al. 1987), indicates that location has little effect on mitotic recombination. In contrast, mitotic ectopic recombination is distance-dependent in that the recombination rate between repeats located 20 kb apart on the same chromosome is about ten times higher than the rate between repeats on different chromosomes (Lichten and Haber 1989). Similar findings from other studies (Roeder et al. 1984; Rudin and Haber 1988; Goldman and Lichten 1996) also agree that the physical distance between repeats is a major controlling factor in the frequency of ectopic recombination.

### 1.3.1.3 Length of homology

It has been shown in many organisms that lengthy homology between repeats is one of the major requirements for efficient ectopic recombination. It was first suggested in the study of _E. coli_ (Shen and Huang 1986) that there is a linear relationship between lengths of homology and recombination frequencies. The length is called the 'minimal efficient processing segment' or MEPS. Once the homology length between repeats is shorter than MEPS, recombination frequencies drop rapidly but still occur inefficiently (Gonda and Radding 1983; Watt et al. 1985). Similar phenomena have been seen in mitotic studies in both yeast and mammalian cells, indicating that ~200–300 bp of homology is essential for homologous recombination (Rubnitz and Subramani 1984; Jinks-Robertson et al. 1993), whereas a shorter MEPS was predicted in _E. coli_ (Watt et al. 1985).

### 1.3.1.4 Degree of sequence divergence between repeats

The rate of intrachromosomal recombination in mammalian cells is mainly determined by the length of uninterrupted homology (Waldman and Liskay 1988). However, sequence divergence between repeats can also be a potent barrier to homologous recombination (Datta et al. 1997). Experiments using both wild type and mismatch repair defective strains of _S. cerevisiae_ showed that recombination rates drop sharply with increasing
sequence divergences. They also indicated that the reduction of recombination is dependent on mismatch repair system when the sequence divergence is less than 10%.

1.3.1.5 Sequence motifs
Since ectopic recombination occurs between homologous sequences, most studies have been focussed on transposable elements like P-elements in Drosophila (Nassif et al. 1994), mariner elements (Lohe et al. 2000) and Ty elements in yeast (Roeder et al. 1984). However, no common recombinogenic sequence motifs have ever been identified within these elements.

1.3.1.6 Ectopic gene conversion
As with the DSB repair model developed mainly for allelic recombination, ectopic recombination can generate not only reciprocal products like deletions and duplications but also non-reciprocal gene conversions. However, the mechanisms controlling the proportion of reciprocal and non-reciprocal products have yet to be identified. Analyses on budding yeast showed that about half of the ectopic recombinants are crossovers accompanied with gene conversion (Lichten et al. 1987), sharing a similar extent with allelic gene conversions (Jinks-Robertson and Petes 1986). In contrast, most intrachromosomal ectopic recombinants (~80%) observed in studies using mammalian cells were gene conversions but unassociated with crossovers (Liskay and Stachelek 1986; Liskay et al. 1987).

1.3.2 Other mechanisms for genome rearrangement
Ectopic recombination is one of the major mechanisms creating copy number variation. However, other mechanisms including SSA, SDSA (see section 1.1.4.1) and replication slippage can also give similar rearrangements but usually in a non-reciprocal fashion.

1.3.2.1 Single strand annealing model
The SSA model was first proposed in a study of extrachromosomal recombination in a mammalian repair system (Lin et al. 1984). Following the formation of a DSB between or within two directed repeats, two long complementary single-stranded DNAs are produced by the 5’ to 3’ strand resection of the DSB ends. A recombination intermediate is formed by annealing of duplicated sequences from the two strands, and the unpaired non-homologous 3’ ends are subsequently resected by exonuclease (Fig. 1.6). The
recombinant generated by this repair event contains only one copy of the repeated sequences, creating an identical deletion as by unequal crossover. Similar to other recombination events, the efficiency of SSA also strongly depends on the length of homology and the distance between repeats (Sugawara and Haber 1992; Sugawara et al. 2000). Interestingly, SSA can function efficiently in yeast even with repeats separated by 15 kb (Paques and Haber 1999).

Fig. 1.6. The SSA model. A DSB occurs between (or within) two direct repeats (red empty blocks) and is subsequently followed by 5' to 3' exonucleolytic resection. The two complementary single strands anneal to each other over the repeat sequences. Non-homologous 3' ends are resected and a deletion molecule is formed after repair synthesis and ligation.

1.3.2.2 Replication slippage

Replication slippage, occurring between runs of repeated sequences such as in microsatellites, is another major mechanism causing copy number variation though at local scale. Expansions and deletions can be generated as a consequence of the misalignment of repeats on the template or nascent strand during DNA replication (Fig. 1.7). Both in vitro and in vivo studies have shown that structures like palindromic DNA, hairpin structures formed by inverted repeats, and secondary structures from GC-rich sequences are obstacles for proper DNA replication (Bedinger et al. 1989; Moore et al. 1999; Sinden et al. 1999). Once the process is stalled, DNA polymerase may dissociate from the complex. The 3' end of the newly synthesised strand may subsequently displace from the template and anneal to a complementary position beyond the barrier. DNA
polymerase may reload to the complex, continuing the synthesis and generating deleted products (Fig. 1.7B) (Viguera et al. 2001).

**Fig. 1.7.** The model of replication slippage. (A) Proper DNA replication in a replication fork. Leading strand is shown in red dashed line while DNA repeats are marked by arrows. (B) Deletion caused by the misalignment of two inverted repeats on a template strand. (C) Duplication caused by the misalignment of two inverted repeats on a nascent strand.

### 1.3.3 Evolution of gene families

Segmental duplications are common and dispersed across the human genome, and it is likely that unequal crossover between dispersed repeats could act as the primary mechanism for generating these rearrangements throughout evolutionary history. Similarly, unequal crossover has also been recognised as a main mechanism for creating gene families. Examples such as the globin gene families (Higgs et al. 1989; Thein 1993) and odorant receptor genes (Kratz et al. 2002) provide evidence that duplicated genes in a gene family can be generated by mechanisms including unequal crossover and gene conversion from a common ancestor. Sequence divergence of gene members in a gene cluster between species is often higher than within species, suggesting that genes within a gene family can evolve in concert rather than independently (Hood et al. 1975; Zimmer et al. 1980). In concert evolution, two duplicated genes, derived from one ancestral gene, diverge by fixation of point mutations in both genes. However, ectopic recombination including unequal crossover and gene conversion between these duplicated genes can lead
to information transfer, resulting in sequence homogenisation which can in principle create a chromosome bearing two identical duplicates different from the ancestral gene. In practice, sequence homogenisation is often incomplete, leading to a patchwork of homology blocks shared by the duplicated genes.

1.3.4 Copy number variation in the human genome

Despite the fact that 99.9% of DNA sequences between two humans are identical (Reich et al. 2002; Feuk et al. 2006), these slight differences are more than enough to give variation between individuals in phenotype and susceptibility to disease. Although the focus has long been on SNPs, accumulating evidence indicates that structural genomic variants including copy number polymorphisms (CNPs), inversions and translocations, contribute significantly to genome variation, controlling gene dosage and regulating gene expression (Sharp et al. 2005; Tuzun et al. 2005; Feuk et al. 2006). Both computational and experimental analyses of CNPs in humans have suggested that ~5% of the human genome is composed of duplicated segments (Cheung et al. 2001; Bailey et al. 2002). In silico analyses have shown that ~10% of the entire genome contains potential rearrangement hot spots. Some of these sites have been shown to be associated with genomic disorders while some are polymorphic or are associated with disease-causing regions (Bailey et al. 2002). These hot spots are copy number polymorphic and most likely contain segmental duplications prone to rearrangement by ectopic recombination/NAHR (Sharp et al. 2005).

1.3.4.1 Array-based comparative genome hybridisation

Besides using karyotyping to detect structural variants in the human genome, other experimental approaches like array- and PCR-based strategies (Pinkel et al. 1998; Hollox et al. 2002; Schouten et al. 2002) have been used mostly in recent studies (Iafrate et al. 2004; Sebat et al. 2004). Array-based comparative genome hybridisation (array-CGH) is an array-based system used to compare copy number variants between DNA samples from different individuals (Pinkel et al. 1998; Ishkanian et al. 2004). In this approach, reference and test DNA samples are labelled with different fluorescent tags and hybridised to a genomic array spotted with cloned DNA fragments, for example, BACs. After hybridisation, copy number differences of a particular clone between reference and test DNA samples are simply analysed by comparing the signal ratios.
To improve resolution and the signal-to-noise ratio, another modified CGH was developed. Representational oligonucleotide microarray analysis (ROMA) is similar to traditional array-CGH approaches but screens only a proportion of the genome (Lucito et al. 2003; Sebat et al. 2004). The genome is firstly digested by restriction enzymes; PCR amplification is subsequently used to selectively amplify small adaptor-ligated fragments (<1.2 kb). These amplicons are then hybridised with an array of computationally designed oligonucleotide probes for signal comparison.

1.3.4.2 PCR-based approaches for targeted screening

PCR-based approaches have been designed to analyse CNPs at targeted regions of the genome instead of the whole genome. Systems like multiplex amplifiable probe hybridisation (MAPH) (Armour et al. 2000; Hollox et al. 2002) and multiplex ligation-dependent probe amplification (MLPA) (Schouten et al. 2002) share the similar principle that copy number variation is detected by analysing the yield differences of PCR products of the target probes by gel electrophoresis.

1.3.4.3 Confirmation using fluorescence in situ hybridisation

Although hybridisation- and PCR-based methods can give exact locations and boundaries of variants in a genome, results are usually validated by the traditional cytogenetic analysis, fluorescence in situ hybridisation (FISH) (Ishkanian et al. 2004; Sebat et al. 2004; McCarroll et al. 2006). FISH acts as a secondary confirmation and provides information about copy number variation and also chromosome position, but only for large rearrangements.

1.3.5 Classic examples of ectopic rearrangements in the human genome

Recent evidence has revealed that a significant proportion of genomic rearrangements is caused by aberrant recombination between segmental duplications such as low copy repeats (LCRs) generating recombinants including duplications, deletions and translocations (Emanuel and Shaikh 2001; Inoue and Lupski 2002). These rearrangements, whether arising from distantly separated or local repeats, can alter copy number in the human genome and also disrupt gene functions leading to genomic disorders. Common human genetic diseases showing CNP include DiGeorge plus velocardiofacial syndrome (DGS/VCFS), Smith-Magenis syndrome (SMS), Charcot-Marie-Tooth disease type 1A plus hereditary neuropathy with pressure palsies (CMT1A/HNPP) and α+-thalassaemia.
(Higgs 1993; Stankiewicz and Lupski 2002; Lupski and Stankiewicz 2005). They all share a similar mechanism for recurrent rearrangement: ectopic recombination between highly homologous LCRs. The following sections describe three classic examples of ectopic recombination occurring between megabase- or kilobase-separated repeats or repeat arrays.

### 1.3.5.1 CMT1A and HNPP

Disorders CMT1A and HNPP are the best-characterised examples of ectopic recombination caused by the misalignment of megabase-separated repeats. Unequal crossover between two 1.4 Mb-separated CMT1A-REP repeats, which share 98.7% nucleotide sequence identity (Reiter et al. 1997), can give rise to neuropathies CMT1A (duplications) and HNPP (deletions) by altering the \textit{PMP22} (peripheral myelin protein 22) gene dosage on chromosome 17p11.2 (Lupski and Stankiewicz 2005). In addition, common breakpoints mapped to CMT1A-REPs in unrelated CMT1A and HNPP patients indicate that they are likely to be reciprocal products of ectopic recombination (Pentao et al. 1992; Chance et al. 1994). Exchanges between repeats in CMT1A/HNPP occur non-randomly along the 24-kb CMT1A-REPs, and cluster into a 0.7 kb hot spot showing >0.5 kb of perfect sequence identity (Lopes et al. 1998; Reiter et al. 1998). This agrees with the argument that MEPS is essential for efficient ectopic recombination as seen in various organisms (Shen and Huang 1986; Jinks-Robertson et al. 1993). Although a \textit{mariner} transposon-like element is located near the hot spot, it is still unclear whether the inflated frequency of ectopic recombination at the hot spot is due to the presence of this putative transposon (Reiter et al. 1996).

Population studies and direct sperm analyses both indicate that the germ-line frequency of \textit{de novo} CMT1A duplication is fairly high, at \( \sim 1.5 \times 10^{-5} \) per sperm (Wise et al. 1993; Han et al. 2000). However, the origins of these recombinants might be different between males and females, since the majority of paternal CMT1A duplications is produced by unequal crossover between homologous chromosomes, while both maternal CMT1A and HNPP rearrangements are generated by intrachromosomal ectopic exchanges (Lopes et al. 1998). These differences indicate that there are distinct sex-dependent mechanisms, although no molecular information exists about the control of these mechanisms.
1.3.5.2 β-thalassaemia caused by ectopic deletions in the β-globin gene cluster

With high levels of sequence homology between genes in a gene cluster, globin gene families are good candidates for the study of unequal crossover. Although most forms of β-thalassaemia are caused by point mutation, unequal crossover occurring between kilobase-separated genes (ε, Gγ, Aγ, ψβ, δ and β) in the β-globin gene cluster can also lead to β-thalassaemia. For example, the Hb Lepore deletion is the product of ectopic exchange between homology blocks shared by the δ and β genes, whereas haemoglobin Kenya is formed by the fusion of Aγ and β genes (Metzenberg et al. 1991). Studies of ectopic exchange between δ and β genes, which generates Hb Lepore deletions and anti-Lepore duplications, have shown that Hb Lepore can be classified into different forms such as Hb Lepore-Boston, Hb Lepore-Baltimore and Hb Lepore-Hollandia based on their amino acid sequences and inferred exchange regions (Hardison et al. 2002). The occurrence of these exchanges seems positively correlated to the length of identical sequences in homology blocks, again supporting the involvement of MEPS (Metzenberg et al. 1991). Likewise, the fact that de novo ectopic deletions between Gγ and Aγ genes in sperm DNA (V. Lawson and A. Jeffreys, unpublished data) are far more frequent than those occurring between δ and β genes (see below) supports this relationship, since Gγ and Aγ genes share a longer homology region than δ and β genes do.

Direct analyses of de novo ectopic exchange between δ and β genes in sperm DNA showed that the ectopic recombination rate between these two regions is extremely low (6.2 × 10⁻⁸ per sperm) (Holloway et al. 2006). This rarity contrasts with the high frequency of allelic recombination within the gene cluster, which is triggered by the β-globin hot spot located at the promoter region of the β-globin gene. In addition, all Hb Lepore exchanges mapped to the 3′ side of the longest uninterrupted identical sequence between δ and β genes, away from the β-globin hot spot, again opposing the argument that ectopic recombination is controlled by the local allelic recombination hot spot.

1.3.5.3 Unequal crossover in minisatellite MS32

Highly unstable minisatellites provide one of the best-studied examples of unequal crossover between human repeat arrays. Unlike simple tandem repeat loci (Warren 1996), minisatellite instability is not hindered by sequence divergence between repeats (Buard et al. 2000). Human minisatellite MS32 is a typical highly unstable locus with a mean mutation rate of 0.8% per gamete (Jeffreys et al. 1994). Small pool PCR (SP-PCR)
analysis of sperm DNA showed that most recombinational events are germ-line-specific complex conversions, creating variable repeat arrays but without exchange of DNA markers flanking the array (Jeffreys et al. 1994). Additional single-DNA-molecule strategies (Jeffreys and Neumann 1997) have shown that unequal crossovers with exchange of flanking markers also existed in the germ-line but at a much lower frequency of about $5 \times 10^{-5}$ per sperm (Jeffreys et al. 1998b). Most of these unequal exchanges are simple, with only a minority showing patchy gene conversion. In addition, equal crossovers, yielding an isometric recombinant array equal in length to one or other progenitor allele, were also detected at a frequency (~$1.9 \times 10^{-5}$ per sperm) similar to unequal crossovers. Both crossover and conversion show polarity towards the beginning of the repeat array, indicating that these processes arise by a common mechanism possibly triggered by the minisatellite-associated hot spot (section 1.2.5.1) (Jeffreys et al. 1998a).

1.4 Human α-globin genes

DNA sequence similarities shared between members of a gene family provide targets for inducing DNA instability by mechanisms such as ectopic recombination. The duplicated α-globin genes (α2 and α1 genes) in the α-globin gene cluster are one of the best examples of unequal crossover, with the most common monogenic disease, α+-thalassaemia, often arising as a consequence of the interaction between the two duplicates. Despite extensive surveys of α-thalassaemia patients, detailed and systematic analyses of the mechanisms and dynamics of these mutations have never been reported. The following sections summarise the properties of the α-globin gene cluster, including its structure, function, and gene expression in normal individuals as well as rearrangements seen in α-thalassaemias.

1.4.1 Organisation of the human α-globin gene cluster

The human α-globin gene cluster locates near the tip of the short arm of chromosome 16 (16p13.3). The whole cluster carries an embryonic gene ($\zeta$), two foetal-adult genes (α2 and α1), three pseudogenes (ψζ, ψα2 and ψα1) and one θ gene with an unknown function, arranged in the order 5'-$\zeta$-ψζ-ψα2-ψα1-α2-α1-θ1-3' (Fig. 1.8) (Lauer et al. 1980; Higgs et al. 1989). The α-globin gene cluster is in a GC-rich isochore and the
overall GC-content is ~60%, which is much higher than the average for the human genome (~40%). The cluster is also associated with non-methylated CpG islands (Bird et al. 1987), many Alu-family repeats (Nicholls et al. 1987) and some hypervariable regions (Higgs et al. 1981).

![Diagram of the α-globin gene cluster and GC-content](image)

Fig. 1.8. Location of the α-globin gene cluster and GC-content of the ψα1, α2 and α1 genes. (A) Location and organisation of the α-globin gene cluster. (B) GC-content across the gene cluster. Structures of the ψα1, α2 and α1 genes, with exons and introns indicated by filled and empty boxes, respectively. Homologous regions are highlighted with coloured blocks. The percentage GC-content across the region is shown below.

### 1.4.2 Evolution of genes in the α-globin gene cluster

Genetic processes like deletion, duplication, gene conversion and base substitution are the major mechanisms responsible for the evolution of a gene family. Calculations of the percent corrected divergence of amino acid replacement sites (Perler et al. 1980) between the embryonic ζ- and the adult α-globin genes indicate that the divergence of ζ/α began approximately 380–400 million years ago (Proudfoot et al. 1982). Likewise, calculations of the divergence of silent substitution sites in ψα1 and α2 genes suggest that they diverged approximately 60 million years ago (Proudfoot and Maniatis 1980).

ψζ and ζ genes show extreme sequence identity, with only six base changes within a 2 kb region if excluding the repeat sequence variants. The significant sequence homology suggests that ψζ was originally a functional gene, subsequently inactivated by the appearance of a termination mutation in codon 6.
Nucleotide sequence similarity and homology block arrangements among the three α-globin genes, ψα1, α2 and α1 proved that they evolved from a common ancestral α-globin gene. Comparison of the foetal-adult α-globin genes (α2 and α1 genes) between human and most vertebrates suggested that the duplicated state has existed for at least 300 million years (Zimmer et al. 1980). Concerted evolution (Zimmer et al. 1980; Michelson and Orkin 1983) involving genetic processes like unequal crossover, gene conversion and point mutation, diverged the duplicates from other species, and these genes were finally fixed within a species (see section 1.3.3). In contrast, the ψα1 gene diverged from a functional α-globin gene under selection and then followed by base changes or deletions which inactivated its gene function (Proudfoot and Maniatis 1980). Time estimation of divergence of the ψα2 gene (the latest identified pseudogene within the cluster) by comparison of sequence homology reveals that it evolved even earlier than the ψα1 gene (Hardison et al. 1986).

The θ1 gene is likely to be functional (Shaw et al. 1987), though no corresponding protein has yet been identified. The θ-like gene has been found in many primate species, and comparisons of replacement- and silent-site substitutions between the α-like globin genes and the θ1 gene in different primates indicated that the θ1 gene has diverged from the α-globin genes about 260–280 million years ago (Hsu et al. 1988).

1.4.3 Gene expression in normal individuals
The three α-like globin genes, ζ, α2 and α1 are expressed at different stages during development. The ζ gene is expressed together with the γ and ε genes, located in the β-globin gene cluster on chromosome 11, to form embryonic haemoglobins Hb Portland (ζ2γ2) and Hb Gower-I (ζ2ε2), respectively. The expression of the ζ gene stops after the embryonic stage, with the α genes taking on the role of globin synthesis. The production of haemoglobin transits from embryonic Hb (Hb Gower-II, α2ε2), to foetal Hb (HbF, α2γ2) and finally to adult Hb (HbA, α2β2 and HbA2, α2δ2) (Higgs et al. 1989).

In normal adults, it is impossible to determine the difference of expression level between the two globin genes, α2 and α1 purely by analysing the identical α-globin proteins. By exploiting the sequence divergence of the 3' untranslated regions, the ratio of α2:α1 mRNA transcripts was found to be 3:1 (Orkin and Goff 1981). Likewise, studies of normal α-globins and α-globin structural mutants also suggest that α2 genes produce two- to three-fold more protein than the α1 gene (Liebhaber et al. 1986).
1.4.4 α-thalassaemia caused by α-globin gene deletions

The two duplicated α-globin genes (α2 and α1) are approximately 3.8-kb apart, and are embedded in highly homologous regions which are divided into several homology blocks (X, Y and Z) with different degree of sequence similarity (Fig. 1.9A). Unequal crossover between misaligned homology blocks can give rise to deletion of one α-globin gene in conjunction with a triplicated α chromosome or an excised circular α-globin gene by either an inter- or intramolecular exchange pathway, respectively (Fig. 1.9B) (Embury et al. 1980). In addition, other mechanisms such as NHEJ, which needs only a very short stretch of identical sequence or no homology at all, can give rise to deletion of one or both α-globin genes (Higgs et al. 1989). Deletions generated by these mechanisms can lead to α-thalassaemia with increasing severity positively correlated with the number of deleted α-globin genes.
Fig. 1.9. Unequal crossover between misaligned α-globin genes. (A) The structure of the α-globin gene cluster carrying ψα1, α2 and α1 genes, as described in Fig. 1.8B. Homologous sequences are grouped into three sets of homology blocks (X, Y and Z). Levels of sequence similarity between a coloured homology block and its paralogue immediately to the right are shown below. (B) Possible pathways for generating α-globin gene rearrangements. (i) An example of −α3.7 deletion arising from inter- or intramolecular ectopic exchange between misaligned Z homology blocks, accompanied with αcoxert α3.7 duplication or production of an extrachromosomal circle, respectively. (ii) Two ectopic pathways as described in (i) but caused by the misalignment of X homology blocks, giving −α4.2 deletion and reciprocal products.
1.4.4.1 \(\alpha^+\)-thalassaemia caused by unequal crossover

The two common classes of single \(\alpha\)-globin gene deletion found in \(\alpha^+\)-thalassaemics are \(-\alpha^{3.7}\) and \(-\alpha^{4.2}\) deletions (Embry et al. 1980), accompanied by a deletion of \(\sim 3.8\) kb and \(\sim 4.2\) kb of DNA, respectively. They are usually produced by unequal crossover between misaligned homology blocks (Fig. 1.9B). The \(-\alpha^{3.7}\) deletion can be classified into three subtypes, \(-\alpha^{3.7}_{II}, -\alpha^{3.7}_{III}\) and \(-\alpha^{3.7}_{IV}\) corresponding to different exchange regions along the \(Z2/Z1\) homology block (Fig. 1.10) (Higgs et al. 1984). These \(-\alpha^{3.7}\) deletions are prevalent in most areas especially in malarial regions whereas the leftward \(-\alpha^{4.2}\) deletions caused by exchange between misaligned \(X\) homology blocks are largely restricted to some populations, for instance Southeast Asia and the Middle East (see section 1.4.6 for details) (Bowden et al. 1987).

![Subtypes of \(-\alpha^{3.7}\) deletion. Structures of the \(\alpha\)-globin gene cluster and the \(-\alpha^{3.7}\) deletion are the same as described in Fig. 1.9. Segments I, II and III corresponding to the exchange breakpoints of \(-\alpha^{3.7}_{II}, -\alpha^{3.7}_{III}\) and \(-\alpha^{3.7}_{IV}\) deletions, respectively, are marked below a magnified \(Z2/Z1\) homology block in which paralogous sequence variants (PSVs) are highlighted in red.](image)

1.4.4.2 Illegitimate recombination causing \(\alpha^0\)-thalassaemia

\(\alpha^0\)-thalassaemia, which is caused by deletion of both \(\alpha\)-globin genes on a chromosome (---), is a severe form of thalassaemia. There are different types of \(\alpha^0\)-thalassaemia, for example, \(-\) \(\text{BRT}, -\text{FIL}, -\text{MED}, -\text{SEA}\) and \(-\) \(\text{THAI}\) (Winichagoon et al. 1984; Higgs et al. 1985; Hill 1992). Most of these deletions remove not only both of the functional \(\alpha\)-globin genes but also the pseudogenes, the \(0\) gene or the \(\zeta\) genes (Higgs et al. 1989). The locations of most \(\alpha^0\) deletion breakpoints are different from those of \(\alpha^+\) deletions, with
breakpoints showing only a short stretch of identical sequence or no sequence homology. Since common sequences and sequence motifs including direct repeats, palindromic sequences, interspersed repetitive sequences, Z-DNA and chi-sequences are usually found around the breakpoints of illegitimate recombination (Smith et al. 1981; Bullock et al. 1985; Kmiec and Holloman 1986), it has been suggested that some α^0-thalassaemia deletions, which carry these sequences around the breakpoints, most likely resulted from illegitimate recombination.

Furthermore, there are some rare deletions which are most likely caused by illegitimate recombination but with only one globin gene deleted, for example, −α^7.9 (Harteveld et al. 2003), (α)α^5.3 (Lacerra et al. 1991) and −α^3.5 deletions (Kulozik et al. 1988). The −α^7.9 deletion, a new deletion type with the entire α2 gene and the ψα1 gene deleted, has been identified in two independent Indian families (Harteveld et al. 2003). The deletion breakpoints imply that this deletion is neither caused by homologous unequal exchange nor recombination between misaligned Alu repeats. Similarly, sequence analysis of breakpoints hinted that the (α)α^5.3 deletion, with the whole ψα1 gene and the 5’ portion of the α2 gene being removed, arose by illegitimate recombination instead of unequal crossover (Lacerra et al. 1991).

1.4.5 Phenotypical differences between normal individuals and α-thalassaemias
Different copy numbers of α-globin genes can result in various levels of clinical and haematological severity with a positive correlation with the reduction of α-globin synthesis. Although reduced levels of haemoglobin, mean corpuscular volume (MCV) and mean corpuscular haemoglobin (MCH) have been observed in mild α-thalassaemias, the reduction is only subtle. Red blood cell indices indicate that −α/αα heterozygotes are phenotypically indistinguishable from normal individuals. The MCV and MCH levels in other α-thalassaemia carriers including −/αα, −α/−α and α-thalassaemia trait (α^T) carriers are reduced compared with those in normal individuals, despite the fact that most of these carriers show only a mild and sometimes asymptomatic form of α^T-thalassaemia (Higgs 1993; Wang 2000). In contrast, severe forms of α-thalassaemia like homozygotes with no α-globin genes (−/−) and individuals with only one α-globin gene (−/−α) can lead to the severe Hb Bart’s hydrops foetalis syndrome and stillbirth, or to HbH disease, respectively (Flint et al. 1986a; Higgs 1993).
1.4.6 Malaria selection and distribution of α-thalassaemia

α+ -thalassaemia, which is caused by the deletion of a single α-globin gene, is commonly found throughout the world. The leftward (−α^4.2) and rightward (−α^3.7) deletions are the most common types. They are prevalent in malarial regions and are most likely maintained by malaria selection (Flint et al. 1986a; Hill 1992). Many speculations in explaining how the growth of the malaria parasites, *Plasmodium falciparum*, is retarded in thalassaemia erythrocytes have been proposed (Nagel and Roth 1989). However, no *in vitro* studies have ever provided conclusive proof for these speculations. Although there is no direct experimental evidence showing how −α chromosomes are selected by malaria, population studies provide strong indirect evidence for this selection. For instance, the frequency of −α alleles and malaria endemicity are significantly correlated, showing a population incidence of α-thalassaemia that decreases in parallel with reduced malaria endemicity in response to inclines of both altitude and latitude in Melanesia (Flint et al. 1986a). Frequencies of −α alleles can be remarkably high in malarial areas, for example, 0.24–0.69 amongst malarious Papua New Guinea coastal populations (Yenchitsomanus et al. 1986) and 0.50 in two groups of children in Kenya (Williams et al. 2005). In contrast, gene frequencies of −α alleles decline in areas with decreasing malaria endemicity (Hill 1992) and remain very low in malaria-free regions, at <0.01 in Iceland and Britain (Flint et al. 1986b). Although the −α chromosome seems to be favoured by malaria selection, whether selection acts on homozygotes (−α/−α) and/or heterozygotes (−α/αα) is still unknown.

Deletions with different breakpoint locations are frequently found in some populations, particularly with −α^3.7 deletions in which three subtypes, −α^3.7_i, −α^3.7_ii and −α^3.7_iii can be distinguished (Fig. 1.10) (Higgs et al. 1984). Furthermore, −α chromosomes with different haplotypic backgrounds can exist in α-thalassaemias; population studies revealed that both −α^3.7 and −α^4.2 deletions distribute differently in regions. −α^4.2 deletions are only prevalent in the north of Papua New Guinea in Melanesia. They are associated with at least three different haplotypes and none of them share the same origin with those surveyed in Southeast Asia. Similarly, −α^3.7_i deletions from the south part of Melanesia were found to have more than one haplotypic origin (Flint et al. 1986a). Interestingly, apart from Melanesia, most of the −α^4.2 deletions distribute mainly in Southeast Asia. −α^3.7_i and −α^3.7_ii deletions are commonly found in most populations, whereas −α^3.7_iii deletions are restricted to Oceania (Flint et al. 1998). Despite the fact that
the reasons for this regional distribution are unclear, \( -\alpha \) deletions must be recurrent in populations.

1.4.7 Allelic recombination in the \( \alpha \)-globin gene cluster

Information about homologous (allelic) recombination in the \( \alpha \)-globin gene cluster is very limited and no direct high-resolution analyses of allelic recombination within this cluster in the human germ-line have yet been reported. Despite the fact that some studies, particularly in males (Saadallah and Hulten 1983; Daniels et al. 2001), predicted that recombination rates in the subtelomeric region of chromosome 16p are high, historical recombination rates within the \( \alpha \)-globin gene cluster seem to be very low. Haplotype studies on normal individuals from nine populations revealed that the \( \alpha \)-globin gene cluster is very polymorphic. However, there is no evidence for distinct differences of recombination frequency within the locus, since the limited combinations of polymorphic markers indicate that the cluster is embedded in a region of strong linkage disequilibrium (Higgs et al. 1986). Furthermore, information from the International HapMap Project (Gibbs et al. 2003) also suggests that the \( \alpha \)-globin gene cluster lies within a strong LD block, although detailed analysis of the region is impeded by limited number of heterozygous SNPs and a huge number of unverified markers, which are most likely paralogous sequence variants (PSVs) between homologous regions that have been incorrectly scored as SNPs. As a whole, although no concrete conclusions about the recombination rate of the \( \alpha \)-globin gene region can be drawn, high allelic recombination rates are not expected.

1.5 Overview of the thesis

Although DNA instability caused by processes including allelic and ectopic recombination can be revealed in global studies of the human genome, details of mechanisms and dynamics can only be analysed by high-resolution studies of specific DNA regions. In this thesis, the human \( \alpha \)-globin gene region was chosen as a representative for the study of ectopic recombination. There are three main reasons for choosing this cluster. Firstly, \( \alpha^+ \)-thalassaemia, which is generally caused by a single \( \alpha \)-globin gene deletion arising from unequal crossover between the duplicated \( \alpha \)-globin genes, is one of the most common genetic disorders. Despite abundant descriptions of the disease-associated deletions, the processes and dynamics of unequal crossover are
unknown. Secondly, being a classic example of gene families, the α-globin gene cluster is a typical prototype for studying CNPs caused by recombination between local repeats, arguably being more relevant in evolution than large genomic rearrangements. Finally, the human α-globin gene cluster is also an excellent candidate for studying the currently unknown relationship between allelic and ectopic recombination in humans.

The studies are reported in six chapters in this thesis. Chapter 3 describes the optimisation of successful PCR amplification of a 12-kb α-globin gene region. This technology was used to detect de novo −α deletions in somatic and germ-line DNA using single-DNA-molecule techniques as described in Chapter 4. The dynamics and processes of ectopic recombination between α-globin genes are discussed in the same chapter. Chapter 5 focuses on the contrast between de novo −α deletion frequencies and the population incidence of −α chromosomes, revealing the strength of natural selection in controlling gene copy number. The detection and characterisation of de novo ααα duplications using a novel inverse PCR strategy are reported in Chapter 6. Despite the substantial level of reciprocity between de novo −α deletions and ααα duplications, the detection of putative extrachromosomal DNAs, described in Chapter 7, may point to the existence of an intramolecular ectopic recombination pathway. Lastly, preliminary strategies for detecting allelic recombinants in this gene region are outlined in Chapter 8.

Some of the results obtained from this work have been published in the following two papers:


Chapter 2
Materials and Methods

2.1 Materials

2.1.1 Chemical reagents
All chemical reagents were supplied by ABgene (Surrey, UK), Amersham Biosciences (Little Chalfont, UK), Applied Biosystems (Warrington, UK), Cambrex Bio Science Rockland (Rockland, USA), Clontech (Mountain View, USA), Edge Biosystems (Gateshead, UK), Fisher Scientific (Loughborough, UK), Fisons (Beverley, USA), FMC Bioproducts (Rockland, USA), Geneflow Ltd. (Fradley, UK), Institute of Biochemistry, University of Muenster (Muenster, Germany), Invitrogen (Paisley, UK), New England BioLabs (Hitchin, UK), Perkin Elmer (Cambridge, UK), Promega (Southampton, UK), Qiagen Ltd. (Crawley, UK), Roche (Welwyn Garden City, UK), Serva (Cambridge, UK), Sigma-Aldrich Company Ltd. (Gillingham, UK), Stratagene (Amsterdam, The Netherlands) and USB Co. (Cleveland, USA).

2.1.2 Consumables and equipment
Consumables and equipment were obtained from Amersham Biosciences (Little Chalfont, UK), Bando Chemical Ltd. (Kobe, Japan), Barloworld Scientific (Stone, UK), BD Biosciences (Oxford, UK), Bio-Rad (Hemel Hempstead, UK), Clare Chemical Research (Dolores, USA), Cecil Instruments (Cambridge, UK), Corning Ltd. (Hemel Hempstead, UK), Duran (Mainz, Germany), Eppendorf Scientific (Cambridge, UK), Flowgen (Nottingham, UK), Fujifilm (London, UK), GE Water & Process Technologies (Heverlee, Belgium), Genetic Research Instrumentation (GRI) (Braintree, UK), Hybaid (Teddington, UK), Millipore (Watford, UK), MJ Research (Cambridge, USA), Nalge Nunc International (Hereford, UK), New Brunswick Scientific Co. (New Jersey, USA), Precisa (Milton Keynes, UK), Sanyo (Watford, UK), Sarstedt (Leicester, UK), Sartorius Ltd. (Epsom, UK), Scilabub Ltd. (Measham, UK), Starlab (Ahrensburg, Germany), Stuart Scientific (Stone, UK), Syngene (Cambridge, UK), Thermo Electron Oy (Vantaa, Finland), Thermo Hybaid (Franklin, USA) and UVP Life Sciences (Cambridge, UK).
2.1.3 Standard solutions
Southern blot solutions (depurinating solution, denaturing solution and neutralising solution), 20 × Sodium chloride-Sodium citrate (SSC) buffer, 10 × Tris-borate-EDTA (TBE) buffer, Luria-Bertani broth (LB) and Luria-Bertani agar were prepared as described by Sambrook and Russell (2001), and supplied by the media kitchen, Department of Genetics, University of Leicester, UK. The standard 11.1 × PCR buffer was prepared and provided by R. Neumann, Department of Genetics, University of Leicester, UK.

2.1.4 Sources of human DNA sample
Genomic DNA was prepared from semen, blood and testis samples. All samples were collected with approval from the Leicestershire Health Authority Research Ethics Committee and with informed consent. Semen samples from anonymous men were supplied by J. Blower (Leicester Royal Infirmary, Leicester, UK). Additional semen and blood samples were provided by volunteers from the Department of Genetics (University of Leicester, Leicester, UK). Testis samples were obtained from testicular cancer and gender reassignment patients through the good offices of Prof. K. Mellon (Leicester General Hospital). Genomic DNA from lymphoblastoid cell lines from individuals in different world populations was supplied by the Centre d’Etude du Polymorphisme Humain (CEPH) (Paris, France). Genomic DNA from buccal cell samples obtained from unrelated men in the United Kingdom identified by grandparental place of birth as northern European in origin was supplied by M. Jobling and T. King, Department of Genetics, University of Leicester, UK.

2.1.5 Oligonucleotides
All oligonucleotides for PCR amplification and ASO hybridisation were supplied by Invitrogen and Sigma-Aldrich Company Ltd., and arrived in a lyophilised form.

2.1.6 Enzymes
Restriction enzymes, T4 polynucleotide kinase and exonuclease I were supplied by New England Biolabs. The Klenow fragment of DNA polymerase I of E. coli was supplied by USB Co. Taq polymerase was obtained from ABgene. Cloned Pfu polymerase was supplied by Stratagene. Proteinase K was provided by Sigma-Aldrich, and shrimp alkaline phosphatase was supplied by Roche.
2.1.7 DNA ladders

λ DNA digested with HindIII and φX174 DNA digested with HaeIII were both obtained from ABgene. Supercoiled DNA Ladder was supplied by Invitrogen.

2.1.8 Computers

The thesis was produced using Windows XP on a PC, an Epson Perfection 1250 or an HP Photosmart C3180 scanner, and was printed on an HP LaserJet 4250tn or an HP Deskjet 880c printer. Data and images were stored, analysed and presented using the software packages Adobe Acrobat, Adobe Photoshop, AutoAssembler, EndNote, Epson Smart Panel, Factura, Freehand, Microsoft Excel, Microsoft Powerpoint and Microsoft Word on either a Macintosh computer or a PC. The Poisson analysis programme was written in True BASIC 4.1 by A. Jeffreys, Department of Genetics, University of Leicester, UK, and the analyses were carried out on a Macintosh computer.

2.2 Methods

2.2.1 DNA extraction

DNA extraction was carried out in a category II laminar flow hood and manipulated as described under conditions designed to minimise the risk of contamination (Jeffreys et al. 1994). The following sections are the general extraction procedures for different tissues.

2.2.1.1 DNA extraction from human sperm

A fresh ejaculate was diluted with 5 ml of 1 × SSC and stored at −80°C. Frozen semen samples were thawed and 1 ml aliquotted into a 1.5-ml screw top tube. Sperm and other cells were pelleted by centrifugation at full speed (at 13200 rpm in an Eppendorf 5415 D centrifuge) for 2 minutes. Supernatant was removed and the pellet was resuspended thoroughly in 1 ml of 1 × SSC. Other cells except sperm heads were lysed by addition of 20 μl 10% (w/v) SDS, then flicking and centrifuged for 2 minutes. The supernatant was discarded; further resuspension and lysis were repeated. The pellet was then resuspended in 1 ml of 1 × SSC and centrifuged for another 2 minutes. In the absence of SDS, sperm heads stick on the wall of the tube instead of pelleting. The supernatant was removed and sperm heads were resuspended in 450 μl 0.2 × SSC. Sperm heads were lysed by the addition of 50 μl 10% (w/v) SDS, 35 μl 2-mercaptoethanol (final concentration 1 M) and
5 μl 20 mg/ml proteinase K (final concentration 200 μg/ml) and incubated at 37°C for 30 minutes with occasional mixing. Proteins were removed by the addition of 350 μl phenol (Fisher Scientific) with gentle mixing to allow emulsification and centrifuged at full speed for 2 minutes. The upper (aqueous) layer of the supernatant was re-extracted by mixing with fresh phenol. The lower (organic) layer was re-extracted once with 1 × SSC and 0.2% (w/v) SDS. Aqueous supernatants containing DNA were pooled, ethanol precipitated using 2.5 volumes of 100% ethanol and swirled gently. The supernatant was removed and the DNA precipitate was washed with 80% ethanol. The DNA pellet was dissolved in 1.8 ml distilled water, and re-precipitated by the addition of 200 μl 2 M sodium acetate (pH 7.0) and 2.5 volumes 100% ethanol and centrifuged for 1 minute. The DNA pellet was further washed with 80% (v/v) ethanol, vacuum dried and redissolved in 5 mM Tris-HCl (pH 7.5).

2.2.1.2 DNA extraction from human blood
Venous blood samples (delivered into equal volumes of 1 × SSC and stored at −80°C), were thawed at 37°C and 1.4-ml-aliquotted into a 1.5-ml screw top tube. The aliquot was centrifuged at full speed for 2 minutes. The supernatant was removed and the pellet was washed twice in 1 ml of 1 × SSC. The pellet was resuspended thoroughly in 300 μl 0.2 × SSC. Cells were lysed by addition of 30 μl 10% (w/v) SDS and proteinase K (final concentration of 200 μg/ml) and incubated at 37°C for 30 minutes with occasional mixing. Trace proteins were removed by addition of phenol with gentle mixing to allow emulsification and centrifuged at full speed for 2 minutes. The re-extraction process was done as for sperm DNA. The DNA pellet was ethanol precipitated, vacuum dried and redissolved in 5 mM Tris-HCl (pH 7.5).

2.2.1.3 DNA extraction from human testis
A testis stored at −80°C was thawed at 37°C. The spermatic cord, tunica albuginea/vaginalis and other peripheral tissues were removed. Seminiferous tubules (~0.6 g) were homogenised in 3 ml of 1 × SSC. An aliquot of homogenate (125–250 μl) was mixed with 1 × SSC in a 1.5-ml screw top tube in a total volume of 925 μl. Cells were centrifuged at 13200 rpm for 2 minutes. The supernatant was discarded and the cell pellet was resuspended thoroughly in 450 μl 0.2 × SSC. The cell suspension was lysed with 50 μl 10% (w/v) SDS, 35 μl 2-mercaptoethanol and 4 μl 20 mg/ml proteinase K and
incubated at 37°C for 30 minutes with occasional mixing. Following incubation, phenol extraction and DNA precipitation were performed as previously described.

2.2.1.4 Plasmid DNA extraction
An isolated colony containing a 3.8-kb plasmid [a 0.8-kb human DNA fragment ligated in a 3.0-kb pGEM-T Easy vector (Invitrogen)] was provided by P. Freeman, Department of Genetics, University of Leicester, UK. The colony was cultured in 5 ml of LB with 200 μg/ml ampicillin at 37°C overnight with shaking at 250 rpm. Plasmid DNA was extracted using a QIAprep Spin Minipred Kit (QIAGEN) according to the manufacturer’s instructions.

2.2.2 Measuring DNA concentration
The DNA concentration was estimated by measuring optical density at wavelength of 260 nm (OD260) at three different dilutions (for example, 1 in 100, 1 in 200 and 1 in 400) using either a Cecil Instruments CE 202 Ultraviolet Spectrophotometer or an Eppendorf BioPhotometer. The DNA concentration was calculated by multiplying the averaged OD260 reading with the dilution factor and 50 (OD of 1 corresponding to approximately 50 μg/ml double-stranded DNA).

2.2.3 PCR amplification
PCRs were set up in a category II laminar flow hood to minimise the risk of contamination. They were performed in 0.2-ml PCR tubes or 96-well plates in either a PTC-225 Tetrad DNA engine (MJ Research) or a PTC-240 Tetrad 2 thermal cycler (MJ Research). The following sections show only general settings for long PCR. Details of specific PCRs, primer sequences, DNA templates and specific PCR thermal cycling conditions are described in Tables 2.3–2.7, in appropriate chapters or in Appendix I.

2.2.3.1 Primer design
PCR primers were 18–20 bases long with 50–70% GC-content. They were designed with at least one G/C but no more than three G/Cs at the last three nucleotides. Allele-specific primers were designed with the target nucleotide at the 3’ end, 15–19 bases long and with at least 50% GC-content.
2.2.3.2 PCR buffer

PCRs were performed using 11.1 × PCR buffer (Jeffreys et al. 1990). Unless stated otherwise, the working PCR buffer concentration was adjusted to 1 ×, as described in Table 2.1.

Table 2.1. Constituents of 11.1 × PCR buffer

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Stock concentration</th>
<th>Volume added (μl)</th>
<th>Final concentration (1×)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris-HCl (pH 8.8)</td>
<td>2 M</td>
<td>167</td>
<td>45 mM</td>
</tr>
<tr>
<td>Ammonium sulphate</td>
<td>1 M</td>
<td>83</td>
<td>11 mM</td>
</tr>
<tr>
<td>Magnesium chloride</td>
<td>1 M</td>
<td>33.5</td>
<td>4.5 mM</td>
</tr>
<tr>
<td>2-mercaptoethanol</td>
<td>100%</td>
<td>3.6</td>
<td>6.7 mM</td>
</tr>
<tr>
<td>EDTA (pH 8.0)</td>
<td>10 mM</td>
<td>3.4</td>
<td>4.4 μM</td>
</tr>
<tr>
<td>dATP</td>
<td>100 mM</td>
<td>75</td>
<td>1 mM</td>
</tr>
<tr>
<td>dCTP</td>
<td>100 mM</td>
<td>75</td>
<td>1 mM</td>
</tr>
<tr>
<td>dGTP</td>
<td>100 mM</td>
<td>75</td>
<td>1 mM</td>
</tr>
<tr>
<td>dTTP</td>
<td>100 mM</td>
<td>75</td>
<td>1 mM</td>
</tr>
<tr>
<td>BSA</td>
<td>10 mg/ml</td>
<td>85</td>
<td>113 μg/ml</td>
</tr>
</tbody>
</table>

2.2.3.3 PCR cocktail

PCRs were generally in 10 μl reactions containing 0.9 μl of 11.1 × PCR buffer (1 × PCR buffer in final concentration), plus 0.2 μM PCR primers, 0.03 unit/μl Taq polymerase, 0.003 unit/μl cloned Pfu polymerase and 12.5 mM Tris base. Carrier salmon sperm DNA (1 μg/ml) was added to the reaction in which the target DNA input was below 1 ng, for instance, in single-molecule PCRs and nested PCRs with 200-fold diluted primary PCR product input, to preferentially coat the surface of plastic-ware and prevent target DNA being sequestered. Details of PCR cocktails are described separately in Tables 2.3–2.7 or in appropriate chapters.

2.2.4 Gel electrophoresis

2.2.4.1 Electrophoresis conditions

DNA samples were mixed with 0.2 volumes of loading dye (44 mM Tris-borate pH 8.3, 1 mM EDTA, 30% vol/vol glycerol and bromophenol blue) and loaded into a 0.8% LE agarose (SeaKem) gel in 0.5 × TBE buffer (44 mM Tris-borate pH 8.3, 1 mM EDTA) with 0.5 μg/ml ethidium bromide for gel electrophoresis. Samples were electrophoresed
alongside DNA ladders of known sizes. Electrophoresis tanks were manufactured in-house and power packs were supplied by Bio-Rad. DNA was visualised using a UV wand (Chromato-vue UVM-57, UVP Life Sciences) or the GeneGenius analysis system (Syngene), or the Dark Reader system (Clare Chemical Research) for band excision without DNA damage.

2.2.4.2 Gel photography
Ethidium bromide-stained gels were visualised using a UV transilluminator inside the GeneGenius analysis system (Syngene). Illuminated DNA images were captured and analysed as inverse images using GeneSnap software (Syngene). Photos were printed out on photographic paper using a Sony digital graphic printer UP-D895 (Syngene).

2.2.5 DNA purification

2.2.5.1 Exo/Sap purification
Homogeneous PCR products, which showed as a single and intense band on gel electrophoresis, were mixed with exonuclease I and shrimp alkaline phosphatase, as described in Table 2.2, and incubated at 37°C for 1 hour and 80°C for 15 minutes to degrade any PCR primer as well as dephosphorylate any dNTP left in the reactions. This purification method was used to clean up PCR products prior to automated sequencing.

Table 2.2. Exo/Sap purification of PCR product

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume added (µl)</th>
<th>Final concentration (units per µl of PCR product)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR product</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>Exonuclease I (20 units/µl)</td>
<td>0.5</td>
<td>2</td>
</tr>
<tr>
<td>Shrimp alkaline phosphatase (1 unit/µl)</td>
<td>1.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Dash indicates not applicable.

2.2.5.2 Electroelution from agarose gels
A target DNA fragment from an agarose gel was visualised by the Dark Reader system and excised with a scalpel. The blue light emitted by the Dark Reader system gives little damage to DNA molecules compared to the UV light. A piece of dialysis membrane was cut to a size slightly wider than the target DNA fragment slice. It was boiled in
10–50 mM EDTA for 10 minutes and placed in a slot cut within a second gel. The target DNA fragment slice was transferred to the gel slot with the dialysis membrane wrapped beneath. The target DNA was monitored using the Dark Reader system, and the gel was run allowing the DNA to electroelute onto the membrane. The membrane was rapidly removed from the gel slot and transferred into a 1.5-ml microcentrifuge tube with an edge of the membrane trapped in the lid. The DNA in droplets of buffer was collected from the dialysis membrane by centrifugation at full speed for 2 minutes and recovered by ethanol precipitation.

### 2.2.5.3 Ethanol precipitation

DNA was precipitated in 1.5-ml microcentrifuge tubes by addition of 0.1 volumes of 2 M sodium acetate pH 5.5 and 2.5 volumes of 100% ethanol. Depending on the size of DNA molecules, tubes containing DNA fragments smaller than 1 kb were chilled at -80°C for 10–30 minutes prior to centrifugation at 13200 rpm for 10 minutes. The supernatant was discarded; the DNA pellet was then washed with 80% ethanol, vacuum dried and redissolved in 5 mM Tris-HCl (pH 7.5).

### 2.2.5.4 Estimating DNA recovery

DNA recovery was either measured by spectrophotometry as described in section 2.2.2 or estimated by comparing band intensity on agarose gels after staining with ethidium bromide. The DNA sample was gel electrophoresed in parallel with serial dilutions of a known amount of DNA, for instance, \( \lambda \) DNA \( \times \) HindIII. The band intensity of PCR products was compared directly with that of fragments in the DNA ladder, and the yield of recovery was estimated.

### 2.2.6 DNA size fractionation

All equipment used in size fractionation, for example, gel tanks, gel combs and glass plates, were cleaned and decontaminated in a bath of bleach and diluted HCl for 1 hour to remove all possible traces of DNA molecules. Processing after electroelution was done in a category II laminar flow hood to minimise the risk of contamination. The following sections describe only general procedures of the strategy; detailed features and descriptions are given in appropriate chapters.
2.2.6.1 Restriction enzyme digestion

Genomic DNA (89-300 µg) was digested with restriction endonuclease to cleave out target DNA fragments. An aliquot of undigested, digested and over-digested (adding excess restriction enzymes to an aliquot of the digested sample) DNA was analysed by gel electrophoresis with staining by ethidium bromide to check for completeness of digestion. In addition, completeness was also checked by addition of 20 ng λ DNA to an aliquot of the digest to ensure that λ DNA was cut to completion under the digest conditions. Digested DNA was ethanol precipitated and redissolved in 400 µl 5 mM Tris-HCl (pH 7.5). The DNA recovery was measured using a spectrophotometer as described above.

2.2.6.2 Preparative gel electrophoresis

A 40-cm-long, 19.5-cm-wide glass plate, whose sides were wrapped with autoclave tape, was used as a gel-casting template. A 1.4-cm-deep 0.8% SeaKem HGT agarose gel (Cambrex Bio Science Rockland) with a 5- x 0.3-cm slot was set on the glass plate. The gel was soaked in 0.5 x TBE buffer with 0.5 µg/ml ethidium bromide. A small amount of loading dye was added to the slot to check if there was any leakage; the dye was then removed by a syringe. The digested DNA mixed with 1/5 volume loading dye plus ethidium bromide added to 400 µg/ml was slowly loaded into the slot. The DNA marker (3 µg λ DNA x HindIII) mixed with 10 µl of loading dye and 4 µl of 5 mg/ml ethidium bromide was also loaded into side slots flanking the main slot. The loaded samples were left for 15 minutes to allow the DNA to sink to the bottom of the slot prior to electrophoresis and to allow gel and sample buffers to equilibrate. The gel was run at 55 V overnight and the voltage was adjusted to 120 V for 2–3 days. The gel tank was covered with a black plastic bag to avoid any possible UV light damage to the DNA. The running buffer was changed every day to provide sufficient ethidium bromide for visualisation. The gel was visualised by a Dark Reader system and the migration distance of the DNA marker was monitored in order to estimate the separation between target DNA mutant and progenitor molecules.

2.2.6.3 Physical enrichment

The migration distance of target mutant and progenitor DNA molecules was estimated from the DNA marker ladder. The position of target mutants and progenitors was marked with loading dye. A total of 8–20 size fractions were collected, with slices closest to the
progenitor collected last to minimise any progenitor contamination. Gel slices were kept separately in petri dishes. The surface 1 mm of each gel slice was trimmed off to remove any progenitor DNA migrating aberrantly near the surface of the gel. Gel slices of each size fraction were divided into two and the DNA was electroeluted onto dialysis membrane in a 1% LE agarose gel as described in section 2.2.5.2. Eluates from the same fraction were pooled and the DNA was ethanol precipitated and redissolved in 50 µl 5 mM Tris-HCl (pH 7.5).

2.2.7 DNA transfer to membrane

2.2.7.1 Southern blot
Following electrophoresis, the target region of agarose gel required for DNA transfer was excised. The gel was soaked in depurination solution (0.25 M HCl) for 5 minutes. The gel was then soaked in denaturation solution (0.5 M NaOH, 1.5 M NaCl) for 8 and 20 minutes to separate double-stranded DNA into single-stranded DNA and to hydrolyse apurinic sites. The gel was finally washed with neutralisation solution (3 M NaCl, 0.5 M Tris-HCl pH 7.4) for 8 and 20 minutes. DNA was transferred to MAGNA nylon membrane (GE Water & Process Technologies) (pre-soaked in 2 x SSC) by capillary action using 20 x SSC as the transfer buffer (Southern 1975). Towels were changed twice in the first 15 minutes to ensure even transfer; the blotting was then left for overnight. The membrane was removed from the blotting apparatus, rinsed in 2 x SSC and dried at 80°C for 10 minutes. The DNA was covalently linked to the membrane by exposure to high intensity UV light at wavelength 302 nm for 2 minutes on a transilluminator.

2.2.7.2 Dot blot
PCR products (10 µl) containing 3–100 ng DNA per kb were mixed with 2 µl loading dye. Samples were mixed with 100 µl denaturation mix (0.5 M NaOH, 2 M NaCl, 25 mM EDTA) to render DNA single-stranded. A sheet of MAGNA nylon membrane (GE Water & Process Technologies) with two additional backing sheets of 3 MM Whatman chromatography paper (all pre-soaked in 2 x SSC) were put in the assembled 96-well dot blot manifold. A vacuum was applied and 50 µl denatured PCR product was loaded in each well. The well was washed with 125 µl 2 x SSC to neutralise the DNA. The membrane was removed from the manifold, dried at 80°C for 10 minutes and exposed to UV light for DNA cross-linking as described above.
2.2.8 Hybridisation

2.2.8.1 Southern blot hybridisation
Double-stranded DNA (8 ng) (target probe and/or DNA markers) was dissolved in 30 μl distilled water and denatured at 100°C for 4 minutes. The denatured probe was mixed with 6 μl 5 × oligo labelling buffer [Solution A: 1.2 M Tris-HCl (pH 8.0), 121 mM MgCl₂, 1.8% v/v 2-mercaptoethanol, 0.5 mM dATP, 0.5 mM dGTP, 0.5 mM dTTP; Solution B: 2 M HEPES (pH 6.0); Solution C: d(N)₆ hexadeoxynucleotides dissolved in 3 mM Tris-HCl, 0.2 mM EDTA (pH 7.0) at 90 OD units/ml (Solutions A, B and C were mixed in the ratio 2:5:3 respectively)], 0.8 μl 10 mg/ml BSA, 0.8 μl 5 units/μl Klenow fragment of *E. coli* DNA polymerase I and 15 μCi α-³²P-dCTP (Amersham Biosciences) and incubated at 37°C for 1 hour or at room temperature for overnight. The labelling was terminated by addition of 150 μl oligo stop solution (20 mM NaCl, 20 mM Tris-HCl pH 7.5, 2 mM EDTA, 0.25% SDS). The labelled probe was recovered by ethanol precipitation at room temperature using 90 μg of high molecular weight salmon sperm DNA as a carrier. The probe was dissolved in 400 μl distilled water and denatured at 100°C for 4 minutes before use. The membrane was pre-hybridised at 65°C for 20 minutes in modified Church buffer (Church and Gilbert 1984) (0.5 M sodium phosphate pH 7.2, 7% SDS, 1 mM EDTA pH 8.0). Depending on the amount of DNA on the membrane, hybridisation was carried out at 65°C for 30 minutes up to overnight. The membrane was washed with high stringency wash solution (0.2 × SSC, 0.1% SDS) at 65°C for 15 minutes, blot-dried and exposed to Fuji RX100 X-ray film either at room temperature or at −80°C with an intensifying screen.

2.2.8.2 Dot blot hybridisation
Allele-specific oligonucleotides (ASOs) were designed with the SNP site located at the eighth nucleotide from the 5’ end of the 18 base long oligo. The ASO (8 ng) was end-labelled at 37°C for 2 hours in a 10 μl mixture of 1 × kinase mix (70 mM Tris-HCl pH 7.5, 10 mM MgCl₂, 5 mM spermidine trichloride, 2 mM dithiothreitol), 3.5 units T4 polynucleotide kinase and 1.2 μCi γ-³²P-ATP. The labelling was terminated by addition of 20 μl kinase stop solution (25 mM EDTA, 0.1% SDS, 10 μM ATP). The dot blot membrane was pre-hybridised in 2 ml hybridisation solution [3M TMAC (tetramethylammonium chloride), 0.6% SDS, 1 mM EDTA, 10 mM sodium phosphate...
pH 6.8, 5 × Denhardt’s solution (0.1% ficoll 400, 0.1% polyvinylpyrrolidone, 0.1% BSA), 4 μg/ml yeast RNA] at 49.5°C for 10 minutes. The membrane was then hybridised with the labelled ASO plus 320 ng of unlabelled ASO of the opposite allele in 2.5 ml of fresh hybridisation solution at 49.5°C for 1 hour. Following hybridisation, the membrane was washed with three changes of 2.5–3 ml TMAC wash solution (3 M TMAC, 0.6% SDS, 1 mM EDTA, 10 mM sodium phosphate pH 6.8) at 56°C for a total of 20 minutes. The membrane was finally rinsed with 2 × SSC, blot-dried and either exposed to a storage phosphor screen (Amersham Biosciences) at room temperature or to Fuji RX100 X-ray film at −80°C with an intensifying screen.

2.2.8.3 Probe removal from membrane
Probes on membranes were removed by washing in boiling 0.1% SDS. The washing was continued (5–6 changes) and monitored by a Geiger counter until probe removal was complete. Membranes were rinsed in 2 × SSC and stored damp at 4°C.

2.2.9 Automated DNA sequencing
PCR product was either purified by Exo/Sap purification or electroelution as previously described. Sequencing reactions (10 μl) containing 20–30 ng/kb PCR product, 1 μl 3.2 μM sequencing primer, 1.5 μl 5 × Big Dye Terminator buffer and 1 μl ABI PRISM BigDye Terminator v3.1 Ready Reaction Mix (Applied Biosystems) were cycled at 96°C for 10 sec, 53°C for 5 sec and 60°C for 4 min for 25 cycles. Reactions were then mixed with 12 μl of 0.37% SDS and incubated at 98°C for 5 min and 25°C for 10 min. Dye Terminators were removed by DyeEx 2.0 spin kit (Qiagen) or PERFORMA DTR Gel Filtration Cartridges (EdgeBioSystems). Sequencing was performed on an ABI 3730 Capillary Sequencing System by the Protein and Nucleic Acid Chemistry Laboratory (PNACL), University of Leicester. DNA sequences were analysed, aligned and compared using software packages Factura and ABI AutoAssembler.

2.2.10 Statistical tests
All P-values were calculated using chi-square tests unless specified otherwise.
Table 2.2. PCR conditions used and described in Chapter 3

<table>
<thead>
<tr>
<th>Section</th>
<th>Target molecule (size, kb)</th>
<th>DNA template</th>
<th>Primers used</th>
<th>PCR buffer concentration</th>
<th>Additive</th>
<th>Cycling conditions</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1, 3.2.2</td>
<td>α-globin genes (1.4)</td>
<td>Man 1 sperm DNA</td>
<td>A13.6F + A15.0R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 34 x (96°C, 20°; 57/60/63/66°C, 30°; 64/66°C, 2.5°)</td>
<td>For optimisation, 5° and 3° probes</td>
</tr>
<tr>
<td>3.2.1</td>
<td>α-globin genes (1.4)</td>
<td>Man 1 sperm DNA</td>
<td>A13.6F2 + A15.0R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 34 x (96°C, 20°; 57/60/63/66°C, 30°; 64/66°C, 2.5°)</td>
<td>For optimisation, 5° and 3° probes</td>
</tr>
<tr>
<td>3.2.1</td>
<td>α-globin genes (0.9)</td>
<td>Man 1 sperm DNA</td>
<td>A24.7F + A25.6R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 34 x (96°C, 20°; 57/60/63/66°C, 30°; 64/66°C, 2.5°)</td>
<td>For optimisation, 5° and 3° probes</td>
</tr>
<tr>
<td>3.2.1</td>
<td>α-globin genes (1.0)</td>
<td>Man 1 sperm DNA</td>
<td>A24.7F + A25.7R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 34 x (96°C, 20°; 57/60/63/66°C, 30°; 64/66°C, 2.5°)</td>
<td>For optimisation, 5° and 3° probes</td>
</tr>
<tr>
<td>3.2.1</td>
<td>α-globin genes (12.0)</td>
<td>Man 1 sperm DNA</td>
<td>A13.6F + A25.6R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 63/66°C, 30°; 64/66°C, 50°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.1</td>
<td>α-globin genes (12.1)</td>
<td>Man 1 sperm DNA</td>
<td>A13.6F2 + A25.7R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 63/66°C, 30°; 64/66°C, 50°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.1</td>
<td>α-globin genes (1°; 12.1; 12.0)</td>
<td>Man 1 sperm and blood DNA, Man 2 blood DNA</td>
<td>1°: A13.6F + A25.7R 2°: A13.6F2 + A25.6R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) (1°: 23; 2°: 32) x (96°C, 20°; 63/66°C, 30°; 64/66°C, 16°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (4.7)</td>
<td>Man 1 sperm DNA</td>
<td>A13.6F + A18.2R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 5°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (4.6)</td>
<td>Man 1 sperm DNA</td>
<td>A13.6F2 + A18.2R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 5°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (9.3)</td>
<td>Man 1 sperm DNA</td>
<td>A13.6F + A22.9R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 10°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (7.6)</td>
<td>Man 1 sperm DNA</td>
<td>A18.1F + A25.7R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 10°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (7.5)</td>
<td>Man 1 sperm DNA</td>
<td>A18.1F + A25.6R2</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 10°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (3.4)</td>
<td>Man 1 sperm DNA</td>
<td>A22.3F + A25.7R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 5°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (3.3)</td>
<td>Man 1 sperm DNA</td>
<td>A22.3F + A25.6R2</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 5°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (0.2)</td>
<td>Man 1 sperm DNA</td>
<td>A18.1F + A18.2R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 2°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (0.6)</td>
<td>Man 1 sperm DNA</td>
<td>A22.3F + A22.9R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 60/63/66°C, 30°; 64/66°C, 2°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.2</td>
<td>α-globin genes (4.9)</td>
<td>Man 1 sperm DNA</td>
<td>A18.1F + A22.9R</td>
<td>0.2 x-1</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 50-66°C, 30°; 64/66°C, 5°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.3</td>
<td>α-globin genes (4.9)</td>
<td>Man 1 sperm DNA</td>
<td>A18.1F + A22.9R</td>
<td>0.2 x-1</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 50-66°C, 30°; 64/66°C, 5°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
<tr>
<td>3.2.3</td>
<td>α-globin genes (3.4)</td>
<td>Man 1 sperm DNA</td>
<td>A22.3F + A25.7R</td>
<td>0.2 x-1</td>
<td>-</td>
<td>1 x (96°C, 1°) 32 x (96°C, 20°; 50-66°C, 30°; 64/66°C, 5°)</td>
<td>Refer to Table 3.1. for annealing temperatures</td>
</tr>
</tbody>
</table>
| 3.2.4 | α-globin genes (3.4) | Man 1 sperm DNA | A22.3F + A25.7R | 0.7 or 1 x | 0.5-2 M DMSO | 1 x (96°C, 1')
|       |                     |                 |              |            |          | 32 x (96°C, 20'; 63/66°C, 30'; 68°C, 5/15')
| 3.2.4 | α-globin genes (3.4) | Man 1 sperm DNA | A22.3F + A25.7R | 0.7 or 1 x | 0.5-1.3 M betaine | 1 x (96°C, 1')
|       |                     |                 |              |            |          | 32 x (96°C, 20'; 63/66°C, 30'; 68°C, 5/15')
| 3.2.4 | α-globin genes (3.4) | Man 1 sperm DNA | A22.3F + A25.7R | 0.7 or 1 x | 0.1-0.5 M homoeoctolone | 1 x (96°C, 1')
|       |                     |                 |              |            |          | 32 x (96°C, 20'; 63/66°C, 30'; 68°C, 5/15')
| 3.2.4 | α-globin genes (3.4) | Man 1 sperm DNA | A22.3F + A25.7R | 0.7 or 1 x | 0.1-0.8 M sulfolane | 1 x (96°C, 1')
|       |                     |                 |              |            |          | 32 x (96°C, 20'; 63/66°C, 30'; 68°C, 5/15')
| 3.2.4 | α-globin genes (3.4) | Man 1 sperm DNA | A22.3F + A25.7R | 0.7 or 1 x | 1-5% glycerol | 1 x (96°C, 1')
|       |                     |                 |              |            |          | 32 x (96°C, 20'; 63/66°C, 30'; 68°C, 5/15')
| 3.2.4 | α-globin genes (12.0) | Man 1 sperm DNA | A13.6F2 + A25.6R2 | 0.7 x | 0.8-3 M betaine | 1 x (96°C, 1')
|       |                     |                 |              |            |          | 32 x (96°C, 20'; 63°C, 30'; 68°C, 16')

1.4 M betaine chosen to be the best concentration
Table 2.3. PCR conditions used and described in Chapter 4

<table>
<thead>
<tr>
<th>Section</th>
<th>Target molecule (size, kb)</th>
<th>DNA template</th>
<th>Primers used</th>
<th>PCR buffer concentration</th>
<th>Additive</th>
<th>Cycling conditions</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.1</td>
<td>α-globin genes (11: 12.1, 21: 12.0)</td>
<td>Man 1 sperm DNA</td>
<td>1°: A13.6F + A25.7R</td>
<td>0.7 x 1.4 M betaine</td>
<td>1 x (96°C, 1')</td>
<td>(1°: 30, 2°: 34) x (96°C, 20'; 63°C, 30'; 68°C, 16')</td>
<td>Single-molecule PCR, with 3-12 pg DNA per PCR</td>
</tr>
<tr>
<td>4.2.2</td>
<td>α-globin genes (11: 12.0, 21: 4.5, 4.9, 3.2)</td>
<td>Sperm DNA from 32-donors panel</td>
<td>1°: A13.6F + A25.6R</td>
<td>0.7 x 1.4 M betaine</td>
<td>1 x (96°C, 1')</td>
<td>(1°: 30, 2°: 34) x (96°C, 20'; 63°C, 30'; 68°C, 16')</td>
<td>ASO hybridization for SNP screening</td>
</tr>
<tr>
<td>4.2.3</td>
<td>α-globin genes (11: 12.1, 21: 12.0, 8.2, 7.8)</td>
<td>BamHI-digested DNA fractions</td>
<td>1°: A13.6F + A25.7R</td>
<td>0.7 x 1.4 M betaine</td>
<td>1 x (96°C, 1')</td>
<td>(1°: 30, 2°: 34) x (96°C, 20'; 63°C, 30'; 68°C, 16')</td>
<td>Mutant detection, with 0.03-1.7 µl of fraction per PCR</td>
</tr>
<tr>
<td>4.2.4</td>
<td>α-globin genes (9.7)</td>
<td>2° PCR product of deletions and progenitors A13.7F + A25.5R</td>
<td>0.7 x 1.4 M betaine</td>
<td>1 x (96°C, 1')</td>
<td>(1°: 30, 2°: 34) x (96°C, 20'; 63°C, 30'; 68°C, 16')</td>
<td>allele-specific PCR for separating mixed deletions</td>
<td></td>
</tr>
<tr>
<td>4.2.3.1</td>
<td>α-globin genes (6.3)</td>
<td>2° PCR product of deletions ASPA19.3C/TF + A25.5R</td>
<td>0.7 x 1.4 M betaine</td>
<td>1 x (96°C, 1')</td>
<td>(1°: 30, 2°: 34) x (96°C, 20'; 63°C, 30'; 68°C, 16')</td>
<td>allele-specific PCR for separating mixed deletions</td>
<td></td>
</tr>
</tbody>
</table>

59
| 4.2.4.3 | α-globin genes (5.0) | 2° PCR product of deletions | ASPA20.6C/GF + ASPA25.5R | 0.7 x | 1.4 M betaine | 1 x (96°C, 1') | 10 x (96°C, 20': 66°C, 30': 68°C, 11') | 14 x (96°C, 20': 65°C, 30': 68°C, 11') | 6 x (96°C, 20': 84°C, 30': 68°C, 11') | Allele-specific PCR for separating mixed deletions |
| 4.2.4.3 | α-globin genes (5.7) | 2° PCR product of deletions | A13.7F + ASPA19.4A/GR | 0.7 x | 1.4 M betaine | 1 x (96°C, 1') | 10 x (96°C, 20': 66°C, 30': 68°C, 11') | 14 x (96°C, 20': 65°C, 30': 68°C, 11') | 6 x (96°C, 20': 84°C, 30': 68°C, 11') | Allele-specific PCR for separating mixed deletions |
| 4.2.4.3 | α-globin genes (11.7) | 2° PCR product of deletions | A13.7F + ASPA25.5A/GR | 0.7 x | 1.4 M betaine | 1 x (96°C, 1') | 10 x (96°C, 20': 61(G)/64(A)°C, 30': 68°C, 11') | 14 x (96°C, 20': 60(G)/63(A)°C, 30': 68°C, 11') | 6 x (96°C, 20': 59(G)/62(A)°C, 30': 68°C, 11') | Allele-specific PCR for separating mixed deletions |
Table 2.4. PCR conditions used and described in Chapter 5

<table>
<thead>
<tr>
<th>Section</th>
<th>Target molecule (size, kb)</th>
<th>DNA template</th>
<th>Primers used</th>
<th>PCR buffer concentration</th>
<th>Additive</th>
<th>Cycling conditions</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.1.1</td>
<td>α-globin genes (≤12.0)</td>
<td>Sperm DNA from LRI panel (170 donors)</td>
<td>A13.6F2 + A25.6R2</td>
<td>0.7 x</td>
<td>1.4 M betaine</td>
<td>1 x (96°C, 1')&lt;br&gt;33 x (96°C, 20''; 63°C, 30''; 68°C, 16'')</td>
<td>40 ng DNA per PCR</td>
</tr>
<tr>
<td>5.2.1.2</td>
<td>α-globin genes (≤12.0)</td>
<td>Genomic DNA from CEPH panel (121 donors)</td>
<td>A13.6F2 + A25.6R2</td>
<td>0.7 x</td>
<td>1.4 M betaine</td>
<td>1 x (96°C, 17')&lt;br&gt;33 x (96°C, 20''; 63°C, 30''; 68°C, 16'')</td>
<td>40 ng DNA per PCR</td>
</tr>
<tr>
<td>5.2.1.3</td>
<td>α-globin genes (1°: ≤12.0; 2°: ≤11.8)</td>
<td>Genomic DNA from M. Jobling's panel (185 donors)</td>
<td>1°: A13.6F2 + A25.6R2&lt;br&gt;2°: A13.7F + A25.5R</td>
<td>0.7 x</td>
<td>1.4 M betaine</td>
<td>1 x (96°C, 1')&lt;br&gt;(1°: 36) x (96°C, 20''; 63°C, 30''; 68°C, 16'')&lt;br&gt;(2°: 34) x (96°C, 20''; 63°C, 30''; 66°C, 16'')</td>
<td>4 µl of DNA per PCR (1° PCR)</td>
</tr>
<tr>
<td>5.2.2</td>
<td>α-globin genes (≤11.8)</td>
<td>2° PCR product of deletions and progenitors</td>
<td>A13.7F + A25.5R</td>
<td>0.7 x</td>
<td>1.4 M betaine</td>
<td>1 x (96°C, 1')&lt;br&gt;22 x (96°C, 20''; 63°C, 30''; 66°C, 11/16'')</td>
<td>3° PCR for restriction mapping</td>
</tr>
<tr>
<td>Section</td>
<td>Target molecule (size, kb)</td>
<td>DNA template</td>
<td>Primers used</td>
<td>PCR buffer concentration</td>
<td>Additive</td>
<td>Cycling conditions</td>
<td>Remarks</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------------------</td>
<td>---------</td>
<td>--------------------</td>
<td>---------</td>
</tr>
<tr>
<td>6.2.1</td>
<td>α-globin genes (1.5: 3.6; 2: 3.8)</td>
<td>Genomic DNA (man 1 and man 2)</td>
<td>1°: A22.5F + A22.5F 2°: A22.5R + A22.5F</td>
<td>0.7 x</td>
<td>1.4 M betaine</td>
<td>1 x (96°C, 1') (1°: 30) (96°C, 20'; 54°C, 30'; 66°C, 6') (2°: 34) (96°C, 20'; 52°C, 30'; 66°C, 6')</td>
<td>1-16 ng DNA per PCR</td>
</tr>
<tr>
<td>6.2.2.2</td>
<td>α-globin genes (1.5: 3.3, 2: 3.3)</td>
<td>SpH-XbaI-digested DNA fractions (man 1)</td>
<td>1°: A21.8R2 + A22.3F 2°: A21.8R3 + A22.4F</td>
<td>0.9 x</td>
<td>0.5 M betaine</td>
<td>1 x (96°C, 1') (1°: 26, 2°: 34) (96°C, 20'; 68°C, 4.5')</td>
<td>Duplication detection, with 0.06-1.33 µl of fraction per PCR</td>
</tr>
<tr>
<td>6.2.2.2</td>
<td>α-globin genes (1.5: 3.7, 2: 3.6)</td>
<td>SpH-XbaI-digested DNA fractions + Affi digestion (man 2)</td>
<td>1°: A22.6R2 + A22.8F 2°: A22.6R + A22.8F2</td>
<td>0.9 x</td>
<td>0.5 M betaine</td>
<td>1 x (96°C, 1') (1°: 26, 2°: 34) (96°C, 20'; 68°C, 4.5')</td>
<td>Duplication detection, with 0.06-0.56 µl of fraction per PCR</td>
</tr>
<tr>
<td>6.2.2.3</td>
<td>11.0-kb SpH-XbaI-digested fragment from chromosome 11 (5.5)</td>
<td>10-fold diluted SpH-XbaI-digested DNA fractions</td>
<td>target4F + target4R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 15-27 x (96°C, 20'; 63°C, 30'; 66°C, 8')</td>
<td>Mutant distribution, 1 µl of 10-diluted fraction per PCR</td>
</tr>
<tr>
<td>6.2.2.3</td>
<td>11.0-kb SpH-XbaI-digested fragment from chromosome 11 (5.5)</td>
<td>Pools of DNA fractions</td>
<td>target4F + target4R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 15-27 x (96°C, 20'; 63°C, 30'; 66°C, 8')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
<tr>
<td>6.2.2.3</td>
<td>11.0-kb SpH-XbaI-digested fragment from chromosome 11 (5.5)</td>
<td>SpH-XbaI-digested DNA</td>
<td>target4F + target4R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 15-27 x (96°C, 20'; 63°C, 30'; 66°C, 8')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
<tr>
<td>6.2.2.3</td>
<td>11.0-kb SpH-XbaI-digested fragment from chromosome 11 (5.5: 2: 8.54)</td>
<td>SpH-XbaI-digested DNA fractions</td>
<td>1°: target4F + target4R 2°: target4F2 + target4R2</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') (1°: 26) (96°C, 20'; 63°C, 30'; 66°C, 8') (2°: 34) (96°C, 20'; 58°C, 30'; 66°C, 8')</td>
<td>Poisson estimation of amplifiable molecules, with 1.5-12 pg DNA per PCR</td>
</tr>
<tr>
<td>6.2.2.3</td>
<td>α-globin genes (4.9)</td>
<td>10-fold diluted SpH-XbaI-digested DNA fractions</td>
<td>A18.1F + A22.9R</td>
<td>0.7 x</td>
<td>0.75 M betaine</td>
<td>1 x (96°C, 1') 22-33 x (96°C, 20'; 63°C, 30'; 66°C, 8')</td>
<td>Poisson estimation of amplifiable molecules, with 1.5-12 pg DNA per PCR</td>
</tr>
<tr>
<td>6.2.2.3</td>
<td>α-globin genes (4.9)</td>
<td>Pools of DNA fractions</td>
<td>A18.1F + A22.9R</td>
<td>0.7 x</td>
<td>0.75 M betaine</td>
<td>1 x (96°C, 1') 22-33 x (96°C, 20'; 63°C, 30'; 66°C, 8')</td>
<td>Poisson estimation of amplifiable molecules, with 1.5-12 pg DNA per PCR</td>
</tr>
<tr>
<td>6.2.2.3</td>
<td>α-globin genes (4.9)</td>
<td>SpH-XbaI-digested DNA</td>
<td>A18.1F + A22.9R</td>
<td>0.7 x</td>
<td>0.75 M betaine</td>
<td>1 x (96°C, 1') 22-33 x (96°C, 20'; 63°C, 30'; 66°C, 8')</td>
<td>Poisson estimation of amplifiable molecules, with 1.5-12 pg DNA per PCR</td>
</tr>
<tr>
<td>6.2.3</td>
<td>α-globin genes (1.5: 3.2)</td>
<td>2° PCR products of duplication (man 1)</td>
<td>A21.8R4 + A22.4F2</td>
<td>0.9 x</td>
<td>0.5 M betaine</td>
<td>1 x (96°C, 1') 22-33 x (96°C, 20'; 60°C, 30'; 66°C, 7')</td>
<td>3° PCR for exchange point mapping</td>
</tr>
<tr>
<td>6.2.3</td>
<td>α-globin genes (1.5: 3.6)</td>
<td>2° PCR products of duplication (man 2)</td>
<td>A22.6R0 + A22.8F3</td>
<td>0.9 x</td>
<td>0.5 M betaine</td>
<td>1 x (96°C, 1') 24-33 x (96°C, 20'; 63°C, 30'; 68°C, 5')</td>
<td>3° PCR for exchange point mapping</td>
</tr>
<tr>
<td>6.2.5</td>
<td>3.8-kb AfII-digested DNA fragment from the MHC (1.8)</td>
<td>10-fold diluted AfII-digested DNA fractions</td>
<td>R85.7F + R87.5R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 24-33 x (96°C, 20'; 62°C, 30'; 64°C, 2')</td>
<td>Mutant distribution, 1 µl of 10-diluted fraction per PCR</td>
</tr>
<tr>
<td>6.2.5</td>
<td>3.8-kb AfII-digested DNA fragment from the MHC (1.8)</td>
<td>Pools of DNA fractions</td>
<td>R85.7F + R87.5R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 24-33 x (96°C, 20'; 62°C, 30'; 64°C, 2')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
<tr>
<td>6.2.5</td>
<td>3.8-kb AfII-digested DNA fragment from the MHC (1.8)</td>
<td>AfII-digested DNA</td>
<td>R85.7F + R87.5R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') (1°: 26) (96°C, 20'; 62°C, 30'; 64°C, 2') (2°: 34) (96°C, 20'; 60°C, 30'; 64°C, 2')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
<tr>
<td>6.2.5</td>
<td>3.8-kb AfII-digested DNA fragment from the MHC (1.8)</td>
<td>AfII-digested DNA</td>
<td>R85.7F + R87.5R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') (1°: 26) (96°C, 20'; 62°C, 30'; 64°C, 2') (2°: 34) (96°C, 20'; 60°C, 30'; 64°C, 2')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
</tbody>
</table>

Table 2.5. PCR conditions used and described in Chapter 6.
<table>
<thead>
<tr>
<th>Section</th>
<th>Target molecule (size, kb)</th>
<th>DNA template</th>
<th>Primers used</th>
<th>PCR buffer concentration</th>
<th>Additive</th>
<th>Cycling conditions</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2.1.3</td>
<td>4.2-kb EcoRV digested MHC fragment (3.6)</td>
<td>10-fold diluted EcoRV-digested DNA fractions</td>
<td>R55.4F + R59.0R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 19-27 x (96°C, 20'; 63°C, 30'; 64°C, 3.5')</td>
<td>Recovery estimation, 1 µl of 10-diluted fraction per PCR</td>
</tr>
<tr>
<td>7.2.1.3</td>
<td>4.2-kb EcoRV digested MHC fragment (3.6)</td>
<td>Pools of DNA fractions</td>
<td>R55.4F + R59.0R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 19-27 x (96°C, 20'; 63°C, 30'; 64°C, 3.5')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
<tr>
<td>7.2.1.3</td>
<td>4.2-kb EcoRV digested MHC fragment (3.6)</td>
<td>EcoRV-digested genomic DNA</td>
<td>R55.4F + R59.0R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 19-27 x (96°C, 20'; 63°C, 30'; 64°C, 3.5')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
<tr>
<td>7.2.1.3</td>
<td>4.2-kb EcoRV digested MHC fragment (3.6)</td>
<td>EcoRV-digested DNA fractions</td>
<td>1*: R55.4F + R59.0R 2*: R55.5F + R58.9R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') (1°: 26'; 2°: 34') x (96°C, 20'; 63°C, 30'; 64°C, 3.5')</td>
<td>Poisson estimation of amplifiable molecules, with 1.5-12 pg DNA per PCR</td>
</tr>
<tr>
<td>7.2.1.3</td>
<td>4.2-kb EcoRV digested MHC fragment (3.6)</td>
<td>Pools of DNA fractions</td>
<td>R-10.9F + R-8.8R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 19-27 x (96°C, 20'; 63°C, 30'; 64°C, 3.5')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
<tr>
<td>7.2.1.3</td>
<td>4.2-kb EcoRV digested MHC fragment (3.6)</td>
<td>EcoRV-digested genomic DNA</td>
<td>R-10.9F + R-8.8R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') 19-27 x (96°C, 20'; 63°C, 30'; 64°C, 3.5')</td>
<td>Recovery estimation, serial diluted input</td>
</tr>
<tr>
<td>7.2.1.3</td>
<td>4.2-kb EcoRV digested MHC fragment (3.6)</td>
<td>EcoRV-digested DNA fractions</td>
<td>1*: R-10.9F + R-8.8R 2*: R-10.7F + R-8.9R</td>
<td>1 x</td>
<td>-</td>
<td>1 x (96°C, 1') (1°: 26'; 2°: 34') x (96°C, 20'; 63°C, 30'; 64°C, 3.5')</td>
<td>Poisson estimation of amplifiable molecules, with 1.5-12 pg DNA per PCR</td>
</tr>
<tr>
<td>7.2.2.1</td>
<td>a-globin genes (3.8)</td>
<td>10-fold diluted EcoRV-digested DNA fractions</td>
<td>A18.1F + A21.8R2</td>
<td>0.9 x 0.5 M betaine</td>
<td>1 x (96°C, 1') 23-34 x (96°C, 20'; 60°C, 30'; 68°C, 4')</td>
<td>Progenitor distribution, 1 µl of 10-diluted fraction per PCR</td>
<td></td>
</tr>
<tr>
<td>7.2.2.1</td>
<td>a-globin genes (3.8)</td>
<td>Pools of DNA fractions</td>
<td>A18.1F + A21.8R2</td>
<td>0.9 x 0.5 M betaine</td>
<td>1 x (96°C, 1') 23-34 x (96°C, 20'; 60°C, 30'; 68°C, 4')</td>
<td>Estimation of progenitor contamination, serial diluted input</td>
<td></td>
</tr>
<tr>
<td>7.2.2.1</td>
<td>a-globin genes (3.8)</td>
<td>EcoRV-digested genomic DNA</td>
<td>A18.1F + A21.8R2</td>
<td>0.9 x 0.5 M betaine</td>
<td>1 x (96°C, 1') 23-34 x (96°C, 20'; 60°C, 30'; 68°C, 4')</td>
<td>Estimation of progenitor contamination, serial diluted input</td>
<td></td>
</tr>
<tr>
<td>7.2.2</td>
<td>a-globin genes (1°: -3.3; 2°: -3.3)</td>
<td>EcoRV-digested DNA fractions + DraI digestion (testis donor and man 2)</td>
<td>1°: A21.8R2 + A22.3F 2°: A21.8R3 + A22.4F</td>
<td>0.9 x 0.5 M betaine</td>
<td>1 x (96°C, 1') (1°: 26'; 2°: 34') x (96°C, 20'; 68°C, 4.5')</td>
<td>Circle detection, with 0.167-0.833 µl of fraction per PCR</td>
<td></td>
</tr>
<tr>
<td>7.2.2</td>
<td>a-globin genes (-3.2)</td>
<td>2° PCR products of linearised circular DNA molecules (testis donor and man 2)</td>
<td>A21.8R4 + A22.4F2</td>
<td>0.9 x 0.5 M betaine</td>
<td>1 x (96°C, 1') 22 x (96°C, 20'; 60°C, 30'; 68°C, 7')</td>
<td>3° PCR for exchange point mapping</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.7. PCR conditions used and described in Chapter 8

<table>
<thead>
<tr>
<th>Section</th>
<th>Target molecule (size, kb)</th>
<th>DNA template</th>
<th>Primers used</th>
<th>PCR buffer concentration</th>
<th>Additive</th>
<th>Cycling conditions</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2.2.1</td>
<td>α-globin genes (1°: ≤12.1; 2°: =12.0)</td>
<td>Sperm/blood DNA from 19-donor panel</td>
<td>1°: A13.6F + A25.7R 2°: A13.6F2 + A25.6R2</td>
<td>0.7 x 1.4 M betaine</td>
<td>1 x (96°C, 1') (1°: 30; 2°: 34) x (96°C, 20'; 63°C, 30'; 68°C, 16')</td>
<td>Test for α-globin gene deletions</td>
<td></td>
</tr>
<tr>
<td>8.2.2.1</td>
<td>α-globin genes (3.3)</td>
<td>Sperm/blood DNA from 19-donor panel + positive control</td>
<td>A21.8R3 + A22.4F</td>
<td>0.9 x 0.5 M betaine</td>
<td>1 x (96°C, 1') 36 x (96°C, 20'; 60°C, 4.5')</td>
<td>Test for ααα-mi3 duplication</td>
<td></td>
</tr>
<tr>
<td>8.2.2.1</td>
<td>α-globin genes (3.5)</td>
<td>Sperm/blood DNA from 19-donor panel</td>
<td>A18.2R + A19.1F</td>
<td>0.9 x 0.5 M betaine</td>
<td>1 x (96°C, 1') 36 x (96°C, 20'; 68°C, 6.5')</td>
<td>Test for ααα-mi4 duplication</td>
<td></td>
</tr>
<tr>
<td>8.2.2.2</td>
<td>α-globin genes (4.9)</td>
<td>2° PCR products of the test for α-globin gene deletions</td>
<td>A18.1F + A22.9R</td>
<td>0.7 x 1.4 M betaine</td>
<td>1 x (96°C, 1') 20 x (96°C, 20'; 63°C, 30'; 66°C, 16')</td>
<td>3° PCR for resequencing</td>
<td></td>
</tr>
<tr>
<td>8.2.2.2</td>
<td>α-globin genes (3.2)</td>
<td>2° PCR products of the test for α-globin gene deletions</td>
<td>A22.3F + A25.5R</td>
<td>0.7 x 1.4 M betaine</td>
<td>1 x (96°C, 1') 20 x (96°C, 20'; 63°C, 30'; 66°C, 16')</td>
<td>3° PCR for resequencing</td>
<td></td>
</tr>
<tr>
<td>8.2.3</td>
<td>α-globin genes (4.2)</td>
<td>Sperm/blood DNA from 19-donor panel</td>
<td>A24.7F + A28.9R</td>
<td>0.9 x 1.2 M betaine</td>
<td>1 x (96°C, 1') 36 x (96°C, 20'; 68°C, 7.5')</td>
<td>3° extended region for resequencing</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3

PCR Amplification of Human α-Globin Genes

3.1 Introduction

The α-globin gene cluster is located near the telomeric end of the short arm of human chromosome 16 (16p13.1). It contains an embryonic gene (ζ), two foetal-adult genes (α2 and α1), three pseudogenes (ψζ, ψα2 and ψα1) and one θ gene with an unknown function (Fig. 1.8). The cluster also shows a high density of Alu-family repeats and high GC-content (60%). These genes, particularly the ζ- and α-globin genes in company with the β-related globin genes (ε, γ, δ, β), play an important role in globin-chain synthesis in normal haemoglobins. Possible ectopic exchanges between homology blocks in the α-globin gene cluster can lead to α⁺-thalassaemia by deletion of a single α-globin gene. This prevalent monogenic disease has been well studied in α⁺-thalassaemics, although carriers of −α chromosomes usually do not have significant haematological differences compared to normal individuals (Higgs et al. 1989; Higgs 1993; Wang 2000). Since studies on somatic samples from α-thalassaemia patients first started in the late 1940s (Weatherall 2004), many prenatal diagnoses and clinical detection techniques have been well established. Despite the excellent development of clinical studies and extensive description of DNA rearrangements in α-thalassaemia patients, basic research on ectopic recombination between local repeats, investigating the dynamics and mechanisms of ectopic exchange, have never been performed. In addition, possibly owing to the length of the gene cluster and its high GC-content, PCR amplification of α-globin genes carrying all homology blocks (12.1 kb) has never been reported. Although this amplification has been proved to be extremely challenging, it is essential not only for detecting de novo deletions of α-globin genes arise by ectopic recombination, but also for studying the relationship between progenitor DNA molecules (αα) and deletion mutants (−α). This chapter describes strategies used to develop successful PCR amplification of the entire α-globin gene region.

3.1.1 Long PCR amplification

Long PCRs amplifying several kilobases (from 5 kb up to 40 kb) usually require more than one kind of DNA polymerase (Kainz et al. 1992; Barnes 1994). Taq in conjunction with Pfu is the most common combination of polymerases used in long PCRs. Taq
provides efficient amplification, while *Pfu*, with 3' to 5' exonuclease proofreading activity, ensures high fidelity during reactions and overcomes stalled extensions by *Taq* polymerase. Denaturation and extension time are also crucial factors. The success of long PCR amplification can be hampered in the early stage of PCR if double-stranded DNA templates are not thoroughly denatured and/or if insufficient extension time is given for amplifying the target molecule.

### 3.1.2 PCR amplification of GC-rich targets

Amplifying GC-rich DNA templates is always cumbersome in long PCRs. With the formation of secondary structures and the high melting temperature of DNA templates, high denaturation temperature is needed for complete template separation. However, this may reduce the stability of DNA polymerases (Yap and McGee 1991). Similarly, the salt concentration in the PCR cocktail is also important for template separation. The repulsive force between two negatively charged DNA strands can be balanced by positive ions such as Mg\(^{2+}\), allowing double-stranded DNA templates tangled in a stable condition. Hence, reducing the ionic strength of the PCR reaction can give a similar effect to increasing the denaturation temperature. Furthermore, the addition of co-solvents, for example, dimethyl sulfoxide (DMSO) and betaine has been reported to improve the specificity and efficiency of PCR amplification (Winship 1989; Henke et al. 1997). However, the performance of these chemicals is unpredictable. Systematic analyses of the effects of buffer concentration, cycling conditions and PCR additives on the efficiency of PCR amplification of the α-globin gene cluster were therefore performed.

### 3.2 Results

#### 3.2.1 Attempts to amplify across the α-globin gene locus

To detect any *de novo* deletion mutants arising from unequal crossover between homology blocks, the PCR amplification strategy was to amplify the region of the α-globin gene cluster carrying the ψα1, α2 and α1 genes using PCR primers in flanking regions (Fig. 3.1). PCR primers were designed in regions outside homology blocks to prevent any mispriming from paralogous sequences.

The goal was to optimise PCR to the point where nested PCR with primers A13.6F plus A25.7R followed by A13.6F2 plus A25.6R could be used to efficiently
amplify the 12.0-kb region containing the \(\psi\alpha1\), \(\alpha2\) and \(\alpha1\) genes at the single-DNA-molecule level. These primers were tested by amplifying short amplicons in the 5' and 3' regions using combinations A13.6F plus A15.0R and A24.7F plus A25.7R, respectively (Fig. 3.1A). All primers proved to be efficient.

The first attempts to amplify across the entire locus (12.0 kb) used different primer combinations. However, multiple bands were detected but none of them showed the correct size (Fig. 3.1B). These bands suggested that they were products of ectopic priming from elsewhere in the human genome or that the man tested might carry large deletions within the \(\alpha\)-globin gene region giving PCR products with reduced size. Nested PCRs with DNA templates from two men were then performed to improve the specificity of PCR and also to confirm the absence of any local deletions in the first man. Very similar profiles were detected in both blood and sperm samples from the two men (Fig. 3.1C). However, the multiple bands suggested that they were ectopic PCR products from the carrier DNA, since same profiles were observed in the absence of human DNA. Although the profiles of PCR product were almost identical between two men, no PCR products of the correct \(\alpha\)-globin size could be seen.

To test whether the target PCR product might exist but in a very low amount, a Southern blot hybridisation was done to detect any trace of target (Fig. 3.1D). No signal of the target amplicon was detected, indicating no amplification of the entire \(\alpha\)-globin gene region.
Fig. 3.1. Assays for amplification of the α-globin gene locus. (A) The region of the α-globin gene cluster analysed, carrying ψα1, α2 and α1 genes. Homologous DNA sequences within the region are highlighted in colour. PCR primers are indicated by arrows below the region. (B) PCR amplification with different primer combinations. PCR products were analysed by gel electrophoresis and visualised by staining with ethidium bromide. Two annealing temperatures (63 and 66°C) were tested; the first two lanes are duplicates, and the negative control (no DNA template) is shown in the last lane. M, λ DNA × *Hind*III and φX174 DNA × *Haelll*. (C) Nested PCRs across the α-globin gene region. Secondary PCR products of nested PCRs (1° PCR, A13.6F + A25.7R; 2° PCR, A13.6F2 + A25.6R) were analysed by gel electrophoresis. They were amplified at two annealing temperatures (63°C shown in the first lane and 66°C for the second) with different DNA templates; 1, sperm DNA from man 1; 2, blood DNA from man 1; 3, blood DNA from man 2; 4, a replicate of 3; 5, carrier DNA only. (D) Southern blot detection of the α-globin gene region. PCR products shown in Fig. 3.1B (b) and Fig. 3.1 C (c) were blotted and hybridised with the 5′ (i) and 3′ (ii) probes.
3.2.2 Breaking up the α-globin gene cluster

DNA fragments with regions of high GC-content are always an obstacle for PCR amplification. The entire α-globin gene cluster is a well-known example of a GC-rich region that has never been amplified in its entirety. In particular, the two GC-peaks (80% GC) in the 5' regions of the α2 and α1 genes (Figs. 1.8 and 3.2A) are likely to make simple long PCRs not feasible. These putative PCR blockers were mapped by dividing the 12.1-kb target into subamplicons that could be amplified with different primer combinations (Fig. 3.2B). Primers were designed outside homology blocks as well as the two GC-peaks to test whether the failure of amplification was due to blockage by GC-peaks. Results revealed that only amplicons outside GC-peaks were successfully amplified, while the amplification of regions carrying α2 and/or α1 genes failed (Fig. 3.2C). These implied that the failure of the whole cluster amplification was mainly due to the presence of GC-peaks, since regions with either one or two GC-peaks could not be amplified.
Fig. 3.2. PCR amplification of subamplicons. (A) The region analysed as described for Fig. 3.1, with GC-peaks and primer positions shown. (B) Location of subamplicons. The position of each subamplicon is marked with a horizontal line in red or black to represent the success or failure of amplification, respectively. (C) Gel electrophoresis of PCR products from subamplicons. Three different annealing temperatures (60, 63 and 66°C from left to right) were tested in each combination. The primer combination and the size of amplicon are listed in Fig. 3.2B corresponding to the number. The locations of predicted PCR product are shown by red arrows. M, λ DNA × HindIII and φX174 DNA × HaeIII.
3.2.3 Alteration of PCR buffer concentration

Further optimisation of long PCR focussed on relatively short amplicons (3–5 kb) separating the α2 or α1 gene and containing a single GC-peak. GC-rich DNA is likely to be difficult to denature and could also adopt secondary structures that might block PCR extension. Since the denaturation temperature used (96°C) was already very high, further increase seemed not feasible not only for the integrity of DNA templates but also for the PCR machine. The PCR buffer concentration was therefore systematically reduced in order to alleviate this denaturation problem. The relation between PCR buffer concentration and the melting temperature of DNA templates is logarithmic, and a drop of 18° in melting temperature can be achieved by a 10-fold reduction of buffer concentration (Table 3.1) (A. Jeffreys, personal communication).

PCRs with serial dilutions of PCR buffer were performed in an attempt to amplify regions carrying a single GC-peak. These regions were now successfully amplified, but only in a particular range of PCR buffer concentration (Fig. 3.3A). For instance, the 4.9-kb amplicon appeared when 0.8 × – 0.4 × PCR buffer was used, while the 3.4-kb amplicon was only amplified in a narrower range of PCR buffer concentration (0.8 × – 0.6 ×). The authenticity of these amplicons was further checked by restriction digestion with HindIII to produce the correct α-globin fragmentation patterns (Fig. 3.3B).

Although DNA regions with one GC-peak were successfully amplified, amplification of the entire α-globin gene cluster still failed, but instead generated intense non-globin ectopic PCR products (data not shown). These results showed that the problem of amplifying regions with high GC-content could be resolved by reducing salt concentration in PCR cocktails, but indicated that this alteration was not enough for successful amplification of the entire α-globin gene cluster. Further optimisations were necessary.
Table 3.1. The logarithmic relationship between PCR buffer concentration and DNA melting temperature

<table>
<thead>
<tr>
<th>PCR buffer concentration</th>
<th>Reduction of melting temperature (°)</th>
<th>Corresponding changes of melting temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ×</td>
<td>0</td>
<td>63 96</td>
</tr>
<tr>
<td>0.9 ×</td>
<td>0.8</td>
<td>62.2 95.2</td>
</tr>
<tr>
<td>0.8 ×</td>
<td>1.7</td>
<td>61.3 94.3</td>
</tr>
<tr>
<td>0.7 ×</td>
<td>2.8</td>
<td>60.2 93.2</td>
</tr>
<tr>
<td>0.6 ×</td>
<td>4.0</td>
<td>59.0 92.0</td>
</tr>
<tr>
<td>0.5 ×</td>
<td>5.4</td>
<td>57.6 90.6</td>
</tr>
<tr>
<td>0.4 ×</td>
<td>7.2</td>
<td>55.8 88.8</td>
</tr>
<tr>
<td>0.3 ×</td>
<td>9.4</td>
<td>53.6 86.6</td>
</tr>
<tr>
<td>0.2 ×</td>
<td>12.6</td>
<td>50.4 83.4</td>
</tr>
<tr>
<td>0.1 ×</td>
<td>18</td>
<td>45 78</td>
</tr>
</tbody>
</table>

Two melting temperatures are shown, arbitrarily chosen as 63° and 96° in 1 × PCR buffer.
Fig. 3.3. PCR amplification with serial dilutions of PCR buffer. (A) The region analysed as described in Fig. 3.2, with locations of HindIII restriction sites shown. Amplification targets are marked with horizontal lines, and the amplicon size is listed below. (i) The 4.9-kb fragment amplified using primers A18.1F plus A22.9R was analysed by gel electrophoresis and indicated by an arrow. Two annealing temperatures (63 and 66°C for left and right lanes, respectively) were used for every set of PCRs except the positive control (+ve). (ii) Another fragment (3.4 kb) carrying one GC-peak was amplified using primers A22.3F plus A25.7R. M, λ DNA × HindIII and 8X174 DNA × HaeIII. (B) Restriction profiles of HindIII digests analysed by gel electrophoresis. Sizes of intact target amplicons (U) and HindIII digested fragments (D) were indicated by black and red arrows, respectively. The HindIII fragment sizes confirm to those predicted from the α-globin gene sequences.
3.2.4 Successful amplification with PCR additives

Two GC-rich amplicons were successfully amplified but with different ranges of optimised PCR buffer concentration. Overall, $0.7 \times$ PCR buffer was chosen to be the most favourable working concentration for both amplicons. Nevertheless, PCR was still erratic, with the target amplicon sometimes not being amplified (Fig. 3.4A). Some PCR additives have been reported as enhancing agents that can improve the specificity, consistency and product yield of PCR amplification (Bookstein et al. 1990; Henke et al. 1997; Iakobashvili and Lapidot 1999; Chakrabarti and Schutt 2001; Schnoor et al. 2004). Depending on the length of amplicons, GC-content and potential secondary structures, PCR additives can give different levels of improvement to the amplification (Baskaran et al. 1996; Schnoor et al. 2004). Hence, a range of additives (DMSO, betaine, homoectoine, sulfolane and glycerol) was tested at different working concentrations. It was clear that PCR additives acted differently in amplifying the same target fragment (Fig. 3.4A). For instance, $0.5 \ M$ betaine gave target products with strong specificity and intensity, whereas similar results were achieved with homoectoine and sulfolane but at lower concentration. Moreover, with optimised concentrations of PCR additives, the subamplicon was amplified in substantial quantity without any aspecific PCR products. These results proved that PCR additives improved the specificity of amplification and greatly increased the yield of target PCR products. However, successful PCR with additives could not be obtained without reduction of PCR buffer concentration (Fig. 3.4A). This indicated that successful amplification of the entire $\alpha$-globin gene cluster might only be possible in the presence of PCR additives and with reduced salt concentration in the PCR reactions.

Further optimisations were performed using selected concentration of PCR additives obtained from pilot experiments along with increased extension time (Fig. 3.4B). Results revealed that the yield of PCR products increased significantly when prolonged extension time was used (15 minutes for the 3.4-kb fragment). This suggests inefficient extension perhaps caused by secondary structures in the templates that can be ameliorated by long extension time. Since PCR amplification with betaine guaranteed the most specific and abundant yield of target amplicon, betaine was chosen as the PCR additive used for amplifying the 12.1-kb $\alpha$-globin gene cluster. PCRs with serial dilutions of betaine were performed in an attempt to obtain target products with the best specificity, efficiency and quantity (Fig. 3.4C). The 12.0-kb $\alpha$-globin gene target was efficiently amplified in the optimised conditions using $0.7 \times$ PCR buffer and 1.4 M betaine with 16-minute extension time (see Table 2.2).
Fig. 3.4. Gel electrophoretic analyses of PCR products amplified in the presence of PCR additives and reduced PCR buffer concentration. (A) PCR amplification of the GC-rich target amplified with primers A22.3F and A25.7R, using 0.7 × or 1 × PCR buffer in conjunction with PCR additives. M, λ DNA × HindIII and φX174 DNA × HaeIII. (B) PCR amplification using two extension times in the presence of 0.7 × PCR buffer and selected concentrations of PCR additives. (C) Successful amplification of the whole α-globin gene cluster using 0.7 × PCR buffer and serial dilutions of betaine. The 12.0-kb target and the positive control (3.4 kb) are marked with arrows.
3.3 Discussion

3.3.1 GC-content of human α-globin genes

Compared to the average GC-content of the human genome (40% GC), the human α-globin gene complex, located in the GC-isochore at the telomeric region, is very GC-rich (~60% GC) (Fischel-Ghodsian et al. 1987a). Although the GC-composition varies a lot across the human genome, regions with high GC-content are commonly found in mammalian housekeeping genes (mostly associated with non-methylated CpG islands) (Gardiner-Garden and Frommer 1987). In contrast, tissue specific α-globin genes, with high GC-content and non-methylated CpG islands in the 5' flanking regions of the α2 and α1 genes (Bird et al. 1987), share this feature with most housekeeping genes. This contrasts with β-globin genes (AT-rich and lack of CpG dinucleotides) (Fischel-Ghodsian et al. 1987b), created a challenging problem for PCR amplification of the entire cluster. Successful amplification of the 12.0-kb α-globin gene locus using long PCRs with the reduction of salt concentration and supplement of PCR additives was demonstrated for the first time in this chapter. Although the optimisation process was very challenging and laborious, this success was crucial for the detection of potential de novo deletions.

3.3.2 Importance of salt concentration and PCR additives

Complete denaturation of double-stranded DNA is essential for the success of PCR amplification. This strand-separating process mainly depends on temperature and salt concentration, particularly for Mg$^{2+}$ ions. Mg$^{2+}$ ions play a crucial role in PCR amplification since they chelate with dNTPs, PCR primers and DNA templates. In addition, the repulsion from negatively charged phosphate groups in double-stranded DNAs can be neutralised by cations such as K$^+$ and Mg$^{2+}$ ions, stabilising the helix structure. Results have revealed that the amplification of regions with high GC-content was significantly affected by the salt concentration (see section 3.2.3), and that optimisations by altering PCR buffer concentration was essential.

Although details of how PCR additives interact with DNA molecules are not fully known, many additives have been widely used in amplifying GC-rich regions to counterbalance limitations caused by high melting temperature and to improve the yield and specificity of PCR. For example, without altering the conformation of double-stranded DNAs, betaine can eliminate the base pair composition dependence of DNA thermal melting transitions, giving a similar stability between GC and AT base pairs.
This ‘isostabilising’ effect reduces the formation of secondary structure and enhances the amplification of GC-rich templates (Henke et al. 1997). In addition, betaine has also been reported to protect proteins against heat and repeated freeze/thaw treatment (Goeller and Galinski 1999; Knapp et al. 1999). These properties might explain why the addition of betaine was crucial in the success of α-globin genes amplification. In contrast, some of these organic co-solvents (Fig. 3.5) have been reported to enhance the efficiency of amplification by unwinding the DNA double-helix (Lee et al. 1981). Organic additives like DMSO and sulfolane were proposed to enhance the amplification of GC-rich regions by destabilising the double-helix (Cheng et al. 1994; Chakrabarti and Schutt 2001). The oxygen group in these compounds may bind strongly to the major and minor grooves of template DNA and unwind the helix structure by breaking hydrogen bonds between bases on complementary strands. Differences between effective ranges of these additives may also be determined by their structures and sizes (Chakrabarti and Schutt 2001). The smaller the size of additives, the more molecules can bind in the grooves. The groove is comparatively easy to saturate by bigger additives at lower concentrations, giving a narrower effective range. This hypothesis might explain why DMSO has a broader effective range than sulfolane in amplifying the subamplicon (Figs. 3.4A and 3.5).

![Fig. 3.5. Structures of PCR additives tested.](image)

---

77
3.4 Conclusions

PCR amplification of the entire α-globin gene cluster is essential not only for the detection of all potential ectopic α-globin gene deletions but also for the study of the relationships between progenitors (αα) and deletion mutants (−α). This crucial step has been achieved systematically by amplifying subamplicons within the locus, reducing PCR buffer concentration and with supplement of PCR additives. The target 12.0-kb amplicon was finally amplified in an optimised condition using 0.7 × PCR buffer and 1.4 M betaine.
Chapter 4

The Dynamics of α-Globin Gene Deletion

4.1 Introduction

Ectopic recombination initiated by the misalignment of non-allelic but highly similar DNA sequences can result in deletions, duplications, inversions and translocations (Inoue and Lupski 2002; Shaw and Lupski 2004; Feuk et al. 2006). It is of fundamental importance in generating copy number variations in human DNA (Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005), changing gene number in gene families and often leading to pathological disorders (Bailey et al. 2002; Stankiewicz and Lupski 2002). The dynamics of ectopic recombination have been extensively studied in model organisms such as yeast (Petes and Hill 1988; Mieczkowski et al. 2006), whereas studies of these processes in humans have only focused on characterising pathological DNA changes seen in patients (Baumer et al. 1998; Potocki et al. 2000; Potocki et al. 2007).

Analysis of familial and de novo DNA rearrangements in genomic disorders, both in patients and by accessing mutations in sperm, has established the importance of ectopic recombination between widely separated repeats in driving large-scale duplications and deletions in the human genome (Han et al. 2000; Emanuel and Shaikh 2001; Yatsenko et al. 2005). For example, ectopic recombination between two 24-kb DNA repeats separated by 1.4 Mb intervening DNA sequence can lead to duplications and deletions, resulting in the inherited disorders Charcot–Marie–Tooth type 1A (CMT1A) and hereditary neuropathy with liability to pressure palsies (HNPP), respectively (Pentao et al. 1992; Reiter et al. 1998). Ectopic exchange points within the repeat elements cluster within a 0.7-kb hot spot and are often associated with patchy gene conversion (Reiter et al. 1996; Reiter et al. 1998; Lopes et al. 1999; Han et al. 2000). In addition, duplications of paternal origin are found to arise by unequal exchange between homologous chromosomes, while maternal duplications and deletions are produced by ectopic exchanges both within and between homologous chromosomes. These differences provided evidence that there are distinct sex-dependent mechanisms of CMT1A/HNPP ectopic recombination (Lopes et al. 1998).

In contrast, little is known about the dynamics of ectopic recombination between locally repeated DNA sequences. Ectopic exchanges between homology blocks of kilobase-separated duplicated genes, for instance in the β-globin gene family, can lead to duplications, deletions and the creation of fusion genes as seen in Hb Lepore (Efremov
Although extensive work on these rearrangements in patients has been done, little is known about the dynamics and processes of these ectopic exchanges.

The \( \alpha \)-globin gene family is another classic system for studying ectopic recombination between locally repeated DNA sequences (Embury et al. 1979; Higgs et al. 1989). Single \( \alpha \)-globin gene deletions (\( -\alpha \) chromosomes) arise by unequal crossover between homology blocks associated with the \( \alpha 2 \) and \( \alpha 1 \)-globin genes, and can lead to \( \alpha^+ \)-thalassaemia. These deletions have been well characterised in \( \alpha^+ \)-thalassaemics and classified into different types based on the sites of unequal exchanges and the lengths of deletions. \( -\alpha^{3.7} \) and \( -\alpha^{4.2} \) are the most common genotypes worldwide found in \( \alpha^+ \)-thalassaemics (Embury et al. 1980; Higgs et al. 1984; Hill et al. 1985). Their prevalence is most likely maintained by malaria selection since \( \alpha^+ \)-thalassaemia frequencies correlate strongly with the endemicity of malaria in some populations such as Melanesians (Flint et al. 1986a; Hill 1992). However, owing to the selective advantage of \( -\alpha \) chromosomes, the actual deletion frequencies, dynamics and processes that generate these DNA rearrangements can only be determined from population prevalence. To elucidate these issues, germ-line analyses on individuals have to be performed.

To study the dynamics and processes of \textit{de novo} deletion arising from ectopic recombination, mutants could be screened at the single-DNA-molecule level from sperm DNA using small pool PCR (SP-PCR). However, it would be difficult to estimate the authenticity of any deletion mutants detected by this approach; some or all PCR products might be SP-PCR artefacts arising by jumping PCRs between paralogous sequences during the early stage of amplification. Previous studies of human minisatellite MS32 showed that these potential problems could be solved by a combination of physical enrichment and single-molecule PCR strategies (Jeffreys and Neumann 1997). These strategies have shown that authentic deletion/duplication mutant DNA molecules can be recovered by single-molecule PCR after complete physical separation of mutants from progenitor molecules. Furthermore, their authenticity can be validated by size analysis.

The work described in this chapter shows that it is possible to use the strategies of physical enrichment and single-molecule PCR to detect deletion mutants arising from unequal crossover between duplicated \( \alpha \)-globin genes directly in genomic DNA. This chapter also shows how these \textit{de novo} deletions have been used to reveal basic features of ectopic exchange between locally repeated DNA sequences, including the germ-line specificity of deletion, the role of meiotic recombination in the deletion process, and the dynamics of ectopic exchange.
4.2 Results

4.2.1 Single-molecule PCR of α-globin genes

To understand the dynamics and processes of ectopic recombination occurring within α-globin genes, *de novo* deletion mutants have to be detected and analysed at the single-DNA-molecule level. Chapter 3 established that PCR amplification of the entire ψα1-α2-α1 interval (12.1 kb) was possible. The first question therefore was whether this amplification was possible at the single-DNA-molecule level. Using the optimised PCR conditions described in Chapter 3, nested PCRs (details in Table 2.3) with different very low DNA inputs were performed (Table 4.1). Clear progenitor PCR products were generated increasing in frequency with DNA input (Fig. 4.1). The maximum likelihood number of amplifiable molecules per haploid genome (3 pg DNA) was estimated at 0.72 (95% confidence intervals 0.4–1.2) by Poisson analysis (programme written in True BASIC 4.1 by A. Jeffreys). Poisson analysis describes the probability functions of phenomena distributed at varying levels in populations. For single-DNA-molecule analysis, the relationship between the mean number of amplifiable DNA molecule (m) per PCR and the proportion of PCR reactions that are negative (P₀) is given by P₀ = e⁻ᵐ. The Poisson analysis programme allows data from different DNA inputs to be combined to give an overall maximum likelihood estimate of the number of molecules per unit of DNA (Table 4.1). These results indicated that this PCR system allowed the efficient amplification of the 12.1-kb progenitor molecule at the single-DNA-molecule level with, on average, one amplifiable molecule detected per 4.2 pg of genomic DNA and giving a single-molecule PCR efficiency of 72%, typical of the amplification efficiency of other long but not GC-rich DNA targets (Jeffreys et al. 2000).

Table 4.1. An example of Poisson analysis of single-molecule PCR

<table>
<thead>
<tr>
<th>DNA input in each 1° PCR (pg)</th>
<th>Total no. of nested PCRs</th>
<th>No. of +ve 2° PCRs</th>
<th>No. of -ve 2° PCRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>16</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>
Fig. 4.1. Gel electrophoretic analysis of Poisson analysis of single-molecule PCR products. The α-globin gene cluster (12.1 kb) was amplified using optimised conditions with approximately 3, 6 or 12 pg DNA input per PCR in 1° PCR. 2° PCR products (12.0 kb, indicated by arrows) were analysed by agarose gel electrophoresis and visualised by staining with ethidium bromide. M, λ DNA × HindIII.

4.2.2 Selection of informative donors

In order to have detailed understanding of the processes of ectopic recombination including the haplotype origin of recombinants, it is necessary to select donors who carried informative SNP heterozygosities across the α-globin gene cluster for analysis. Surveying the NCBI SNP database revealed many putative SNPs across the α-globin gene region. Some of these dbSNPs have been validated, for example, in the International HapMap Project (Gibbs et al. 2003), while many were invalidated, with a substantial proportion located within homology blocks; such variants might not be SNPs but instead PSVs. Since DNA donors with informative SNP heterozygosities along the α-globin genes are essential for the study of dynamics and mechanisms of ectopic recombination, a panel of 32 semen donors was screened for heterozygous SNPs by ASO hybridisation to PCR products of three subamplicons derived from the 12.1-kb α-globin gene target (Fig. 4.2). Homology blocks along the α-globin gene region were separated into three subamplicons to avoid any ASO mispriming to paralogous sequences. Based on information from the NCBI SNP database and resequencing data of the three subamplicons from man 1 and man 2, a total of 67 pairs of ASOs were designed. However, only 22 genuine SNPs including three identified by resequencing were found along the cluster in these donors. Man 1 and man 2 were chosen for further analyses because of the abundance of informative SNP heterozygosities (Fig. 4.2C), plus the availability of semen and blood samples.
Man 1 was of north European origin, while man 2 was of mixed north European and Indian ancestry. To ensure these two men lacked any constitutional \(\alpha\)-globin gene deletions or duplications as well as to determine the linkage phase of SNPs, single progenitor molecules (12.1-kb) were amplified from extreme dilutions of genomic DNA and genotyped by dot blot hybridisation (Fig. 4.3). Both men showed two distinct SNP haplotypes each on a 12.1-kb long molecule, establishing linkage phase and proving that neither man carried a constitutional \(\alpha\)-globin gene rearrangement.

**Fig. 4.2.** Identification of informative semen donors. (A) PCR amplifications for dot blot hybridisation. The 12.1-kb \(\alpha\)-globin gene region was amplified in a 1\(^{\text{st}}\) PCR and three subamplicons were amplified in 2\(^{\text{nd}}\) PCRs using diluted 1\(^{\text{st}}\) PCR product as DNA template to separate homology blocks within the cluster. Target amplicons are marked with horizontal lines. Heterozygous SNPs within the region analysed are marked with white circles. (B) Examples of autoradiographs of ASO hybridisation in man 1. (C) Genotyping results of the 32-donor panel. SNP heterozygosities are highlighted in red. \(K = G + T, M = A + C, R = A + G, S = C + G, Y = C + T.\)
Fig. 4.3. Genotyping of single-molecule-amplified progenitors. (A) Examples of autoradiographs of ASO hybridisation for characterising the linkage phase of SNPs. 5’ and 3’ flanking SNPs were characterised by ASOs, A15.9A/G and A25.5A/G respectively, in eight single progenitor molecules from man 2. Three of these molecules are from one haplotype and five from the other. (B) All SNP heterozygosities along the α-globin genes were separated into two haplotypes in two men. SNPs with an asterisk are examples shown in A.

4.2.3 Physical enrichment of deletion mutants

4.2.3.1 Preparative gel electrophoresis for separating deletion molecules from progenitors

Studies of patients with α-thalassaemia have revealed two classes of deletion that can arise by ectopic exchange between homology blocks in the α-globin gene region: –α^4.2_ deletions that extend leftward from the α2-globin gene and lose 4.2 kb of DNA and –α^3.7_ deletions that extend rightward (Fig. 4.4A) (Embury et al. 1980; Higgs et al. 1989). Genomic DNA was digested with BamHI to release 14.1-kb progenitor molecules carrying ψα1-, α2- and α1-globin genes, and also any shorter (9.9, 10.3 kb) de novo –α^4.2_ and –α^3.7_ deletion molecules. BamHI, which has no star activity, is a cheap and robust enzyme. In addition, with pairs of restriction sites located in both 5' and 3' flanking regions (Fig. 4.4A), the potential problem of incomplete digestion can be minimised. Thus, BamHI was chosen for DNA digestion prior to preparative gel electrophoresis. The single-molecule amplification efficiency of progenitor molecules from BamHI-digested DNA was also tested and found to be very similar to the efficiency obtained from undigested DNA (data not shown).
Bulk sperm and blood DNA from both men were BamHI-digested and fractionated to enrich for deletion mutants. Given the unknown but probably low frequency of deletion molecules, large amounts (100–180 µg) of DNA were processed. BamHI-digested DNA was loaded into a preparative gel and electrophoresed to separate mutant DNA molecules from progenitors (Fig. 4.4B) (Jeffreys and Neumann 1997). Both classes of deletion mutant could be simultaneously recovered and physically enriched in size fractions in which progenitor DNA molecules were heavily depleted. An aliquot of pools of DNA size fractions and serial dilutions of BamHI-digested DNA were analysed in parallel by gel electrophoresis to estimate the DNA recovery after size fractionation (Fig. 4.4C). The overall recovery was estimated at 30% to 40%.
Fig. 4.4. Size fractionation of deletion molecules. (A) The region analysed, as described for Fig. 3.1, plus BamHI cleavage sites used for size fractionation of genomic DNA and nested PCR primers (arrows) used to amplify deletion molecules. Ectopic exchange between homology blocks can give leftward and rightward deletions. SNP heterozygosities [white (haplotype A) and black (haplotype B) circles] make it possible to distinguish interchromosomal recombinants showing exchange of flanking markers (AB and BA) from intrachromosomal deletions (AA and BB). (B) Positions of progenitors, -α3.7 and -α4.2 deletion mutants in a preparative gel electrophoresis. M, λ DNA × HindIII. (C) Gel electrophoretic analysis of DNA recovery. Pools of fractionated DNA (P and 0.5P) were compared with known amounts of BamHI-digested DNA (100, 50, 25, 13 and 6% yield, corresponding to 180, 90, 45, 22.5 and 11.3 μg DNA). The intensity of ethidium bromide staining of P should be the same as that of the corresponding position of 100% BamHI-digested DNA if the recovery is 100%.
4.2.3.2 Recovery of deletion mutants and size validation

PCR of fractionated DNA did indeed reveal the presence of α-globin deletion molecules in both sperm and blood DNA. Deletion molecules in size-fractionated genomic DNA were recovered by nested PCR amplification (as described in Chapter 3) of multiple aliquots of each fraction, with each aliquot containing at most 1.2 deletion molecules and no more than 0.9 progenitor molecules as established from pilot experiments performed on each fraction to gain an initial estimate of deletion frequency. Secondary PCR products were analysed by agarose gel electrophoresis and visualised by staining with ethidium bromide. Nested PCR analysis revealed deletions in these fractions plus some remaining progenitor molecules (Fig. 4.5A). Analyses of both blood and sperm DNA from the two men showed that in total ~875 progenitor DNA molecules were amplified in those DNA fractions where deletion molecules should have been located. This survey analysed fractionated DNA derived from $3.12 \times 10^7$ amplifiable haploid genomes. Thus size fractionation had depleted >99.997% of progenitor DNA molecules. The full screening of blood and sperm DNA in both men involved the analysis of 3125 nested PCRs.

Comparison of the size ranges of each DNA fraction (Fig. 4.5B) with the frequency of deletion mutant molecules (Fig. 4.5C) showed that these mutants were almost completely restricted to the correct size fractions, providing strong evidence for their authenticity. Further validation of the mutant distribution used a control 10.0-kb BamHI DNA fragment from the MHC that was matched in size to $\alpha^{42}$ deletion mutants. Amplification of an 8.0-kb interval from this MHC fragment across all fractions showed a fragment distribution very similar to that of the $\alpha^{42}$ deletion molecules. Likewise, the $\alpha^{37}$ deletions showed a similar distribution but shifted to somewhat larger DNA fractions, as expected because these deletion molecules are 0.4 kb larger than $\alpha^{42}$ mutants (Fig. 4.5C). The correct distribution plus the evidence of extremely low progenitor contamination established the authenticity of >99% of detected deletion molecules and showed that they were not PCR artefacts arising from broken or intact progenitor DNA molecules.

DNA recovery could also be estimated by comparing band intensities of the PCR amplified MHC interval (8.0-kb) from known amounts of unfractionated DNA with those from the pooled fractions in a titration series. Furthermore, the number of amplifiable target molecules in recovered DNA across all fractions could also be estimated by PCR amplification of the same MHC fragment at the single-DNA-molecule level using Poisson
analysis. Yields so estimated for the four fractionations performed (blood and sperm DNA from man 1 and man 2) varied from 20% to 50%. These yields agreed well with the yields estimated from bulk DNA recovery (Fig. 4.4C), indicating little damage to DNA during fractionation.

![Diagram A](image1)

**Fig. 4.5.** Mutant distribution across DNA size fractions. (A) Examples of detection of deletion molecules. Aliquots of one of the fractions of blood DNA from man 1 (10.2- to 11.4-kb BamHI DNA fragments), each derived from $1.1 \times 10^5$ amplifiable haploid genomes and containing ~0.5 remaining 14.1-kb progenitor DNA molecules, were amplified by nested PCR, and products were analysed by agarose gel electrophoresis. M, X DNA x HindIII. Some reactions show PCR products derived from deletion mutants, some others show progenitor molecules and some show no PCR products. Differential amplification results in mutant molecules being substantially overamplified relative to the longer progenitor. (B) Gel electrophoresis of BamHI-digested sperm DNA fractions from man 2. The location of progenitors and deletions was estimated with reference to the DNA marker. (C) Cumulative frequencies of deletion molecules across size fractions of BamHI-digested sperm DNA from man 2, determined from a total of 90 $-\alpha^{3.7}$ and 19 $-\alpha^{4.2}$ deletion mutants recovered from $6.7 \times 10^6$ amplifiable haploid genomes. Size ranges covered by fractions are shown as grey bars. The control is a 10.0-kb genomic BamHI fragment matched in size to the $-\alpha^{4.2}$ deletion mutant.
4.2.4 Characterisation of deletion mutants

4.2.4.1 Restriction mapping of deletion mutants

Having recovered mutant DNA molecules, the next stage was to classify them as $-\alpha^{3,7}$ or $-\alpha^{4,2}$ mutants and to roughly map ectopic exchange points. All 2° PCR products derived from deletion mutants plus some progenitor molecules were therefore collected and reamplified in tertiary PCR to a saturated level (Fig. 4.6B). The ectopic exchange point location in each mutant was roughly localised by digesting mutant- and progenitor-amplified DNA with \(Pvull\) or \(Rsal\) and comparing electrophoretic profiles of restriction fragments (Fig. 4.6C, D). Two distinct profiles were seen in these mutants; the deletion length and exchange location in each mutant (Fig. 4.6A) were estimated by comparing the restriction profiles between progenitors and mutants, indicating the recovery of both $-\alpha^{3,7}$ and $-\alpha^{4,2}$ mutants.

---

**Fig. 4.6.** Restriction digestion for ectopic exchange point mapping. (A) The region analysed with coloured homology blocks, \(Pvull\) and \(Rsal\) cleavage sites. Ectopic exchange breakpoints (arrowed lines) of each mutant are roughly located by restriction mapping. (B) Tertiary PCR products of twelve mutants detected in one fraction of sperm DNA from man 2. M, \(\lambda\) DNA \(\times\) HindIII. (C) Electrophoretic profiles of \(Pvull\)-digested mutants (1–12) and the progenitor (P). (D) Electrophoretic profiles of \(Rsal\)-digested mutants and the progenitor.
4.2.4.2 Refining the mapping of ectopic exchange breakpoints

Restriction mapping revealed that almost all mutants were either $-\alpha^3$ or $-\alpha^4$ deletions with exchanges within the Z and X homology blocks, respectively. Ectopic exchange breakpoints were refined further by hybridisation with oligonucleotides specific to PSVs along the homology blocks (Fig. 4.7). SNPs were likewise typed by hybridisation of allele-specific oligonucleotides to these PCR products to characterise their haplotype of origin (details in section 4.2.6). A total of 34 PSVs (including 12 haplotype-specific PSVs created by SNPs within homology blocks), were typed and the exchange points were refined to intervals of perfect sequence identity shared by homology blocks that ranged in length from 34 bp to 1143 bp. Since the 3’ end regions of homology blocks were very divergent (Fig. 4.7), only few PSVs in these regions were typed by hybridisation. In total, 17 mutants with ectopic exchange breakpoints located in these regions were confirmed by sequencing. In addition, 131 assorted deletions were also sequenced to confirm exchange regions as well as any switching of PSVs near the ectopic exchange breakpoints.
Fig. 4.7. Refinement of ectopic exchange breakpoints of $-\alpha^{4.2}$ and $-\alpha^{3.7}$ deletions. (A) The region analysed with coloured homology blocks and SNPs (grey circles) in man 1. PSVs in the magnified X and Z homology blocks on both haplotypes are marked with red and blue bars, respectively. (B) Examples of $-\alpha^{4.2}$ and $-\alpha^{3.7}$ interchromosomal deletions arising by unequal exchange between homology blocks. The haplotype of origin and the ectopic exchange breakpoints were characterised by ASO hybridisation of flanking SNPs (white and black circles) and PSVs (red and blue bars) along the homology block, respectively. The deletions illustrated show a simple unequal exchange between haplotype B and haplotype A that mapped to the longest interval of sequence identity shared by the homology blocks.
4.2.4.3 Poisson correction of the frequency of deletion mutants

In some cases (10–14% of positive reactions), two different mutants were seen in the same PCR as shown by mixed PSV and/or SNP sites. In cases of a single mixed site, the constituent mutants could be deduced unambiguously. In other cases, sequence-specific PCR directed to a mixed PSV or SNP site was used to separate mutants prior to mutant characterisation. Only 8 out of 459 positive PCRs gave complex mixtures that could not be resolved; these mutant pools were excluded from further analyses.

To correct for instances of more than one molecule of a given type being present in a PCR, all positive PCRs were catalogued according to their haplotype(s) of origin and exchange breakpoints. A full inventory of PCRs that were positive or negative for each type of mutant was established for each size fraction and was used to estimate Poisson-corrected numbers of molecules. An example of such an inventory for blood mutants detected in man 2 is given in Table 4.2. These Poisson corrections were insignificant for most mutant classes and only 1.7-fold for the most abundant type of mutant.
Table 4.2. Poisson-corrected deletion mutants recovered from size fractions of blood DNA in man 2

<table>
<thead>
<tr>
<th>Mutant type</th>
<th>Size fraction</th>
<th>Volume of fraction per PCR (µl)</th>
<th>Total no. of PCRs</th>
<th>No. of -ve PCRs</th>
<th>No. of +ve PCRs</th>
<th>Poisson corrected no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AZ12</td>
<td>7</td>
<td>0.417</td>
<td>48</td>
<td>44</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.781</td>
<td>32</td>
<td>31</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AZ3</td>
<td>8</td>
<td>0.417</td>
<td>48</td>
<td>47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.267</td>
<td>30</td>
<td>28</td>
<td>2</td>
<td>10.6</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.292</td>
<td>24</td>
<td>23</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.417</td>
<td>48</td>
<td>41</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>AZ5</td>
<td>7</td>
<td>0.267</td>
<td>30</td>
<td>29</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.267</td>
<td>30</td>
<td>28</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.781</td>
<td>32</td>
<td>30</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.333</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.417</td>
<td>24</td>
<td>23</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0.5</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AZ6</td>
<td>8</td>
<td>0.417</td>
<td>48</td>
<td>47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AZ7</td>
<td>7</td>
<td>0.333</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AZ8</td>
<td>8</td>
<td>0.417</td>
<td>48</td>
<td>47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BX1</td>
<td>5</td>
<td>1</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.267</td>
<td>30</td>
<td>29</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.781</td>
<td>32</td>
<td>30</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>BZ12</td>
<td>8</td>
<td>0.267</td>
<td>30</td>
<td>29</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.267</td>
<td>30</td>
<td>28</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.417</td>
<td>48</td>
<td>47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BZ34</td>
<td>8</td>
<td>0.267</td>
<td>30</td>
<td>29</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.267</td>
<td>30</td>
<td>28</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.417</td>
<td>48</td>
<td>47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BZ5</td>
<td>8</td>
<td>0.267</td>
<td>30</td>
<td>29</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.267</td>
<td>30</td>
<td>28</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.417</td>
<td>48</td>
<td>47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.417</td>
<td>24</td>
<td>23</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.781</td>
<td>32</td>
<td>31</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>BZ7</td>
<td>8</td>
<td>0.267</td>
<td>30</td>
<td>29</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td>48</td>
<td>48</td>
<td></td>
</tr>
</tbody>
</table>

These mutants were recovered from $7.1\times10^6$ amplifiable haploid molecules, separated into 11 size fractions (1–11 in increasing size, 50 µl per fraction). Fractions 1–4 and 6 yielded no mutants and are not shown. The nomenclature of mutants is assigned according to their haplotypes of origin and exchange point locations. For example, AZ3 represents intrachromosomal mutants exchanged between/within haplotype(s) A at the third segment of uninterrupted identical sequence in the Z homology block (Fig. 4.7).
4.2.5 Deletion frequencies
Deletion mutants were analysed in blood and sperm DNA from two men who have been proved to carry neither a constitutional α-globin gene deletion nor duplication (section 4.2.2). Since the incidence of −α chromosomes in northern Europe is very low [<0.01 in some northern European populations (Flint et al. 1986b)], low deletion frequencies were therefore expected. However, substantial instability was seen in all samples (Table 4.3). Blood-deletion frequencies were very similar in both men, at $6.7 \times 10^{-6}$ and $6.8 \times 10^{-6}$ per haploid genome in man 1 and man 2, respectively. Sperm-deletion frequencies were significantly higher than in blood in both men ($P < 0.001$), with man 1 showing a 4-fold higher frequency than man 2 ($6.8 \times 10^{-5}$ vs. $1.6 \times 10^{-5}$ per sperm, $P < 0.001$).

Table 4.3. Ectopic deletion frequencies in sperm and blood in two men

<table>
<thead>
<tr>
<th></th>
<th>Ectopic deletion frequency ($\times 10^{-6}$ per haploid genome)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Man 1</td>
</tr>
<tr>
<td>Sperm</td>
<td></td>
</tr>
<tr>
<td>Interchromosomal</td>
<td>14.8</td>
</tr>
<tr>
<td>Intrachromosomal</td>
<td>52.9</td>
</tr>
<tr>
<td>Total</td>
<td>67.7</td>
</tr>
<tr>
<td>Blood</td>
<td></td>
</tr>
<tr>
<td>Intrachromosomal</td>
<td>6.7</td>
</tr>
</tbody>
</table>

4.2.6 Structures of deletions
All deletion mutants recovered from blood and sperm were characterised, as described in section 4.2.4, to locate ectopic exchange points. Deletion molecules were also analysed for SNP heterozygosities to determine whether exchanges were intrachromosomal (AA or BB type) or had occurred between homologous chromosomes (interchromosomal, AB or BA type) (Fig. 4.4A). Some of these SNPs mapped within homology blocks and created haplotype-specific PSVs (Fig. 4.7), allowing investigation of the effects of sequence mismatches on ectopic recombination.

4.2.6.1 Intra- and interchromosomal exchanges
Most deletions in blood and sperm were of the $-\alpha^{3.7}$ class, with only 4% showing the leftward $-\alpha^{4.2}$ deletion (Fig. 4.8). All of the 132 blood mutants were intrachromosomal, with none showing exchange of flanking SNP markers. These mutants were most likely products of unequal mitotic recombination, although a possible contribution from
replication slippage cannot be excluded. In contrast, 26% of sperm mutants (145 of 550) were interchromosomal and showed exchange of flanking markers (344 intra- vs. 96 interchromosomal exchanges in man 1, \( P < 0.001 \); 61 intra- vs. 49 interchromosomal exchanges in man 2, \( P = 0.253 \)) (Table 4.3), which indicates a clear difference in somatic and germ-line ectopic recombination pathways and suggests a significant role of homologue pairing and meiotic recombination in the generation of sperm deletions. The sperm specificity of these interchromosomal exchanges plus the size validation of mutants proved that these must be authentic and not jumping PCR artefacts derived from progenitor DNA molecules.

Fig. 4.8. Distribution of ectopic exchange points. Progenitor haplotypes in each man have heterozygous SNPs marked as white and black circles. Each interval of sequence identity containing exchange breakpoints is marked by a tied horizontal line, in blue for \(-\alpha^{4.2}\) deletions and black for \(-\alpha^{3.7}\) deletions, with the number of exchanges seen in sperm and blood DNA indicated in black and red, respectively, above the tie (\(-\), no mutants). Intra- and interchromosomal deletions are shown separately. Some SNPs are located within homology blocks and, depending on the allele, can break long regions of identity into smaller regions, creating differences between haplotypes in the intervals within which ectopic exchanges can be mapped. These mutants were recovered from \(6.5 \times 10^6\) and \(12.9 \times 10^6\) amplifiable haploid genomes from sperm and blood DNA, respectively, from man 1 and \(6.7 \times 10^6\) and \(7.1 \times 10^6\) molecules from man 2.
4.2.6.2 Simple and complex exchanges

Almost all deletions had arisen by simple exchange between homologous DNA sequences within intervals of sequence identity (Fig. 4.9Bi, Ci). Only four exceptional types were seen. Firstly, nine $-\alpha^{4.2}$ deletions seen in sperm from man 2 that mapped to the same interval all showed the same switching of PSVs near the region of exchange, indicating patchy gene conversion accompanying ectopic exchange. This gene conversion was most likely generated by mismatch repair of heteroduplex DNA after the resolution of Holliday junctions. These mutants were all intrachromosomal exchanges, and the switching of PSVs made the ectopic exchange breakpoints ambiguous (Fig. 4.9Bii). A similar type of conversion was seen in five $-\alpha^{3.7}$ deletions (Fig. 4.9Cii). These five deletion mutants were intrachromosomal exchanges, with switching of haplotype-specific PSVs in both men. Three of them were detected in blood from man 1 while two were sperm mutants from man 2.

The third type of exceptional deletion was a simple intrachromosomal $-\alpha^{4.2}$ deletion accompanied by a 33-bp microdeletion located ~1 kb away from the ectopic exchange point (Fig. 4.9Biii). Sequencing data indicated that the microdeletion was located in a distal *Alu* repeat at the end of the X1 homology block. The origin of this microdeletion is uncertain and the possibility of a PCR artefact cannot be excluded, though the artefact would have had to arise very early during PCR given that no trace of the progenitor in the 33-bp deletion was present in this mutant.

Finally, one blood mutant with a deletion of ~6.5 kb of DNA was found from man 1 (Fig. 4.9D). The deleted fragment extended from ~1.9 kb upstream of the $\psi\alpha 1$ gene to the end of exon 2 of the $\alpha 2$ gene. Deletion of the 5' flanking SNP meant that the haplotype of origin could not be identified. This deletion was apparently not the product of homologous recombination since no homologous sequences were found in the vicinity of the exchange points (Fig. 4.9D). Size validation revealed that this deletion was located correctly in the smallest DNA fraction in man 1, although again, the possibility of a PCR artefact cannot be excluded. The frequency of this deletion type might be underestimated since some mutants were likely to have been lost in smaller DNA fractions that were not collected.
Fig. 4.9. Structures of deletion mutants. (A) The region analysed, as described for Fig. 3.1. Homology blocks X and Z are indicated above. (B) Structure of $\alpha^{4.2}$ deletions, with PSVs marked as bars. (i) Example of a simple $\alpha^{4.2}$ deletion mapping to a single interval between PSVs. A total of 18 such exchanges were seen in both men. (ii) A complex $\alpha^{4.2}$ mutant with PSV switching (*) near the site of ectopic exchange. Nine such $\alpha^{4.2}$ deletions were identified in the sperm of man 2. (iii) A simple exchange accompanied by a microdeletion in a distal Alu repeat (red sequence lost) seen in one blood mutant. (C) Structure of $\alpha^{3.7}$ deletions. (i) A total of 649 such simple $\alpha^{3.7}$ deletions were seen in sperm and blood DNA. (ii) A total of five complex $\alpha^{3.7}$ mutants with PSV switching were seen in both men. (D) A mutant with $\sim6.5$-kb of DNA deletion including the $\psi\alpha1$ gene and the 5' end of the $\alpha2$ gene was detected in blood DNA from man 1. The deleted region is marked by a bracket. Sequences around the 5' breakpoint (top) and 3' breakpoint (bottom) are compared with the deletion mutant (middle), with sequence matches indicated by lines.
4.3 Discussion

4.3.1 Substantial somatic and germ-line deletion frequencies

Despite the extensive characterisation of α-globin gene rearrangements in thalassaemia (Embury et al. 1979; Embury et al. 1980; Higgs et al. 1984; Higgs et al. 1989) and the detailed studies of haematological profiles in α-thalassaemia patients (Dozy et al. 1979; Johnson et al. 1982; Bowden et al. 1987), the frequency of α-globin gene deletions has never been established. In addition, direct analyses of de novo −α frequencies have never been reported, since parents of the α-thalassaemic individuals in most pedigree studies were −α carriers (Dozy et al. 1979). The work described in this chapter, using individuals without α-globin rearrangements, revealed that de novo germ-line deletion frequencies were surprisingly high, and contrasted to the rarity of −α chromosomes in north European populations (Flint et al. 1986b). In addition, high deletion frequencies identified in blood DNA from two men indicated that ectopic recombination between local DNA repeats is not a germ-line-specific process. This observation contrasts strongly to allelic recombination at meiotic recombination hot spots defined by sperm typing in which no genuine mitotic recombinants have ever been seen. (Jeffreys et al. 2000; Jeffreys et al. 2001; Jeffreys and Neumann 2002; Jeffreys and May 2004; Kauppi et al. 2004; Jeffreys et al. 2005).

The consistently low incidence of −α chromosomes in non-malarial regions (Flint et al. 1986b; Yenchitsomanus et al. 1986; Hill 1992) and the absence of de novo deletion mutants identified in familial studies hinted that the occurrence of de novo α-globin deletions might be rare. However, the germ-line deletion rates obtained in this study (4.2 × 10⁻⁵ per sperm averaged over two men) suggested that the big contrast between the remarkable germ-line deletion rates and the rarity of α⁺-thalassaemia in north European populations is the result of strong selective forces against carriers of α-globin gene deletions (both −α/−α homozygotes and −α/αα heterozygotes) despite their lack of prominent haematological differences from normal individuals. This issue is discussed in more detail in Chapter 5.

Deletion frequency has also been characterised in the β-globin gene cluster. In contrast to the α-globin genes, the ectopic deletion rate in the δ- and β-globin gene region is extremely low (6.2 × 10⁻⁸ per sperm) (Holloway et al. 2006). Possible explanations for this significant difference between the two globin families are the higher sequence
similarity between homology blocks, and longer uninterrupted identical sequences within homology blocks in α-globin genes compared to the δ- and β-globin gene region (Fig. 4.10). These criteria appear to be components for triggering ectopic exchanges although other factors like nuclear folding (Nicholls et al. 1987) might influence the exchange frequency.

While significant somatic mutation frequencies have been found in the α-globin gene region, analyses of human minisatellite MS32 showed that the α-globin gene cluster is not the only region with bulk genome instability maintained by unequal crossover. Studies of human minisatellite MS32 revealed that these repeat arrays were unstable in somatic tissues with mutation frequencies as high as $10^{-5}$ per haploid genome (Jeffreys and Neumann 1997). However, the germ-line mutation frequency at MS32 is far higher (~1000 fold), in major contrast to the similar deletion frequencies seen for α-globin genes in somatic and germ-line DNA.

![Fig. 4.10. Homology blocks in α-globin genes, the δ- and β-globin genes.](image)

(A) Structures of the α-globin gene cluster (Left), with ψα1, α2 and α1 genes shown and of the δ- and β-globin genes (Right). Exon and intron are marked with filled and empty boxes, respectively. Homologous regions are highlighted with coloured blocks. (B) Uninterrupted identical sequences within homology blocks. PSVs on each pair of misaligned homology blocks are marked with red bars. The longest uninterrupted identical sequences within homology blocks are indicated by arrowed lines. Structures of homology blocks in α-globin genes and in the δ- and β-globin gene region are shown in different scales.
4.3.2 Distinct mechanisms in soma and germ-line

4.3.2.1 Mitotic recombination frequency altered by mutational mosaicism in soma

Somatic deletions arise by what appeared to be a strictly intrachromosomal pathway of homologous exchange; no interchromosomal exchanges were observed in the survey of 132 blood deletions (Fisher’s exact test, $P > 0.95$). These mutants were most likely generated by mitotic recombination in which exchanges occurred only between/within sister chromatids. These ectopic exchanges were expected to occur on both haplotypes at similar frequencies (AA-type = BB-type). However, significant haplotype distortions were observed. For instance, the most common type of $-\alpha^{3.7}$ deletion in blood from man 1 showed an excess on haplotype B (42 mutants vs. 18 on haplotype A, $P = 0.002$). Similarly, somatic $-\alpha^{3.7}$ deletions exchanged in the first two uninterrupted identical regions of the Z homology blocks in man 1 also revealed haplotype asymmetry (16 BB-type mutants vs. 4 AA-type mutants, $P = 0.007$). These fluctuations most likely arise from mutational mosaicism. The level of this mosaicism detected in deletion mutants depends not only on the mutation frequency per cell division but also on the timing of mutation, with significant haplotype distortions arising if the mutation appeared early in development.

4.3.2.2 Mechanisms in germ-line

Contrasting to the strictly intrachromosomal pathway operating in the soma, recombinational interactions between homologous chromosomes were frequently found in sperm deletions, with 26% of deletions showing flanking marker exchange. Interchromosomal exchanges, which generate either AB- or BA-type recombinant, showed distinct haplotype symmetry. For example, 96 interchromosomal $-\alpha^{3.7}$ deletions distributed equally on both haplotypes in sperm from man 1 (48 AB-type vs. 48 BA-type, $P = 1$), while frequencies of interchromosomal sperm exchanges from man 2 on both haplotypes were very similar (27 AB-type vs. 22 BA-type, $P = 0.475$). Significant haplotype symmetry of these interchromosomal sperm exchanges indicated that they were most likely unique recombinational products generated at meiosis, contrasting to examples of haplotype distortions identified in intrachromosomal exchanges (section 4.3.2.1). These findings provide clear evidence that ectopic recombination between locally repeated DNA sequences can arise by two different pathways, each of which contributes a significant level of genome instability.
Despite the significant proportion of interchromosomal exchanges detected in the germ-line, intrachromosomal exchange was the predominant exchange type in sperm DNA from both men (Fig. 4.8). These intrachromosomal exchanges could be pre-meiotic recombinants or meiotic exchanges between sister chromatids; these would be indistinguishable in this study. If intrachromosomal sperm exchanges were solely generated by meiotic unequal exchange between sister chromatids, haplotype asymmetry would not be expected. However, similar to what was observed in blood mutants, some significant haplotype distortions were seen. Fluctuations were seen in sperm from man 2 for the most common class of \(-\alpha^{3,7}\) deletion, all nine of which were derived by intrachromosomal exchange on haplotype B and shared the same unusual conversion event (Fig. 4.8). This germ-line mosaicism strongly suggests a pre-meiotic intrachromosomal component to sperm deletion, with in this case one sperm in 700,000 sharing the same pre-meiotic mutation that must have arisen in a single progenitor cell. Similar levels of sperm mosaicism have been seen for minisatellite deletion mutants (Jeffreys et al. 1990) and can distort estimates of mutation rate. A potentially similar observation was seen in the unusually frequent \(-\alpha^{3,7}\) intrachromosomal deletions found on both haplotypes in the longest region of sequence identity in sperm from man 1 but not man 2 (Fig. 4.8). Since there were no DNA sequence differences between the two men in this region that could account for this frequency discrepancy, the most reasonable explanation is that pre-meiotic mosaicism has inflated these frequencies in man 1 and, by chance, to the same extent on both haplotypes. Together, these examples of frequency fluctuation and inflation indicate that mutational mosaicism might contribute substantially to the variation in sperm deletion frequencies between man 1 and man 2.

This pattern of germ-line deletion via a predominantly intrachromosomal pathway contrasts with sex-dependent mechanisms seen in the study of \textit{de novo} CMT1A and HNPP (LeGuern et al. 1996; Lopes et al. 1997; Lopes et al. 1998). Analyses of CMT1A/HNPP patients revealed that the rearrangements of paternal origin, particularly for CMT1A, were generated by interchromosomal unequal exchanges while both CMT1A and HNPP of maternal origin were found to be the products of intrachromosomal exchanges. Sex-dependent mechanisms were further supported by the exchange point distribution of deletions and duplications; most paternal duplications mapped to a 741-bp interval defining a hot spot for unequal exchange, in contrast to maternal rearrangements which exchanged outside the interval (Lopes et al. 1998). These examples demonstrate distinct differences between processes of ectopic exchange between locally repeated
DNA sequences in α-globin genes and recombination between distant repeat elements such as CMT1A/HNPP ectopic exchanges. However, since very few examples have been seen to date, more examples of both are needed for studying generality.

4.3.3 Random distribution of exchange points
The bulk of ectopic exchanges mapped as expected to the longest regions of identical sequence shared between homology blocks (Fig. 4.8). However, exchange distributions showed that ectopic exchange points, although avoiding regions of high divergence, were distributed fairly randomly, with the cumulative frequency of exchanges rising fairly uniformly across homology blocks (Fig. 4.11). The only exceptions were α3.7 intrachromosomal deletions in sperm from man 1, which were unusually frequent on both haplotypes in the longest region of sequence identity and largely responsible for the high overall sperm-deletion frequency seen in this man compared with man 2 (Table 4.3).

CMT1A/HNPP studies on patient samples revealed that exchanges mapped preferentially in a ratio of approximate 53:1 to a 1.7-kb region within the surrounding ~28 kb of homologous region (Reiter et al. 1996). A mariner transposon-like element (MITE), which was proposed to trigger strand exchange events, has been located near the 1.7-kb recombination hot spot (Reiter et al. 1996). This identification suggested that biased exchange location of ectopic recombinants in CMT1A/HNPP might be promoted by surrounding sequences. In contrast, despite the fact that mapping ectopic recombination resolution points cannot exclude the possibility of a localised recombination-initiating region, no local exchange hot spots within homology regions in the α-globin gene cluster were identified. This provides further evidence for a major difference of the processes of ectopic exchange between local DNA repeats and recombination between distant repeat elements.
Fig. 4.11. Distribution of ectopic exchange breakpoints across homology blocks. (A) The $\alpha$-globin gene cluster as described in Fig. 3.1. (B) The X (Left) and Z (Right) homology regions analysed, with PSVs marked with bars. (C) Cumulative number of ectopic exchanges, per $10^6$ haploid genomes, across homology blocks for all $-\alpha^42$ exchanges combined and for each type of $-\alpha^37$ exchange as defined in the table at lower left, with intrachromosomal sperm exchanges in man 1 shown separately. The interval with a comparatively low ectopic exchange frequency per base pair is marked with an asterisk (described in Fig. 4.12).
4.3.4 A challenge to minimum homology requirement

Extensive studies of homologous mitotic recombination and unequal crossover have both been done in vitro and in vivo in bacteria and mammalian somatic cells (Watt et al. 1985; Ayares et al. 1986; Metzenberg et al. 1991). Studies focussed on recombination between plasmid and chromosomal DNA indicated that minimum amount of homology (minimal efficient processing segment, MEPS) seems to be necessary for efficient recombination. For instance, the recombination frequency between the πVb plasmid and the λ bacteriophage decreased dramatically with reduced length of perfect homology (<74 bp) (Watt et al. 1985). Likewise, the efficiency of intrachromosomal recombination in mammalian cells dropped sharply when the amount of uninterrupted homology decreased to ~100–200 bp (Rubnitz and Subramani 1984; Liskay et al. 1987; Waldman and Liskay 1988). These observations provided evidence that extensive regions of sequence identity between homologous DNA sequences are needed for efficient recombination. Furthermore, studies using mammalian cells indicated that the length of uninterrupted sequence identity rather than the degree of mismatches within a homology block determines the efficiency of recombination (Liskay et al. 1987; Waldman and Liskay 1988).

However, although most ectopic exchanges in the α-globin genes occurred at the longest regions of uninterrupted region within homology blocks (Fig. 4.11), ectopic exchange frequencies per base pair in each interval of sequence identity showed only a weak dependence on uninterrupted homology length in the α-globin gene cluster (Fig. 4.12). Ectopic exchange occurred in intervals as short as 34 bp and the general exchange frequency was found to be ~8 × 10^-9 per base pair even for the longest intervals. Therefore, there is no evidence that ectopic exchanges are inhibited by modest sequence divergence between homology blocks. This lack of inhibition is also demonstrated by haplotype-specific SNPs that can act as PSVs and alter homology block lengths. For instance, man 1 shows an 1143-bp interval of sequence identity for AB-type exchanges, whereas two SNPs divide this interval into three for BA-type exchanges (Fig. 4.8). Despite this disruption, the same numbers of exchanges mapped to this region in AB and BA recombinants (42 in each case), with BA exchanges distributed across the three intervals in proportion to their length (observed vs. expected exchanges: 6, 2, and 34 vs. 8, 2, and 32 in intervals 220, 66, and 855 bp long, respectively, P = 0.5). These observations challenge the notion of minimum pairing segments of uninterrupted homology required.
for homologous exchanges (Gonda and Radding 1983; Rubnitz and Subramani 1984; Waldman and Liskay 1988).

**Fig. 4.12.** Ectopic exchange frequencies per base pair estimated for each interval of sequence identity, with the best-fit logarithmic curve after excluding the outlying points (arrows). The outlying points are exchanges with exceptionally high frequencies, which are most likely triggered by mutational mosaicism. The circled points with a low frequency correspond to the interval marked by an asterisk in Fig. 4.11.

### 4.3.5 Sequence identity is not the sole factor controlling ectopic recombination

The frequencies of various types of $\alpha$ deletion in $\alpha$-thalassaemia patients correlate with the length of identical sequences shared between homology blocks, for example $-\alpha^{3.7I}$ (1436 bp) $>-\alpha^{4.2}$ (1339 bp) $>-\alpha^{3.7II}$ (171 bp) $>-\alpha^{3.7III}$ (46 bp) (Fig. 1.10) (Higgs et al. 1984). The present study revealed that ectopic recombination frequencies, however, are not governed solely by lengths of high sequence similarity shared between homology blocks. Thus, a 166-bp interval at the 3' end of the Z2/Z1 homology block consistently showed a low ectopic exchange frequency (6 observed vs. 44 expected mutants, $P < 0.001$) (Fig. 4.11C). Similarly, the X2/X1 homology block contains an 800-bp-long region of strong sequence similarity, compared with 1700 bp for the Z2/Z1 homology block. If ectopic recombination frequencies were determined purely by homology length, then 32% of ectopic exchanges should generate $-\alpha^{4.2}$ deletions, compared with 4% observed ($P < 0.001$). The rarity of $-\alpha^{4.2}$ chromosomes (clustered mostly in Southeast Asia and the Middle East, Higgs and Weatherall 1983) compared with high frequencies of $-\alpha^{3.7}$ chromosomes in most malarial populations suggested that factors other than sequence identity between homology blocks may control the prevalence of $-\alpha^{3.7}$ and $-\alpha^{4.2}$ deletions.
in populations (Higgs et al. 1984; Flint et al. 1998). However, whether this discrepancy caused by mechanistic differences of unequal exchange or selection is still unknown.

It is possible that recombination initiation events are, in fact, preferentially targeted to the longest regions of sequence identity, particularly within the Z2/Z1 homology block, and that subsequent migration of Holliday junctions leads to substantial diffusion of exchange resolution points. This diffusion, however, would have to be greater than that seen at hot spots for allelic recombination, given the narrow width of the latter (Jeffreys et al. 2001). Instead, it seems likely that other factors, such as features of nuclear architecture (Vanin et al. 1983; Nicholls et al. 1987), play an important role in controlling the frequency of recombinational interactions between different paralogous sequences. For example, the clustering of 5' and 3' breakpoints within a 10-kb region in different types of a°-thalassaemia deletions may indicate that those regions are preferred recombination target sites or that the sequences are brought together by nuclear topology (Nicholls et al. 1987). Similar suggestions were raised in studies of the β-globin gene in which the folding of chromatin and the location of nuclear matrix attachment points may play a significant role in controlling which homologous sequences can interact efficiently (Vanin et al. 1983; Collins et al. 1987; Jarman and Higgs 1988).

### 4.4 Conclusions

The work presented in this chapter demonstrated that it is possible to use single-DNA-molecule analysis to recover \textit{de novo} α-globin gene deletion mutants and to gain direct insights into the dynamics and processes of ectopic recombination that operate within human gene families (Potocki et al. 2000). This study revealed that \textit{de novo} deletion frequencies of single α-globin gene deletions were surprisingly high in both sperm and blood. Distinct recombination pathways have been revealed in this study. Somatic deletions arise by a strictly intrachromosomal pathway of homologous exchange that also operates predominantly in the germ-line and can generate mutational mosaicism which significantly influences deletion rates between individuals. A second pathway generating some sperm deletions involves recombinational interactions between homologous chromosomes and most likely occurs at meiosis. Despite avoiding high sequence divergences along homology blocks, ectopic exchange breakpoints distribute fairly randomly with no evidence of local ectopic recombination hot spots. Ectopic exchange also shows surprisingly little dependence on uninterrupted identical homology,
challenging the notion of minimum pairing requirements suggested in the studies of mammalian cells (Gonda and Radding 1983; Rubnitz and Subramani 1984; Waldman and Liskay 1988). Finally, the relationship between homology block lengths and the prevalence of various deletion types suggested that other factors besides local homology play an important role in controlling the frequency of recombinational interactions between different paralogous sequences in ectopic recombination.
Chapter 5
Deletion Frequencies and Population Fitness

5.1 Introduction
While many hypotheses have been proposed to explain why haemoglobinopathies have reached high frequencies in some populations but not in others, it is believed that malaria selection is the major force for controlling this differential geographical distribution. Although malaria selection cannot account for the distribution of all haemoglobinopathies, this hypothesis seems to fit well for most of these monogenic disorders and particularly for α- and β-thalassaemias. (Flint et al. 1986a; Hill et al. 1988; Flint et al. 1998). For example, β-thalassaemia is prevalent in most malarial regions like the Mediterranean, North Africa and the Southeast Asia (Flint et al. 1998). Only limited numbers of β-thalassaemia haplotypes have been found in these malarious regions, while a diversity of β-thalassaemia mutations has been identified in some relatively malaria-free areas where β-thalassaemia is rare. This contrast implies that the β-haplotypes were positively selected in malarial regions whereas those mutations in malaria-free areas have arisen independently and are of recent origin (Thein 1993). Furthermore, despite evidence against the ‘malarial hypothesis’ from in vitro experiments (Friedman 1979; Ifediba et al. 1985; Modiano et al. 1991), population studies revealed that the prevalence of α-thalassaemia is selected by this parasitic disease, showing for example a remarkable altitude- and latitude-dependent correlation with malaria endemicity in Melanesia (Flint et al. 1986a).

Clinical studies suggested that the red blood cell indices including MCH and MCV of −α/αα heterozygotes are slightly but not significantly lower than those of normal individuals (Higgs 1993). Despite the drop of these haematological parameters, most α*-thalassaemias are phenotypically indistinguishable from normal individuals. In addition to the advantage of malaria selection, it is not difficult to explain why α-thalassaemia is so worldwide and particularly prevalent in malarious regions (Flint et al. 1998).

The previous chapter showed that de novo deletion frequencies in sperm are substantial at $4.2 \times 10^{-5}$ per sperm averaged from two men. Since the process of −α deletion from the normal αα chromosome is irreversible (ignoring exchanges between ααα and −α chromosomes), the population frequency of αα chromosomes would have
been substantially reduced (<0.8) before the time of modern human evolution with these remarkable deletion frequencies (A. Jeffreys, personal communication). However, population studies in non-malarial regions have recurrently shown that \(-\alpha\) chromosomes are very rare in populations like Japanese (Nakashima et al. 1990), Koreans (Shimizu et al. 1989), Icelandics and Britons (Flint et al. 1986b). This rare incidence of \(-\alpha\) chromosomes is almost impossible to be maintained unless significant selection operates in these populations. In an attempt to quantify the strength of selection, systematic analyses of the fitness of normal individuals, \(-\alpha\) carriers and \(-\alpha/-\alpha\) homozygotes in relation to the incidence of \(-\alpha\) chromosomes in the absence of malaria selection were performed. The work in this chapter provides evidence of significant selection on the fitness of \(-\alpha/-\alpha\) homozygotes and/or \(-\alpha\) carriers, revealing completely new insights of historical levels of selection operating in human populations.

5.2 Results

5.2.1 Screening of \(-\alpha\) chromosomes

To compare the \textit{de novo} \(-\alpha\) deletion frequency with the incidence of \(-\alpha\) chromosomes, a survey of the frequency of \(-\alpha\) chromosomes in malaria-free areas was necessary. Previously reported data on \(-\alpha\) deletion frequencies in non-malarial populations from other studies were used (Flint et al. 1986b; Higgs et al. 1989; Hill 1992). In addition, the incidence of \(-\alpha\) chromosomes was analysed by PCR amplification. Using the optimised PCR conditions as described in Chapter 4, progenitor and \(-\alpha\) deletion DNA molecules were readily distinguished (Fig. 5.1). A total of 476 DNA donors from three different panels were surveyed. Of all individuals tested, only those reported as being of northern European descent, who were from relatively malaria-free areas, were further analysed. In total, 822 chromosomes were screened and only 18 of them were \(-\alpha\) chromosomes. Details are described in the following sections.
Fig. 5.1. Screening of \(-\alpha\) chromosomes in populations. (A) Possible sizes of the \(\alpha\)-globin gene cluster detected in PCR amplification. (B) Examples of detection of \(-\alpha\) chromosomes. PCR products from ten individuals who were \(\alpha_a/\alpha_a\) homozygotes, plus two heterozygous \(\alpha_a/\alpha\alpha\) carriers detected from the LRI panel were analysed by agarose gel electrophoresis and visualised by staining with ethidium bromide. M, \(\lambda\) DNA \(\times\) HindIII. PCR amplification is in favour of shorter molecules in heterozygotes, giving efficient amplification of \(-\alpha\) chromosomes but not \(\alpha\alpha\) chromosomes.

5.2.1.1 Samples from the Leicester Royal Infirmary panel

\(-\alpha\) chromosomes were screened from a total of 170 sperm DNA samples from semen supplied by the Assisted Conception Unit, Leicester Royal Infirmary (LRI). In this LRI panel, 132 men are self-reported as being of northern European descent, 25 men are of Indian origin and two men are Afro-Caribbean. A total of eight \(-\alpha\) deletion chromosomes were detected; however, six of them were from Indians while the other two were of Afro-Caribbean origin. In total, no \(-\alpha\) deletions were detected from men of northern European descent in this LRI panel.

5.2.1.2 Samples from Centre d'Etude du Polymorphisme Humain (CEPH) panel

DNA samples for \(-\alpha\) chromosome screening were selected from 121 unrelated individuals of Utah, Amish, or French descent from 39 different CEPH family pedigrees. Most family pedigrees were of three generations and priority was given to grandparent samples, or if unavailable, parent samples. A total of 97 samples were unambiguously scored after PCR amplification and only three of them were \(-\alpha/\alpha\alpha\) carriers.
5.2.1.3 **Samples from M. Jobling's panel**
Genomic DNA extracted from buccal cells from 182 unrelated men in the United Kingdom identified by grandparental place of birth as northern European in origin were provided by M. Jobling and T. King, Department of Genetics, University of Leicester, UK. A total of 364 chromosomes were screened and six of them were \(-\alpha\) chromosomes.

5.2.2 **Characterisation of \(-\alpha\) chromosomes**
All 2° PCR products of detected \(-\alpha\) chromosomes were collected and reamplified in tertiary PCR for characterisation by restriction digestion as described in Chapter 4, section 4.2.4.1. Electrophoretic profiles of \(PvuII\) digests revealed that almost all deletions were of the \(-\alpha^{3,7}\) class, with only one \(-\alpha^{4,2}\) deletion identified in the CEPH panel (data not shown).

5.2.3 **Comparison of population studies and \(-\alpha\) chromosome screening**
To obtain a more accurate estimate of the incidence of \(-\alpha\) chromosomes in non-malarial areas, data from reported population studies and from this \(-\alpha\) chromosome screening were compared to see if there was any significant discrepancy. Population studies have reported that one \(-\alpha/\alpha\alpha\) heterozygote was identified in a survey of 155 north Europeans (Beutler and West 2005), whereas no \(-\alpha\) chromosomes in 140 individuals from the United Kingdom and 110 Icelandics were detected (Flint et al. 1986b). Results of \(-\alpha\) chromosome screenings were then compared with those \(-\alpha\) frequencies reported in population studies (Table 5.1). There is no evidence for significant heterogeneity in carrier frequency across these samples \((P = 0.15 \text{ by computer estimation})\), and therefore all were combined to obtain a mean incidence of \(-\alpha\) chromosomes of 0.006 (95% confidence interval, 0.003–0.011). Given that malaria might have existed in some of these populations such as the reported malarious illness in mediaeval time in southeast England (Kuhn et al. 2003; Chin and Welsby 2004), this estimate of the incidence of \(-\alpha\) chromosomes in non-malarial regions is maximal and could be lower if any of the surveyed individuals have partial ancestry from regions with a high frequency of \(-\alpha\) chromosomes.
### Table 5.1. Incidence of \(-\alpha\) chromosomes in \(-\alpha\) deletion screenings and in reported studies

<table>
<thead>
<tr>
<th>(-\alpha) chromosome screening</th>
<th>Reported (-\alpha) chromosome survey</th>
<th>No. of chromosomes screened</th>
<th>No. of (-\alpha) chromosomes detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRI panel</td>
<td></td>
<td>264</td>
<td>0</td>
</tr>
<tr>
<td>CEPH panel</td>
<td></td>
<td>194</td>
<td>3</td>
</tr>
<tr>
<td>M. Jobling's panel</td>
<td></td>
<td>364</td>
<td>6</td>
</tr>
<tr>
<td>155 white subjects</td>
<td>(Beutler and West 2005)</td>
<td>310</td>
<td>1</td>
</tr>
<tr>
<td>140 Britons and 110 Icelandic</td>
<td>(Flint et al. 1986b)</td>
<td>500</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1632</td>
<td>10</td>
</tr>
</tbody>
</table>

Dashes indicate not applicable.

### 5.3 Discussion

#### 5.3.1 Incidence of \(-\alpha\) chromosomes and the effect of fitness

Chromosomes bearing \(\alpha\)-globin gene deletions are very common particularly in Asia and Africa, where the population incidence can be as high as 90% (Dozy et al. 1979; Embury et al. 1979; Hill 1992), and are most likely selected by malaria in favour of \(-\alpha\) chromosomes (Flint et al. 1986a; Hill 1992). However, a much lower incidence of deletion chromosomes (~0.6%) was seen in non-malarial areas. Given that the frequency of ectopic exchanges between misaligned \(\alpha\)-globin genes in sperm, primarily generating \(-\alpha^{3,7}\) deletions, is high, at \(4.2 \times 10^{-5}\) per sperm averaged over the two men tested, the low incidence of \(-\alpha\) chromosomes in malaria-free areas such as northern Europe is most likely constrained by some selective force. To quantify the effect of selection, a selection model is proposed, based on three assumptions. Firstly, since the frequency of ectopic exchange in \(\alpha\)-globin genes was only characterised in the male germ-line, an equal deletion rate in the female germ-line is assumed. Secondly, in addition to the absence of malaria selection in the samples studied, no other selection in favour of \(-\alpha\) chromosomes is present. Finally, it is assumed that \(-\alpha\) deletions are partially or fully recessive, in which
case selection will act more against $-\alpha/-\alpha$ homozygotes than against $-\alpha/\alpha\alpha$ heterozygotes.

The incidence of deletion chromosomes was compared with population frequencies expected for a given germ-line deletion rate, $\mu$. Consider a population with $-\alpha$ chromosomes at frequency $q$ and with normal $\alpha\alpha/\alpha\alpha$ individuals showing a fitness $f_n$ of 1, $-\alpha/\alpha\alpha$ carriers a fitness of $f_c$, and $-\alpha/-\alpha$ homozygotes a fitness of $f_i$. At equilibrium, the gain of new $-\alpha$ deletions from normal individuals and carriers is:

$$[(1 - q)^2 \mu + f_c q(1 - q)\mu].$$

This will be balanced by the loss of deletions from carriers and homozygotes:

$$[(1 - f_c)q(1 - q) + (1 - f_i)q^2].$$

If selection only operates on homozygotes ($f_c = 1, f_i < 1$), then this equilibrium frequency at $q \ll 1$ simplifies to the standard form,

$$q = \sqrt{\mu(1 - f_i)}.$$

In this model, the low incidence of $-\alpha$ chromosomes in northern Europe can only be maintained for a fully recessive deletion by significant selection against $-\alpha/-\alpha$ homozygotes (fitness $< 0.66$) (Fig. 5.2) despite their showing only a mild and sometimes asymptomatic form of $\alpha^+$-thalassaemia. If their fitness were higher, then the population incidence of $-\alpha$ chromosomes would exceed 1.1%, the upper 95% confidence limit of the northern European frequency. If there is a drop of fitness of $-\alpha/\alpha\alpha$ carriers, who can also show subtle haematological changes (Higgs 1993; Wang 2000), then the strength of selection operating against homozygotes could be weaker (Fig. 5.2). However, selection against carriers must be weak (fitness $> 0.985$), irrespective of the fitness of homozygotes, otherwise the incidence of $-\alpha$ chromosomes would fall below 0.3%, the lower 95% confidence interval of observed frequency. In malarial populations, the observed high incidence of $-\alpha$ chromosomes can be maintained by modest balancing selection against normal $\alpha\alpha/\alpha\alpha$ individuals; for example, a drop of fitness of only $1/10^{th}$ relative to $-\alpha/-\alpha$ homozygotes would be sufficient to maintain the incidence of deletion chromosomes in a
population at 10%, with an equilibrium frequency that is largely independent of deletion rate.

If \(-\alpha\) deletion frequencies between the female and male germ-line were very different, for example, instability in the female germ-line was ten-fold lower or higher than the deletion frequency observed in males, there would still need to be significant selection against \(-\alpha/\alpha\) homozygotes (Fig. 5.3). With a negligible deletion frequency in females, significant selection against homozygotes would still be needed but at a weaker level (fitness = 0.81, if fitness of \(-\alpha/\alpha\) carriers is 1) (Fig. 5.3A). However, if the instability were ten-fold higher in females, the fitness of homozygotes and heterozygotes would have to be substantially reduced with even stronger selection against homozygotes (Fig. 5.3B).

![Graph showing the effect of fitness on the incidence of \(-\alpha\) deletion chromosomes in non-malarial populations.](image-url)

**Fig. 5.2.** The effect of fitness on the incidence of \(-\alpha\) deletion chromosomes in non-malarial populations. The expected population incidence of deletions at mutation/selection equilibrium was estimated from the observed germ-line deletion frequency of \(4.2 \times 10^{-5}\) per sperm, averaged over the two men analysed and assumed to be the same in the female germ-line. Frequencies were estimated at various fitness levels of \(-\alpha/\alpha\) homozygotes and with different fitnesses of \(-\alpha/\alpha\alpha\) carriers. The dashed line provides the estimated incidence of \(-\alpha\) chromosomes in north Europeans, with 95% confidence intervals indicated in grey.
Fig. 5.3. The effect of fitness with contrasting deletion frequencies between males and females in non-malarial populations described in Fig. 5.2, but with the fitness relations of \(-\alpha/-\alpha\) homozygotes and \(-\alpha/\alpha\) carriers \((f_c)\) estimated when (A) female deletion frequencies were 10-fold lower than in males or (B) female deletion frequencies were 10-fold higher than in males.

5.3.2 Possible selective forces
The remarkable disparity between the germ-line deletion frequencies and the population incidence of \(-\alpha\) chromosomes indicates that significant selection must have acted against \(-\alpha/-\alpha\) homozygotes and/or \(-\alpha/\alpha\alpha\) heterozygotes. However, the nature of the selective forces operating against these largely normal individuals remains unknown. Speculations on any possible selective forces can be divided into two major areas, physical fitness and reproductive fitness.

Haematologic analyses, for instance, measurements on MCH, MCV and \(\alpha/\beta\) chain synthesis ratio, have revealed that the haematological profiles of silent carriers \((-\alpha/\alpha\alpha)\)
are not significantly different from normal individuals and are considered to have no haematological effects. In contrast, there were significant though subtle reductions of both MCH and MCV values in mild α-thalassaemics (−α/−α or −/−αα) although their range overlaps that of normal individuals (Johnson et al. 1982; Higgs 1993; Wang 2000). In some cases, these subtle changes could give iron-deficiency anaemia, microcytic red cells in peripheral circulation and 5–6% of Hb Bart's in the cord blood (Wills and Londo 1981). These subtle haematological shifts between α-thalassaemics and normal individuals might possibly provide some selective force for maintaining the low incidence of −α chromosomes.

Clinical studies have shown that the deficiency of α-globin chains in the early stage of gestation caused by some severe forms of α⁺-thalassaemia could impair foetal oxygenation and give rise to miscarriage (Higgs et al. 1989). Similarly, adverse effects associated with malaria during pregnancy, for instance, low birthweight caused by intrauterine growth retardation and preterm delivery, were also reported (Brabin et al. 1990; Allen et al. 1998). However, whether maternal α⁺-thalassaemia favours pregnancy by providing protection against malaria or damages pregnancy with reduced reproductive fitness is still unclear. Recently, a report on the studies of α⁺-thalassaemia and pregnancy in a malaria hyperendemic region of Papua New Guinea has indicated that maternal α⁺-thalassaemia does not affect susceptibility to malaria or reproductive fitness during pregnancy (O'Donnell et al. 2006). This study measured several markers of reproductive fitness including the number of miscarriages and stillbirths, Plasmodium falciparum infection of the mother plus placenta, maternal haemoglobin, preterm delivery and birthweight, and showed that α⁺-thalassaemia does not protect against malaria during pregnancy nor obviously affect fertility, despite the association with mild anaemia during pregnancy. These findings imply that reproductive fitness of maternal α⁺-thalassaemics is unlikely to be the major selective force for explaining the disparity between the remarkable −α deletion frequencies in the germ-line and the low population incidence of −α chromosomes.

Although the women recruited in this recent study were claimed to be representative of those living in the community (O'Donnell et al. 2006), there was still the issue that women who had not conceived were by definition excluded from the study. This raises the question of whether α⁺-thalassaemia might lead to infertility in some α⁺-thalassaemics, both men and women. Previous reports have shown that iron overload after frequent blood transfusions or chemotherapy after bone marrow transplantation can lead
to reproductive failure in thalassaemia major patients (De Sanctis et al. 1991; Perera et al. 2002). Despite these studies revealing the potential effects of major thalassaemia on fertility, the issue of reduced fertility in mild thalassaemias is still open given lack of data.

5.4 Conclusions
The work described in this chapter has shown that direct comparisons of the germ-line mutation frequencies and population frequencies can provide insights into historical levels of selection operating in human populations. The model suggested that the remarkable contrast between high germ-line deletion frequencies and the rarity of $-\alpha$ chromosomes is maintained by strong selection operating against $-\alpha/-\alpha$ homozygotes, and potentially on $-\alpha/\alpha\alpha$ heterozygotes too. However, the factors governing these selective forces remain completely unknown, since there is no significant phenotypic difference, in respect to haematological profiles and fertility, between mild $\alpha$-thalassaemics and normal individuals.
Chapter 6
Processes of De Novo \textit{\alpha\alpha\alpha} Duplication

6.1 Introduction

Despite the importance of duplicated DNA sequences in the creation of new genes and the generation of copy number variation and genetic disorders (Lupski and Stankiewicz 2005; Feuk et al. 2006; Lee and Lupski 2006), very little is known about the dynamics of DNA duplication and the reciprocity or otherwise of unequal exchange in human DNA. Examples do exist of reciprocal duplications and deletions detected in patients and in human populations (Chance and Fischbeck 1994; Edelmann et al. 1999; Potocki et al. 2000). Since factors like selection can influence the population frequencies of these exchanges, the dynamics and reciprocity of the processes that generate these rearrangements cannot be investigated directly from population studies. Ectopic recombination between misaligned homologous sequences on different DNA molecules, whether sister chromatids or homologous chromosomes, should generate both deletions and duplications as reciprocal exchange products. However, other processes can produce deletions including single-strand annealing during repair of a broken DNA molecule (Fishman-Lobell et al. 1992), as well as recombination within a single molecule to produce deletions plus excised extrachromosomal DNA circles (Stankiewicz and Lupski 2002), known to exist in mammalian cells (Stanfield and Helinski 1986). In addition, gene duplications can also arise in a non-reciprocal fashion by gene conversion events between chromatids or homologous chromosomes that leave the donor chromosome unaltered (Nassif et al. 1994).

The best-studied example of ectopic recombination between distant DNA repeats, in the disorders CMT1A and HNPP, revealed that unequal crossover breakpoints map to the same narrow recombination hot spot within the repeats in patients with \textit{de novo} CMT1A duplications and in individuals with HNPP deletions (Reiter et al. 1996; Reiter et al. 1998), consistent with reciprocal exchange (Chance et al. 1994). However, there is evidence of sex-dependent pathways of ectopic exchange (Lopes et al. 1998), and it remains unknown whether ectopic recombination in CMT1A/HNPP is a fully reciprocal exchange between different DNA molecules that produces duplications and deletions exclusively and at equal frequency.

Ectopic exchanges can also occur between locally duplicated DNA sequences separated by kilobases rather than megabases; such exchanges play a major role in the
evolution of gene families. Well-known examples are provided by the human globin gene clusters in which numerous copy number changes resulting in haemoglobinopathies have been described (Efremov 1978; Higgs et al. 1989; Thein 1993). De novo deletions between misaligned δ- and β-globin genes, analysed by single-DNA-molecule methods, proved to be extremely rare in sperm DNA, preventing further analysis of the reciprocity of ectopic recombination (Holloway et al. 2006). In contrast, as described in Chapter 4, the α-globin gene cluster shows a far greater propensity for deletion, most likely due to the extended length of homologous regions that encompass the α2- and α1-globin genes on normal αα chromosomes. Chromosomes with a single α-globin gene deletion (−α) are very common in some populations and are most likely maintained by malaria selection (Flint et al. 1986a). As shown in Chapter 4, using physical enrichment and single-DNA-molecule strategies, de novo −α deletions can be detected in normal individuals. These deletions proved to be common in both sperm and blood DNA. Some sperm deletions arise by unequal crossover between homologous chromosomes, as shown by exchange of flanking SNP markers. These exchanges most likely arise by unequal crossover at meiosis and in principle should generate reciprocal duplication products, namely ααα chromosomes carrying triplicated α-globin genes. In contrast, all blood deletions and the majority of sperm deletions arise by a pathway that does not involve any exchange between homologous chromosomes. These intrachromosomal mutants could be generated by reciprocal sister chromatid exchange or by other non-reciprocal processes of intramolecular deletion (Embury et al. 1980; Fishman-Lobell et al. 1992).

Chromosomes carrying three α-globin genes do indeed exist in human populations. These ααα chromosomes are less prevalent than −α chromosomes in most populations (Goossens et al. 1980; Flint et al. 1986b; Hill 1992), but this disparity does not necessarily indicate non-reciprocity in the duplication/deletion process because −α and ααα chromosomes are likely to be affected differently by selection. Since ααα chromosomes are not related to any pathological disorders, few studies have been done on these aberrant rearrangements; the dynamics and processes of duplications that generate ααα chromosomes remain completely unknown.

This chapter describes how direct analyses of de novo α-globin duplications, using physical enrichment and inverse PCR amplification, were performed on human DNA, allowing de novo duplication molecules to be detected and revealing the dynamics and processes of gene duplication. In addition, analyses of duplication frequencies and the distribution of ectopic exchange breakpoints indicated a very substantial degree of
reciprocity of duplication and deletion, suggesting a fully reciprocal process of ectopic recombination in the α-globin gene cluster.

6.2 Results

6.2.1 Amplification of triplicated α-globin gene molecules by small-pool PCR

To study the dynamics of duplication and its reciprocity with deletion caused by ectopic recombination, studies were carried out on the same αα/αα homozygotes in which the dynamics and processes of α-globin deletions had been fully characterised. As described in Chapter 4, various types of deletion can occur depending on the point of ectopic exchange. By far the most common in blood and sperm is the −α3.7 deletion that arises by unequal exchange between homology blocks Z2 and Z1, losing 3.8-kb DNA (Fig. 4.8). Analyses were therefore focussed on the reciprocal αααanti3.7 duplication mutants that have gained 3.8-kb DNA.

The α-globin gene region is very GC-rich and difficult to amplify by PCR, and the recovery of entire αααanti3.7 duplication (~16 kb) mutants by long PCR would almost certainly be impossible. Instead, a new PCR strategy was therefore developed to amplify only the duplicated region of the entire αααanti3.7 duplication. Making use of the structural difference between duplication and progenitor molecules, a novel inverse PCR strategy was developed using primers located between the α2- and α1-globin genes that point away from each other in αα chromosomes (Fig.6.1A). After duplication, these primers point towards each other and allow amplification of a fused region that spans the site of ectopic exchange (Fig. 6.1B). Inverse PCR primers were designed within a tiny region (86 bp) between two SNP heterozygosities shared by the two men, to amplify the exchange interval (3.8 kb) and allowing the haplotype origin of each mutant to be determined. This inverse PCR, in principle, should only amplify the duplicated region in αααanti3.7 duplication mutants but not progenitors. Hence, PCR amplification was first attempted to detect de novo duplication mutants directly in genomic DNA instead of performing any physical enrichment prior to PCR amplification. However, multiple PCR products were detected in almost all PCRs with increasing DNA inputs, despite the fact that genomic DNA was from an αα/αα homozygote (Fig. 6.1C). These PCR products were unlikely to be genuine because of their abnormally high frequencies, despite the fact
that some of them shared the same restriction profile with genuine products amplified from an ααα carrier (data not shown). Instead, they appeared to be PCR artefacts similar in structure to the fusion gene on ααα chromosomes.

\[\text{Fig. 6.1. Inverse PCR strategy for detecting duplication mutants. (A) Structure of progenitor α-globin gene molecules with homologous regions highlighted in coloured blocks. SNP heterozygosities are marked with grey circles in two men. Nested PCR primers for inverse PCR amplification are indicated by red arrows. (B) Structure of } \alpha\alpha\alpha_{\text{α37}} \text{ duplication mutants. PCR primers now point towards each other allowing amplification of the duplicated region. Haplotype-specific SNPs marked with asterisks are not necessarily present in duplication mutants unless ectopic exchange occurs before them. (C) Examples of PCR artefacts amplified in nested inverse PCR amplification. PCR products were amplified from sperm DNA from man 1 with different DNA inputs (1–15 ng per PCR), analysed by gel electrophoresis and visualised by staining with ethidium bromide. M, λ DNA × HindIII.} \]

6.2.2 Size fractionation for duplication mutant detection

6.2.2.1 Preparative gel electrophoresis and DNA recovery
Since the most likely source of these PCR products was as artefacts from progenitors, a size fractionation strategy was therefore applied to physically separate duplication mutants from progenitors. To give a distinct size difference between duplication and progenitor molecules, restriction enzymes were carefully chosen which cleaved only
outside of the fused α1/α2 gene. By digesting human genomic DNA with SphI plus XbaI, it was possible to release the target from αα chromosomes on a 7.3-kb DNA fragment (Fig. 6.2A). Duplication will increase this fragment size to 11.1 kb, allowing separation of progenitor and duplication molecules by gel electrophoresis (Fig. 6.2B). Since the frequency of duplication, in principle, should not be higher than that of deletion, 89–200 μg of double-digested sperm and blood DNA from both men were gel electrophoresed and size-fractionated. A total of twelve size fractions were collected ranging from 6.8 to 22.7 kb to include all potential ααα^anti3.7 duplication molecules.

Total DNA recovery was estimated to be 40–60% by gel electrophoretic comparison of an aliquot of DNA pooled from all fractions with known amounts of digested genomic DNA (as described in Chapter 4, section 4.2.3.1). In addition to the gel electrophoretic comparison (Fig. 4.4C), DNA recovery and size distribution in each fraction was estimated by PCR amplification of a 5.5-kb interval from a control 11.0-kb SphI–XbaI double-digest DNA fragment from chromosome 11. This control molecule identified fractions that should contain duplication molecules and also allowed overall DNA recovery to be estimated, again at ~40%.
6.2.2.2 PCR recovery of duplication molecules by inverse PCR

Pilot assays of inverse PCR amplification were performed using primers A22.5R plus A22.5F (1° PCR) and A22.5R2 plus A22.5F2 (2° PCR) which were designed within the 86-bp interval between two SNP heterozygosities shared by the two men (Fig. 6.1). However, PCR artefacts were still present in the smaller fractions containing progenitor molecules (data not shown). The design of the four primers was limited by the 86-bp interval, and some were not good for achieving efficient and specific PCR amplification. Primers were therefore designed in other intervals in each man where the haplotype origin of each mutant could still be identified after inverse PCR amplification (Fig. 6.3B and C). This strategy worked successfully in man 1 in which what appeared to be genuine mutants were efficiently recovered from triplicated α-globin gene molecules in fractions where they were predicted to land. In contrast, the shifted primer sites in man 2 resulted...
in PCR artefacts from αα chromosomes. PCR analysis of genomic DNA from an αα/αα homozygote showed that these inverse PCR artefacts could be eliminated by digestion with AfIII, which presumably separates the Z1 and Z2 homology blocks in any progenitor molecules and prevents artefacts generated by intramolecular looping back (detailed discussion in section 6.3.1). Each DNA fraction of man 2 was therefore overdigested with AfIII (4–6 units of enzyme per microlitre of fractionated DNA) prior to inverse PCR (Fig. 6.3C); these artefacts were eliminated to reveal authentic duplication molecules. Control overdigestion of genomic DNA from an αα/ααα heterozygote, who was accidentally identified in the −α chromosome screening of the LRI panel, showed no significant effect on the efficiency of recovery of duplication molecules (data not shown).

Poisson analysis of limiting dilutions of genomic DNA from the αα/ααα heterozygote showed that the nested inverse PCR strategy used for man 1 could amplify a single duplicated DNA molecule from 11.3 pg of DNA, indicating a 53% efficiency of PCR in recovering duplication molecules. The efficiency of the protocol used for man 2 was similar, at 66%.
Fig. 6.3. Strategies for detecting de novo duplications in the α-globin gene region. (A) Organisation of normal αα chromosomes and αααα dual3.7 duplications, with X, Y, and Z homology blocks shown in colour, and divergent primers used to amplify duplications in man 1 indicated by red arrows in progenitor. SNP sites used to characterise recombinants are shown as grey circles. Key restriction sites are also shown. (B) Detecting duplications in man 1. Genomic DNA was digested with SphI plus Xbal and fractionated to deplete progenitor molecules. Duplication molecules were recovered by nested inverse PCR amplification, using divergent primers that, after duplication, become convergent. Typing SNPs in the exchange interval (white and black circles for haplotypes A and B, respectively) allows intrachromosomal recombinants (intra AA, intra BB) to be distinguished from interchromosomal exchanges (inter AB, inter BA). (C) Detecting duplications in man 2, with digestion of fractionated DNA with AffI before inverse PCR to eliminate progenitor artefacts.
6.2.2.3 Distribution of duplications (size validation) and progenitor contamination

With both sets of primers customised for each man, the next goal was to detect duplication molecules by nested PCR across all fractions of sperm and blood DNA from both men. The DNA input for each PCR was adjusted to contain at most 1.3 amplifiable duplication molecules as established from pilot experiments performed on each fraction to obtain an initial estimate of duplication frequency. Secondary PCR products were analysed by agarose gel electrophoresis and visualised by staining with ethidium bromide (Fig. 6.4A). A total of 565 positive reactions seen in 5325 nested PCRs of sperm and blood DNA from both men were collected for further analysis.

To validate these duplication mutants, the distributions of progenitor αα molecule and of a control genomic DNA fragment matched in size to the ααα\textsuperscript{ani3.7} duplication molecule across all size fractions were estimated. The frequency of contaminating progenitor molecules in each fraction was determined by PCR amplification of a 4.9-kb segment lying within the SphI–XbaI interval of the α-globin gene region using PCR primers A18.1F and A22.9R. In addition, as mentioned in section 6.2.2.1, a 5.5-kb interval from a control 11.0-kb SphI–XbaI double-digest DNA fragment from chromosome 11 was also amplified to estimate the distribution of duplication molecules. Analysis of different electrophoretic size fractions for the control genomic DNA fragment matched in size to ααα duplication molecules and for progenitor αα molecules showed substantial separation of control DNA and progenitor with >95% depletion of progenitor molecules from fractions that could contain duplications (Fig. 6.4B). These analyses showed that de novo duplication mutants were correctly distributed across the size fractions and implied that these mutants must therefore correspond to authentic ααα chromosomes.
Fig. 6.4. Detection of de novo $\alpha_{\alpha\alpha^\text{3.7}}$ duplication molecules. (A) Examples of $\alpha_{\alpha\alpha^\text{3.7}}$ duplication molecules recovered by inverse nested PCR amplification of sperm DNA fractions 6 and 7 from man 1. Each PCR contained DNA derived from $1.8 \times 10^5$ and $1.1 \times 10^5$ amplifiable haploid genomes, respectively. PCR products were analysed by agarose gel electrophoresis. M, λ DNA × HindIII. (B) Cumulative frequencies of duplication molecules (in total, 249 mutants recovered from $4.0 \times 10^6$ amplifiable haploid genomes), across size fractions of SphI–XbaI-digested sperm DNA from man 1. DNA size ranges covered in each fraction are shown by grey bars. Fragment lengths of progenitor, control, and $\alpha_{\alpha\alpha^\text{3.7}}$ duplication molecules are indicated by dotted lines. The control molecule is an 11.0-kb SphI–XbaI genomic fragment from chromosome 11 matched in size to the $\alpha_{\alpha\alpha^\text{3.7}}$ duplication.

6.2.3 Characterising duplication mutants

Similar to the characterisation of deletion mutants described in Chapter 4 section 4.2.4, all $2^\circ$ PCR products of detected duplication mutant were collected and further amplified to a saturated level in a tertiary nested PCR. Ectopic exchange points were mapped by typing the tertiary PCR products of each mutant by dot-blot hybridisation with oligonucleotides specific to PSVs and to SNPs (details in Chapter 4, section 4.2.4). Haplotypes A and B in man 1 and man 2 differed at eight and eleven SNP sites, respectively, three of which mapped into the amplified interval (Fig. 6.5). A total of 30 mutants with exchange points that could not be located unambiguously by hybridisation were sequenced. An additional
53 mutants (~8.4% of the total) were also sequenced. None showed any rearrangements in addition to exchange of PSVs, and only four single base changes, presumably PCR misincorporations, were seen over 158 kb of DNA sequenced.

A total of 68 positive reactions contained mixed mutants as showed by mixed PSV and/or SNP sites. These mutants were separated by sequence-specific PCR directed to a mixed PSV or SNP site before characterisation. All mixtures were successfully resolved into their two to three constituent molecules. A full inventory of all mutant types over all PCRs was prepared and used to Poisson correct for instances of a PCR containing more than one molecule of a given type of recombinant (as for the deletion mutant example in Table 4.2). These corrections were modest, with only a 1.3-fold increase for the most abundant type of mutant.

Fig. 6.5. Heterozygous SNPs along the α-globin gene cluster in both men. (A) SNP sites on progenitor molecules. The region analysed as described in Fig. 6.3, with SNP sites marked as grey circles, and restriction sites indicated. Eight and eleven SNP sites were found in man 1 and man 2, respectively. Haplotype phasing allowed alleles to be assigned to haplotypes A and B. (B) SNP sites remaining on the amplified interval of Sphl-Xbal-digested αααααααααα duplication molecules. Only three SNP sites were left on the amplified interval after inverse PCR amplification (PCR primers are marked in red arrows). The haplotype-specific SNP marked with an asterisk is not necessarily present on the amplified interval unless the ectopic exchange occurs before it. Exchange types can be identified by typing the three SNPs on the amplified interval.
6.2.4 Duplication frequencies

As with deletions, duplication molecules were detected in both sperm and blood DNA, with mutants 10- to 20-fold more common in sperm than in blood (Table 6.1). Man 1 showed 2- to 5-fold more duplications than man 2 in both tissues analysed ($P < 0.001$ for each tissue), indicating significant interindividual variation in duplication frequency.

The frequencies of $\alpha\beta\delta$ duplications and $-\alpha$ deletions were very similar, both in the sperm and in the blood of man 1 ($P = 0.34$ and 0.37, respectively). This is consistent with reciprocal ectopic exchange and thus a predominantly if not exclusively intermolecular recombination pathway. In contrast, man 2 showed disparities in duplication/deletion frequencies; the 2-fold excess of duplications over deletions seen in sperm and the 6-fold excess of deletions in blood are highly significant ($P < 0.001$ in each tissue) and could point to factors that perturb mutation frequency in the germ-line and soma.

Table 6.1. Frequencies of ectopic $\alpha\alpha^{\text{anti}3.7}$ duplications and $-\alpha^{3.7}$ deletions in sperm and blood DNA in two men

<table>
<thead>
<tr>
<th></th>
<th>Mutation frequencies per haploid genome $\times 10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Man 1</td>
</tr>
<tr>
<td>$\alpha\alpha^{\text{anti}3.7}$</td>
<td>Interchromosomal</td>
</tr>
<tr>
<td></td>
<td>Intrachromosomal</td>
</tr>
<tr>
<td>Sperm</td>
<td>Total</td>
</tr>
<tr>
<td>Blood</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>$A/K$ selection</td>
</tr>
</tbody>
</table>

Percentages of inter- vs. intrachromosomal exchanges in sperm are indicated in parentheses. $A/K$ selection on blood DNA (see section 6.2.5) was capable of recovering duplications plus any potential extrachromosomal circles generated by intramolecular recombination (Fig. 6.6C). Dashes indicate not applicable.
6.2.5 Alternative approach of size fractionation

Since there were significant differences between deletion and duplication frequencies in blood from man 2 (Table 6.1), blood mutant frequencies were further checked by using a different size fractionation strategy capable of detecting not only duplications but also any extrachromosomal circles generated by intramolecular recombination (Fig. 6.6). DNA recovery and distribution of mutants were analysed similarly as described in previous sections (Chapter 4, section 4.2.3 and Chapter 6, section 6.2.2) but using PCR primers R85.7F and R87.5R to amplify a 1.8-kb interval from a 3.8-kb \( \text{Afl} \) DNA fragment from the MHC region.

For man 1, this approach gave an overall mutant frequency very similar to the duplication frequency \((P = 1)\). For man 2, the overall mutant frequency increased significantly \((P < 0.001)\) but not to the level of the \(-\alpha^{3.7}\) deletion frequency \((P = 0.005)\). It is not known whether this shift in frequency (Table 6.1) reflects the existence of genuine extrachromosomal circles. However, the structures of mutants recovered by this approach showed no significant differences from those isolated as genuine duplications (details in section 6.2.6).
Fig. 6.6. Alternative approach for detecting de novo ectopic recombinants. (A) Progenitor molecule of the α-globin gene cluster with homology blocks highlighted in different colours, SNP sites marked as grey circles, and an AflII restriction site. (B) Ectopic recombination between Z homology blocks via inter- and intramolecular pathways. –α\textsuperscript{3.7} deletions could be generated by various pathways involving exchanges between Z homology blocks; only the intermolecular pathway generating the reciprocal αα\textsuperscript{ant3.7} duplication and the intramolecular pathway that produces a deletion plus an excised circle are shown. (C) Blood DNA was digested with AflII only and size fractionated to recover 3.8 kb DNA fragments derived from duplicated αα\textsuperscript{ant3.7} chromosomes and from any linearised extrachromosomal circular DNAs, before amplification by inverse PCR (nested PCR primers are marked in red arrows). Loss of a 5' SNP site in man 1 after AflII digestion prevented the distinction of intra- vs. interchromosomal recombinants.
A flII progenitor homology block

B

intermolecular pathway

\[ \text{A} \]

\[ \text{B} \]

\[ \text{C} \]

\[ \text{man 1} \]

\[ \text{man 2} \]

\[ \text{PCR amplification} \]

\[ \text{exchange type} \]

\[ \text{intra AA} \]

\[ \text{intra BB} \]

\[ \text{inter AB} \]

\[ \text{inter BA} \]
6.2.6 Structures of duplications

6.2.6.1 Intra- and interchromosomal recombinants

All mutants were typed for informative SNPs to differentiate interchromosomal exchanges from intrachromosomal recombinants. Roughly 20% of sperm mutants (Table 6.1) showed exchange of these flanking markers, establishing that they had arisen by ectopic recombination between homologous chromosomes. In contrast, only two blood mutants of 145 (1.4%) showed marker exchange. This lack of marker exchange in blood recombinants was also seen for mutants recovered by Af/III digestion alone. This germ-line specificity of interchromosomal exchanges has also been seen for –α deletions (Fig. 6.7) and is fully consistent with these exchanges arising primarily by unequal exchange at meiosis.
**Fig. 6.7.** Distribution of ectopic exchange points in $\alpha\alpha\alpha^{\text{ant}3.7}$ duplications and $-\alpha^{3.7}$ deletions across the Z2/Z1 homology blocks. (A) Structures of duplications and deletions. The location of PSVs and SNPs within the aligned Z2 and Z1 homology blocks are shown below as black bars. A SNP present only in man 1 is marked with an asterisk. (B) Location of exchange points. The homology block can be divided into various intervals determined by PSV distribution, with SNPs further dividing the region in a haplotype-specific fashion. The resulting intervals of sequence identity (horizontal lines) mark regions to which ectopic exchanges can be mapped. The numbers of sperm and blood exchanges mapping to each interval are shown in black and red respectively under the line (−, no mutants). Additional blood mutants recovered by AflI digestion (Fig. 6.6) are indicated in purple. Duplication mutants were recovered from $4.0 \times 10^6$, $8.3 \times 10^6$ and $10.9 \times 10^6$ amplifiable haploid genomes of DNA from sperm, blood and blood (alternative approach that could also detect extrachromosomal circles) respectively, from man 1, and from $9.0 \times 10^5$, $13.9 \times 10^6$ and $7.7 \times 10^6$ haploid genomes from man 2. For deletions, mutants were recovered from $6.5 \times 10^6$ and $12.9 \times 10^6$ haploid genomes from sperm and blood DNA, respectively, from man 1 and $6.7 \times 10^6$ and $7.1 \times 10^6$ molecules from man 2. The data are taken from Chapter 4. (C) Cumulative number of exchanges, per $10^6$ haploid genomes, across the homology block for each class of $\alpha\alpha\alpha^{\text{ant}3.7}$ duplication and $-\alpha^{3.7}$ deletion. Intrachromosomal $\alpha\alpha\alpha^{\text{ant}3.7}$ duplications in both men and intrachromosomal $-\alpha^{3.7}$ deletions in man 1 identified in sperm DNA were relatively abundant and are shown separately with different scaling.
6.2.6.2 Simple vs. complex exchanges

Each duplication mutant was typed for PSVs that distinguish the α2- and α1-globin homology blocks. As with deletions (Chapter 4, section 4.2.6), almost all mutants (99.4%) were simple duplications with exchange points mapping to a single interval of sequence identity shared between homology blocks. Only four mutants were complex (one blood mutant in man 1 identified by the AffIII selection approach plus two sperm mutants and one blood mutant in man 2); all shared switching of PSVs near the site of ectopic exchange that presumably arose by patchy repair of heteroduplex DNA generated during recombination; a similar low frequency of complex events has been found with deletion mutants (Chapter 4).

6.2.6.3 Ectopic recombinants within Y homology blocks

Since the strategies developed were to detect ectopic recombinants arising from the misalignment of Z homology blocks, αααα*3.7 duplication molecules were expected to be the only mutant. However, four mutants with consistently shorter length (3.6 kb vs. 3.8 kb for αααα*3.7 mutant) were also found in the background of abundant αααα*3.7 duplication molecules. Sequencing showed that these exchanges were located in the very short Y2/Y1 blocks (Fig. 6.3A) and mapped to one or other of the longest regions of sequence identity (25, 39 bp) in these blocks (Fig. 6.8). Such Y2/Y1-driven duplications have yet to be reported in human populations. Nor were any Y2/Y1-mediated deletions detected in blood or sperm DNA (Chapter 4).

Fig. 6.8. Ectopic duplications arising from misaligned Y homology blocks. Four duplicated mutant molecules generated by Y exchanges were identified in man 1. Y1 and Y2 sequences are shown aligned over three lines, with PSVs marked in green. Ectopic exchange intervals are indicated by arrowed lines, with the number of exchanges seen in sperm (black) and blood (red) indicated above.
6.2.6.4 Duplications accompanied by deletion

In addition to the majority of ectopic exchanges arising from misaligned homology blocks, seven intrachromosomal mutants with abnormal lengths were also detected. These mutants were all smaller than the \( \alpha\alpha^{mu3.7} \) duplication molecules and contained 0.3- to 2.2-kb deletions extending over or within the Z1/Z2 homology block (Fig. 6.9). These breakpoints were different from normal unequal exchange breakpoints, with no homology identified between the 5' and 3' breakpoints.

Interestingly, sequencing showed that the deleted blood mutant from the \( \text{AfIII} \) selection had exchanged within the longest uninterrupted sequence (855 bp) in the Z1/Z2 homology block but was accompanied by a 0.3-kb deletion as well as a 7-bp insert at the 3' end of the exchange breakpoint (Fig. 6.10B). The 7-bp insert, TCTCAGG, was found in three locations along the \( \alpha\alpha \) progenitor (Fig. 6.10A); they were at ~1.3 kb upstream of the \( \psi\alpha1 \) gene, in the region extending over the 3' PSV site of the longest uninterrupted identical sequence in the Z2 homology block, and in the region between X1 and Y1 homology blocks. The most likely source of this insert was from the one in the Z2 homology block which was just located 7 bp away from the insert.

Despite the small number of deleted mutants detected in both men, none of these mutants showed exchange of flanking markers and no homology was seen between the 5' and 3' breakpoints (Fig. 6.9). If these mutants were genuine, it points to possible aberrant processing of recombination intermediates, and perhaps to other mechanisms besides unequal crossover (see Chapter 9, section 9.7). However, the authenticity of these mutants was uncertain given the broad DNA size coverage in most fractions.
Fig. 6.9. Duplications accompanied by deletions detected by inverse PCR amplification. (A) Mutants with abnormal length detected in man 1. Mutants were detected by strategies (size fractionation plus inverse PCR amplification) as described in Fig. 6.3. SNP sites are highlighted in grey circles; they can be separated into haplotype A (white) and haplotype B (black). Deleted regions in amplified intervals are shaded in grey. Sequences around the 5’ breakpoint (top) and 3’ breakpoint (bottom) are compared with the duplication mutant (middle), with sequence matches indicated by lines, and heterozygous SNP coloured in red. Regions of the fusion gene in each mutant are marked with Z1 and/or Z2, while exchange breakpoints are indicated by ‘Z1/Z2’. Mutants marked with an asterisk were detected in the correct size fractions of genomic DNA. (B) Duplications (or possibly extrachromosomal circles) with deletions detected in man 2.
A fl\^ selection approach
duplication/extrachromosomal circle

0.3 with 7 bp Insert
Fig. 6.10. Detailed structure of the blood mutant from AfII selection in man 2 which was accompanied by deletion plus insertion. (A) Locations of the 7-bp oligomer, TCTCAGG, along progenitors, with structures described as in Fig. 6.3. The one described in (B) is highlighted with an asterisk. (B) Structure of the blood mutant accompanied by a deletion plus an insertion as shown in Fig. 6.9. A 7-bp insert was located between the 5' and 3' breakpoints of a 0.3-kb deletion within the exchange points of the fusion gene. The nearest 7-bp oligomer (highlighted in green) was located 7-bp away from the insert across a PSV site of the Z2 homology block.

6.3 Discussion

6.3.1 Elimination of inverse PCR artefacts

The inverse PCR strategy, in principle, should only amplify duplication molecules but not progenitors (Fig. 6.3). However, pilot PCR assays showed PCR artefacts in DNA fractions from man 2 increasing with the degree of progenitor contamination. Since PCR primers for man 2 were located closer to the homology block Z1 than those for man 1 (Fig. 6.3), it may be the possible reason why PCR artefacts were only seen in man 2. One possible origin is that the 3' end of an extending DNA strand in homology block Z1 could loop back into block Z2 on the same DNA molecule during an early stage of PCR to create an amplicon from a progenitor molecule identical in size to that expected from an ααα chromosome (Fig. 6.11C). This model is supported by the fact that these inverse PCR artefacts could be eliminated by digesting fractionated DNA prior to PCR with a restriction enzyme located between the divergent primers in progenitor molecules. This
strategy presumably separates the Z1 and Z2 homology blocks in any progenitor molecules and prevents PCR artefacts from this intramolecular looping back.

![Diagram of PCR artefact generation](image)

**Fig. 6.11.** Mechanism of generating the inverse PCR artefact. (A) Inverse PCR strategy on progenitor molecules. Divergent primers between homology blocks Z2 and Z1, in theory, cannot generate any PCR products. (B) Inverse PCR amplification on duplication molecules. Inverse PCR primers become convergent on duplication molecules. The fused homology block (Z1/Z2) on duplication molecules is amplified. (C) PCR artefact arising from intramolecular looping back. The 3' end of an extending DNA strand in the Z1 homology block may displace in an early stage of PCR and misprime onto paralogous sequences in the Z2 homology block. An amplicon which is identical to the fused homology block on duplication molecules is thus generated from progenitor molecules.

### 6.3.2 Reciprocity of −α3.7 deletions and αααanti3.7 duplications

#### 6.3.2.1 Two distinct mechanisms

As with −α3.7 deletions, *de novo* αααanti3.7 duplications are common in blood and sperm and appear to arise by two distinct mechanisms. Ectopic exchanges between homologous chromosomes generated a minority of sperm duplications, while intrachromosomal exchanges account for the majority of duplications in the germ-line (~79%) and especially in the soma (~99%). The germ-line specificity of interchromosomal exchanges indicated that these exchanges had most likely arisen by unequal exchange at meiosis. Moreover,
interchromosomal deletion and duplication frequencies in sperm in man 2 were indistinguishable (Table 6.1), consistent with a fully reciprocal unequal exchange process at meiosis. Although man 1 showed a 1.6-fold excess of deletions, this difference is of marginal significance \( (P = 0.026) \). Further evidence of these mutants being meiotic recombinants was indicated by the presence of haplotype symmetry (Fig. 6.7B). For example, man 1 showed very similar numbers of AB- and BA-type sperm duplications (20 and 17, respectively, \( P = 0.622 \)). Similarly, interchromosomal exchanges in sperm DNA from man 2 were also of similar frequencies (28 for AB- vs. 36 for BA-type, \( P = 0.317 \)). The frequency of these duplications and deletions would therefore simply reflect the frequency of unequal crossover between \( Z \) blocks on homologous chromosomes at meiosis (\( \sim 10^{-5} \) per sperm).

In contrast, and similar to deletions, the frequencies of intrachromosomal exchanges were significantly higher than those of interchromosomal recombinants in sperm from both men \( (P < 0.001 \) for both men). These intrachromosomal sperm exchanges could be mitotic recombinants and/or meiotic exchanges arising from the interaction between sister chromatids; these would be indistinguishable. However, significant haplotype asymmetry of exchange breakpoints excludes the possibility that these exchanges were purely meiotic sister chromatid recombinants (detailed discussion in section 6.3.3). As reported in Chapter 4, section 4.3.2, these erratic fluctuations in frequency were commonly seen in intrachromosomal exchanges in sperm as well as in blood (Fig. 6.7B, C) but not in interchromosomal recombinants. This putative mosaicism again pointed to the distinct characters of mitotic and meiotic mechanisms in both duplications and deletions.

### 6.3.2.2 Similar distribution of ectopic exchange points

Further evidence for reciprocal ectopic exchange in \( \alpha \)-globin genes was revealed by the distribution of exchange breakpoints. As with \( -\alpha^{3.7} \) deletions, unequal exchange points were fairly randomly distributed across the \( Z \) homology blocks (Fig. 6.7), suggesting that most regions are equally prone to exchange, irrespective of the presence of PSVs, and arguing against the presence of a local recombination hot spot of the type that drives unequal exchange in genomic disorders such as CMT1A (Reiter et al. 1996; Reiter et al. 1998). The comparatively low ectopic exchange frequency identified in the 166-bp interval at the 3' end of the \( Z2/Z1 \) homology block (Fig. 4.11C) is also observed in \( \alpha\alpha^{anti3.7} \) duplications (Fig. 6.7). Although this shows some non-randomness in the
distribution of exchange breakpoints, it implies that factors other than local homology, such as topological constraints (Nicholls et al. 1987), can affect both deletions and duplications in the same fashion.

6.3.3 Mechanisms of intrachromosomal αα^anti3.7 duplications

Most duplications in blood and sperm are intrachromosomal without flanking marker exchange (Fig. 6.7B and Table 6.1) and must therefore have arisen by unequal sister chromatid exchange involving crossover or gene conversion. For blood, these events must result from mitotic recombination, and their prevalence in sperm suggests a premeiotic component to germ-line duplication occurring at some stage before spermatogenesis. Unlike meiotic events, these duplications can show substantial distortions in frequency between haplotypes. For example, sperm mutants in man 1 mapping to the longest interval of sequence identity (Fig. 6.7B) were more common on haplotype B than on haplotype A (108 vs. 55, \( P < 0.001 \)). A similar observation was found as well in man 2 but with higher frequency on haplotype A than on haplotype B (100 vs. 44, \( P < 0.001 \)).

As with deletions, the most reasonable explanation for these haplotype asymmetries is mutational mosaicism whereby an early mitotic recombination event can spread to multiple descendant cells. Clear instances of very unusual deletions that should be rare but which were detected repeatedly in deletion surveys have already been noted. A similar phenomenon is observed for duplications, for instance the six haplotype B sperm duplications in man 1 that all mapped to an interval just 8 bp long at the beginning of the Z homology block (Fig. 6.7B). Given the expected rarity of exchange in such a small region, these six mutants most likely derived from a single ancestral exchange event. Mosaicism is therefore not unusual and could provide a reasonable explanation for other haplotype distortions in intrachromosomal rearrangement frequencies. Sperm mosaicism also suggests that a significant proportion of intrachromosomal sperm events must be premeiotic in origin rather than arising by meiotic sister chromatid exchanges.

Although mosaicism could play a major role in influencing the frequency of germ-line and somatic rearrangements, other factors could also contribute to these haplotype asymmetries. Perhaps SNPs that can act as PSVs and disrupt homology blocks in a haplotype-specific fashion could perturb the ectopic exchange process. However, there is no obvious effect of sequence interruptions on unequal exchange rate and distribution for interchromosomal duplications and deletions (Fig. 6.7B). There are instances of an apparent effect on intrachromosomal rearrangements (for example, the reduction of 108...
haplotype B sperm duplications in the longest region of sequence identity in man 1 to 55 haplotype A duplications in the presence of two disrupting paralogous SNPs). These are, however, countered by instances of exchange frequencies showing elevation despite the presence of a disruption (for example the 5' adjacent interval shows four haplotype A duplications and 27 haplotype B duplications, despite an interruption in the latter). Haplotype disparities could also be caused by epigenetic marks or distal regulators in cis that influence ectopic recombination frequencies in a haplotype-specific fashion. There is, however, no clear evidence for a haplotype that shows a consistent enhancement of both duplication and deletion predicted from such regulators (Fig. 6.7B). The dominant factor creating haplotype asymmetries therefore appears to be mutational mosaicism.

6.3.4 Duplication dynamics and the population incidence of $\alpha\alpha^{anti3.7}$ chromosomes

As with $-\alpha$ deletions, the high frequency of de novo $\alpha\alpha^{anti3.7}$ duplications in sperm ($6.2 \times 10^{-5}$ per sperm in man 1, $2.6 \times 10^{-5}$ in man 2) predicts a correspondingly very high incidence of $\alpha\alpha$ chromosomes in human populations. As described in Chapter 5, the low incidence of $-\alpha$ chromosomes in populations not affected by malaria can only be maintained by significant selection against $-\alpha/-\alpha$ homozygotes and/or $-\alpha/\alpha\alpha$ heterozygotes. Precisely the same arguments apply to $\alpha\alpha$ duplications, which also show a very low incidence in most populations (0.002–0.006) (Flint et al. 1986b), which can only be maintained in the face of the strong duplication pressure by selection against $\alpha\alpha\alpha$ carriers and/or homozygotes, with strengths of selection similar to those reported for $-\alpha$ deletions (Chapter 5, section 5.3.1). Given the normal haematological profiles of individuals with $\alpha\alpha\alpha$ chromosomes (Goossens et al. 1980; Trent et al. 1986), the nature of this selection, whether through subtle imbalances between $\alpha$- and $\beta$-globin levels or through chromosomal processes such as segregation distortion (Orioli 1995; Dean et al. 2006), remains completely unclear.

6.4 Conclusions

Reciprocal ectopic recombination has long been postulated as a major driving force in human genome rearrangements and in the evolution of gene families (Glusman et al. 2000; Park et al. 2002; Lee and Lupski 2006; Redon et al. 2006). Although what appear to be reciprocal products of unequal exchange can be found in human populations, for example
-α and ααα globin chromosomes, these do not exclude the existence of distinct deletion and duplication pathways. The same reservations apply to genomic disorders such as CMT1A/HNPP, where reciprocal exchange points in duplications and deletions reflect a common initiating mechanism for unequal exchange but not necessarily to a fully reciprocal exchange process (Chance et al. 1994; Reiter et al. 1996; Reiter et al. 1998).

This chapter reveals that de novo gene duplications can be recovered from human DNA and used to analyse processes of duplication. The overall picture in the α-globin gene cluster is one of substantial reciprocity in the dynamics and processes of duplication and deletion, with instability occurring both in the germ-line and in somatic DNA. Unequal exchange between homologous chromosomes at meiosis appears to play only a minor role in establishing overall germ-line instability levels. Although a possible contribution from meiotic sister chromatid exchange cannot be excluded, the dominant germ-line process appears to be mitotic recombination. The process generates rearrangements whose frequency can be significantly influenced by factors such as mutational mosaicism, which in turn can lead to erratic inflation of various classes of rearrangement, apparently affecting the overall incidence of duplications and deletions. This study also establishes that other pathways such as intramolecular recombination play at best only a minor role in the generation of deletions. Finally, the α-globin genes present a picture of a gene cluster being subjected to very strong forces of duplication/deletion that contrast strongly with the rarity of rearranged chromosomes in most populations, giving clues about the strength of selective forces that must have acted to stabilise gene copy number.
Chapter 7

Detection of Extrachromosomal Circular DNA

7.1 Introduction
As described in previous chapters, intra- and intermolecular exchanges are the two major ectopic recombination pathways, producing deletions plus either extrachromosomal circular DNAs or duplications, respectively. Given that deletions and duplications can sometimes lead to pathogenic disorders like CMT1A/HNPP and Smith–Magenis syndrome/dup(17)(p11.2) (Lupski 1998; Inoue and Lupski 2002) in humans, studies have been focussed mostly on these two products. In contrast, very few studies have been done on extrachromosomal circles. Hence, the existence, topology and characteristics of these ectopic recombinants are completely unknown.

Despite little information on extrachromosomal circles in vivo in humans, small polydisperse circular DNA (spcDNA) has commonly been detected in other systems (Tsuda et al. 1983; Stanfield and Helinski 1986; Fujimoto and Yamagishi 1987; van Loon et al. 1994; Cohen et al. 1999). Extrachromosomal circular DNA has been found in most in vitro studies of higher eukaryotes, for instance, in Chinese hamster ovary (CHO) cells (Stanfield and Helinski 1986) and HeLa cells (van Loon et al. 1994). These circular DNA molecules are homologous to, for example transposons (Jones and Potter 1985; Schindler and Rush 1985) and tandemly repeated DNA sequences (Kunisada and Yamagishi 1987), indicating that they are most likely derived from chromosomal DNA. Many mechanisms have been proposed for generating spcDNA including homologous or illegitimate recombination and reverse transcription; however, exact mechanisms have yet been determined.

The formation of extrachromosomal circles carrying an α1/α2-fusion globin gene via an intramolecular mechanism within the α-globin gene cluster has been proposed (Embury et al. 1980). However, no evidence has yet been found to prove this hypothesis. It is arguable whether these circular DNA molecules, carrying only the α1/α2-globin gene region but no centromere, can be transmitted in cell division, although the transmission of covalently closed circular (CCC) supercoiled viral DNA has been found in dividing hepatocytes in other animal models (Dandri et al. 2000; Schorr et al. 2006). Similarly, with the absence of any replication origins, the chance of detecting these circular DNA molecules is expected to be extremely low unless the frequency of this
intramolecular recombination is very high. Despite de novo ectopic deletion and duplication mutants revealing substantial reciprocity of the dynamics and processes in the $\alpha$-globin genes, blood mutants detected by the $AflIII$ selection approach (described in Chapter 6, section 6.2.5) implied that extrachromosomal circular DNA molecules might be generated in the soma. However, their steady state level is unknown though likely to be very low.

In this chapter, a new strategy for detecting extrachromosomal circles is developed. By using single-molecule techniques and inverse PCR amplification, circular DNA molecules carrying an $\alpha1/\alpha2$-fusion globin gene region were analysed directly in genomic DNA and characterised into different topological forms such as nicked circles and covalently-closed supercoiled circular DNA molecules. This work provides insights for the existence of putative extrachromosomal circles in the soma indicating that they might be recombinants arising from intramolecular ectopic recombination within the $\alpha$-globin gene cluster.

7.2 Results

7.2.1 Selection of donor samples

Since there is no information about extrachromosomal circles arising by intramolecular recombination, whether the process is mitotic and/or meiotic is completely unknown. If circles were produced by a meiotic process, they might be detectable in spermatocytes but could fail to be retained in sperm. Given that normal testes will contain both mitotic and meiotic cells, this provides the optimal tissue for circle detection. Hence, the first attempt was therefore to detect circles in testis.

A panel of 10 testis samples from testicular cancer and gender reassignment patients were provided by the Leicester General Hospital. With an unknown but probably very low frequency of extrachromosomal circles, the selection of donors was therefore focussed on five complete testis samples from gender reassignment patients to provide sufficient material for analysis. These samples were then checked by additional criteria including lack of $\alpha$-globin gene rearrangements, the presence of informative SNP markers in the target region (data not shown), and also whether the testis showed normal levels of meiotic recombination activity as monitored by minisatellite instability plus meiotic recombination hot spot activity (Tamaki et al. 1999; Kauppi et al. 2004; I. Berg and A.
Jeffreys, unpublished data). Only one sample fulfilled these selection criteria and was therefore chosen for further analysis. Preliminary studies of the instability of minisatellites CEB1 and B6.7 in this testis DNA sample revealed a mutation frequency of 1.2% and 0.3% per haploid genome, respectively, in total testis DNA (I. Berg and A. Jeffreys, unpublished data). These mutation frequencies indicated that ~8–18% of total testis cells have completed meiosis I, suggesting that this gender reassignment sample was active in meiosis. Since no mature sperm were observed in this sample (I. Berg, personal communication), this minisatellite instability gives an indication of recombination frequency in post-meiosis I cells (secondary spermatocytes and spermatids).

In addition, as reported in Chapter 6, the discrepancy between the frequencies of blood duplication mutants obtained from man 2 by the two size fractionation strategies might indicate the presence of excised circular DNAs in somatic DNA. Blood DNA from man 2 was therefore also screened for extrachromosomal circles.

### 7.2.2 Collecting extrachromosomal circular DNA from genomic DNA

Since the fusion gene in these extrachromosomal circles is identical to the duplication in \( \alpha \alpha^{anti} \) molecules (Fig. 6.6B), it is impossible to differentiate circular DNA molecules from duplications if inverse PCR amplification is used directly to recover these circular molecules from total genomic DNA. Therefore, similar to most traditional methods including CsCl centrifugation (Sambrook and Russell 2001), pulsed-field electrophoresis (Simske and Scherer 1989) and two-dimensional gel electrophoresis (Oppenheim 1981), circular DNA molecules have to be physically separated away from chromosomal DNA prior to analysis. The strategy chosen was again separation by agarose gel electrophoresis.

#### 7.2.2.1 Mobility of extrachromosomal circles

To track the presence of any extrachromosomal circles derived from the \( \alpha \)-globin genes, it was necessary to know the mobility of circular DNA molecules during agarose gel electrophoresis. A ~0.8-kb DNA insert from chromosome X was cloned into a 3.0-kb pGEM-T Easy vector, transformed and propagated in DH5 alpha competent *E.coli* cells. This ~3.8-kb plasmid DNA was matched in size to any extrachromosomal circles arising from intramolecular ectopic exchange between the Z homology blocks of the \( \alpha \)-globin gene cluster. Since circular DNA molecules exist in different topologies, for example, nicked and supercoiled circles, different forms of the positive control were also analysed.
In general, the majority of plasmid DNA purified by alkaline lysis is supercoiled. To obtain both supercoiled and nicked forms of the positive control, supercoiled DNA was therefore digested with a nicking endonuclease, Nt.BbvCl, which cleaves only once in this plasmid control on one strand of DNA, generating nicked circular DNA. The open and closed circular forms of DNA were checked in parallel with linearised plasmid DNA, produced by cleavage with XbaI, by gel electrophoresis and visualised by staining with ethidium bromide (Fig. 7.1). As expected, supercoiled DNA migrated faster than linearised DNA, while the mobility of the nicked circular DNA was reduced by its open structure. The mobility of these two circular forms of plasmid DNA allowed the gel location of any extrachromosomal circles (~3.8 kb) derived from the α-globin gene cluster to be determined.

**Fig. 7.1.** Three different forms of the 3.8-kb plasmid control. These control DNAs were analysed in parallel with DNA marker (M, λ DNA × HindIII) and supercoiled DNA marker (SM) in 0.8% LE agarose gel electrophoresis and visualised by staining with ethidium bromide.

### 7.2.2.2 Preparative gel electrophoresis

High molecular weight genomic DNA is very hard to electrophorese into an agarose gel, making gel loading and separation for circular DNA extremely difficult. Total genomic (testis or blood) DNA was therefore digested with a restriction enzyme to reduce DNA fragment size. The goal was to use an enzyme which cleaves frequently in the human genome but not in the α-globin gene cluster, maintaining the α-globin gene region on as large a DNA fragment as possible. This strategy can provide efficient gel electrophoresis, ensure undisturbed topologies of extrachromosomal circles and also minimise progenitor contamination. The resequencing data of progenitor haplotypes from the two men analysed, together with Ensembl data showed EcoRV to be highly suitable. There were no
EcoRV cut sites within the 4.6-kb α2–α1 region; the nearest EcoRV cleavage sites were ~16.1 kb upstream of the α2 gene and ~14.3 kb downstream of the α1 gene, generating a 35-kb α2–α1 DNA fragment. In addition, there were no SNPs present in this fragment that could generate an EcoRV cut site. Since the frequency of extrachromosomal circles generated from the α-globin genes was completely unknown but expected to be very low, 265 µg EcoRV-digested testis DNA from the gender reassignment patient and 282 µg EcoRV-digested blood DNA from man 2 were analysed by preparative gel electrophoresis. 10 ng of each circular form (supercoiled and nicked) of the positive plasmid control was added to the EcoRV digest prior to electrophoresis to provide the internal size marker controls. Large amounts of control DNA were also electrophoresed in adjacent slots to monitor migration (Fig. 7.2).

![Fig. 7.2. Size fractionation for circular DNA molecules. (A) Locations of 3.8-kb nicked (open circle, OC) and supercoiled (closed circle, CC) plasmid DNAs in agarose gel electrophoresis. EcoRV-digested genomic DNA (with OC and CC as internal positive controls) was electrophoresed in parallel with DNA marker (M, λ DNA × HindIII) and plasmid DNA in two different topologies (OC and CC). (B) Size fractionation of EcoRV-digested genomic DNA. A total of 20 fractions of EcoRV-digested testis DNA ranged from 1.7 kb to 18.3 kb were collected. Alignments of these fractions were analysed in parallel with DNA marker (M) and the two plasmid DNA controls (OC and CC) by electrophoresis on a 0.8% HGT agarose gel and visualised by staining with ethidium bromide.](image-url)
7.2.2.3 DNA recovery and distributions of circular DNA molecules and progenitors

Since there was a major difference of electrophoretic mobility between supercoiled and nicked DNAs (Fig. 7.2), a total of 16–20 fractions ranging from 1.5 kb to 18.3 kb were collected (Fig. 7.3B). Similar to methods described in Chapters 4 and 6, DNA recovery was estimated both by gel electrophoretic comparison of an aliquot of DNA pooled from all fractions with known amounts of digested genomic DNA (see example in Fig. 4.4C) and by PCR amplification of two control EcoRV-digested MHC fragments (2.1 kb and 4.2 kb). The recovery of these two control fragments was very similar, giving an overall DNA recovery of 44–50% averaged from the two amplicons for testis and blood DNA.

Since 10 ng of each topological form of the positive control plasmid was present in the EcoRV-digested testis/blood DNA (section 7.2.2.2), the distribution of circular DNA molecules was directly estimated by PCR amplification of a 1.2-kb amplicon from the control plasmid across all size fractions. However, because of the broad distribution of the control plasmid especially in the blood DNA fractionation (Fig. 7.3B) and the possibility of different circular DNA topoforms, mutants were potentially detectable across many fractions in contrast to the narrower distribution of deletions and duplications (Figs. 4.5 and 6.4). The distribution of progenitor molecules was also determined by PCR amplification of a 3.8-kb fragment of the α-globin genes using primers A18.1F and A21.8R2 (Fig. 7.3B).
Fig. 7.3. Distribution of control, progenitor and mutant molecules across size fractions. (A) Examples of Dral linearised circular DNA molecules recovered by inverse PCR amplification from testis DNA fraction 14 (Left) and blood DNA fraction 12 (Right). The testis mutant carries a small deletion, while the two blood mutants are of the size expected from an extrachromosomal circle. Each PCR contained DNA derived from $1.5 \times 10^5$ and $1.3 \times 10^5$ amplifiable haploid genomes, respectively. PCR products were analysed by agarose gel electrophoresis and visualised by staining with ethidium bromide. M, X DNA $\times$ HindIII. (B) Cumulative frequency of plasmid controls and $\alpha$-globin progenitor molecules across size fractions of EcoRV-digested DNA. Size ranges covered by fractions are shown as grey bars. The closed circle (CC) and open circle (OC) controls are undigested and Nt.BbvCI-digested 3.8-kb plasmids, respectively. (C) Distribution of putative mutant molecules across size fractions. Mutants are grouped into two classes; those matched in size to extrachromosomal circles generated by intramolecular recombination between $\alpha2$ and $\alpha1$ genes (red blocks), and mutants with smaller size (blue blocks). These mutants were recovered from $15.9 \times 10^6$ amplifiable molecules from testis DNA (Left) and $6 \times 10^6$ molecules from blood DNA from man 2 (Right).
7.2.3 Progenitor digestion and detection of putative mutants

The progenitor α-globin gene cluster was predicted to lie on a ~35-kb EcoRV DNA fragment. However, smaller progenitor molecules were still detected in some of the larger fractions in both samples (Fig. 7.3B). Those are most likely broken DNA molecules, and could generated inverse PCR artefacts as described in Chapter 6. To minimise these PCR artefacts, each fraction was overdigested with Dral (4–6 units of enzyme per microlitre of fractionated DNA) prior to inverse PCR amplification. This digestion cleaved progenitor molecules between the inverse primer sites and also linearised any potential circular DNA molecules, removing any amplification constraint caused by the structure of supercoiled DNA.

Putative de novo mutants were indeed detected across size fractions by nested inverse PCR amplification, using the same primers as for detecting de novo duplications in man 1 (Fig. 6.3). The location of these primers also allowed the haplotype origin of mutants detected in testis DNA to be determined, whereas the loss of 3’ flanking SNPs in man 2 prevented the distinction of interchromosomal recombinants from intrachromosomal exchanges (Fig. 7.4).

**Fig. 7.4.** Dral digestion for minimising PCR artefacts. Structures of progenitor and extrachromosomal circular DNA molecules with coloured homology blocks, Dral restriction site, SNP sites (circles) and PCR primers (red arrows) indicated. Dral digestion was performed in each fraction prior to inverse PCR amplification, to minimise PCR artefacts generated from progenitor DNA molecules. Potential circular DNAs were linearised by this digestion, allowing efficient amplification. Haplotype origins can be determined for both testis and blood mutants, but interchromosomal exchanges (which cannot occur for genuine circles) can only be diagnosed in the testis.
7.2.4 Characterisation of putative mutants

All 2° PCR products of the detected putative mutants were collected and reamplified in tertiary PCR to a saturated level for characterisation by sequencing. A total of 24 candidates were recovered from 15.9 x 10^6 and 6 x 10^6 amplifiable progenitor molecules from the testis sample and from blood of man 2 (12 mutants each) across various DNA size fractions, respectively (Fig. 7.3). These putative mutants could be categorised into two groups: extrachromosomal circles generated from the α2 and α1 genes with the expected size (3.8 kb) and mutants accompanied by a deletion sizing from 0.4–2.8 kb extending over or within the Z1/Z2 homology block (Fig. 7.5), with no homology identified between the 5' and 3' breakpoints (data not shown). Almost all putative 3.8-kb mutant molecules were simple exchanges except one complex recombinant with PSV switching recovered from blood DNA. All molecules carrying both 5' and 3' flanking markers were found to be intrachromosomal recombinants (AA or BB type), consistent with an intramolecular loop-out process. As with deletions and duplications, the exchange breakpoints of most putative 3.8-kb circular mutants (9 out of 14) mapped to the longest region of uninterrupted identical sequence (855 bp) shared between Z homology blocks. However, with the small number and dispersed distribution of these candidates, the authenticity of these putative mutants remained in doubt.

Although putative α1/α2-fusion mutants (3.8 kb), especially in testis, were clustered in correct size fractions for closed circles, there was a possibility that these putative mutants were generated from intramolecular looping back within an αα progenitor molecule (Fig. 6.11). However, this is unlikely to be true, given that the progenitor contamination was very low and all fractions were further digested with DraI to eliminate any potential templates for looping back. Alternatively, those 3.8-kb amplicons distributed in DNA fractions ranging from 1.9–2.8 kb could be PCR artefacts arising by annealing of broken progenitor molecules during an early stage of PCR. These potential PCR artefacts should have appeared, by chance, as both intra- and interchromosomal exchanges. However, all four exchanges were intrachromosomal, although this is only marginally insignificant (P = 0.0625). Given the unknown but probably low concentration of these broken molecules, plus the low annealing kinetics of these templates during PCR, it seems unlikely that these putative mutants were generated by annealing.

In contrast, despite the fact that some putative open circular 3.8-kb mutants were detected in correct size fractions for the OC control in blood DNA (Fig. 7.3B and C), it
was very difficult to judge whether they were genuine mutants because of their rarity and the comparatively higher concentration of progenitor contamination in those fractions.

For mutants accompanied by a deletion (<3.8 kb), detected mainly in testis DNA, it was even more difficult to validate their authenticity. These mutants were distributed broadly across most fractions and there was no obvious relationship with the control circular DNA sizes. These mutants could arise as jumping PCR artefacts between broken molecules, giving both intra- and interchromosomal exchanges; however, no mutants with exchange of flanking marker were observed ($P = 0.0625$), but again with only four putative mutants detected.
Fig. 7.5. Exchange point distribution of putative mutants. (A) Structure of α-globin genes and distribution of PSVs on Z homology blocks. Structure of progenitor molecules with coloured homology blocks (as described in Fig. 6.3), Drai cleavage site and flanking SNP sites (grey circles) are indicated. PSVs differing between Z2 and Z1 homology blocks are shown as red and blue lines, respectively, on both haplotype A and B. (B) Structure of putative mutants recovered from testis DNA and blood DNA from man 2. Deleted regions are marked with brackets. Eight out of 12 mutants recovered from testis DNA showed deletions varying from 0.7 to 2.7 kb, while only two deleted mutants were identified in blood DNA. Exchange breakpoints in 3.8-kb mutants are indicated by regions between exchanges of PSVs (blue to red lines). SNPs that mapped within homology blocks (haplotype-specific PSVs) are indicated by purple lines. PSV switching in a complex mutant is marked with asterisks.
A

**A**

![Diagram of testis DNA and man 2 blood DNA with specific markers and annotations.](image)

**B**

![Table showing fraction size and number of samples.](image)

<table>
<thead>
<tr>
<th>Fraction</th>
<th>Size (kb)</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3.1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3.8</td>
<td>1</td>
</tr>
<tr>
<td>6, 7</td>
<td>3.8</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1.6</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3.8</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>2.3</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>3.4</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>2.3</td>
<td>1</td>
</tr>
</tbody>
</table>
| 20       | 1.8       | 1

---

**man 2 blood DNA**

![Diagram of man 2 blood DNA and specific markers and annotations.](image)

**B**

![Table showing fraction size and number of samples.](image)

<table>
<thead>
<tr>
<th>Fraction</th>
<th>Size (kb)</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 7, 8, 15</td>
<td>3.8</td>
<td>4</td>
</tr>
<tr>
<td>12 (x3), 15</td>
<td>3.8</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>3.8</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>3.8</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1.1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>2.9</td>
<td>1</td>
</tr>
</tbody>
</table>
7.3 Discussion

7.3.1 Low frequencies of putative extrachromosomal circles
The frequencies of putative 3.8-kb extrachromosomal circles detected in testis DNA and blood DNA from man 2 were $2.5 \times 10^{-7}$ and $1.7 \times 10^{-6}$ per haploid genome, respectively. These estimates give no clues about the actual rate of production of intramolecular recombinants in $\alpha$-globin genes, although the blood mutation frequency in man 2 was numerically very similar to the estimate predicted by the AflIII selection approach described in the previous chapter ($1.7 \times 10^{-6}$ vs. $2.1 \times 10^{-6}$ per haploid genome) (Table 6.1). It is likely that the actual mutation frequencies were higher than the observed incidence of circles given their probable lack of replication and segregation functions. However, the low mutation frequencies of extrachromosomal circles arising from ectopic exchange within $\alpha$-globin genes were expected, since the substantial reciprocity of the dynamics and processes of deletion and duplication implied that an intramolecular pathway would play at best only a minor role in deletion. Nevertheless, if these putative mutants are genuine extrachromosomal circles, it would give the first evidence for the existence of an intramolecular ectopic recombination pathway in humans.

7.3.2 Different topoforms of putative circles
Extrachromosomal circular DNA molecules have been isolated from model systems such as Drosophila (Shinomiya and Ina 1991), mice (Yamagishi et al. 1983; Fujimoto et al. 1985) and animal cell lines (Jones and Potter 1985; van Loon et al. 1994). Strategies used in these studies include CsCl density gradient centrifugation, pulsed-field gel electrophoresis, two-dimensional gel electrophoresis and electron microscopy, but are expensive and time-consuming. Another approach using size fractionation in conjunction with inverse PCR amplification was used instead in this study for detecting extrachromosomal circles arising from ectopic recombination within $\alpha$-globin gene molecules.

Despite the fact that the authenticity of the putative 3.8-kb extrachromosomal circles especially those from blood DNA has not been proved, only supercoiled 3.8-kb circles were recovered from testis whereas both topoisomers appeared to exist in blood DNA from man 2 (Fig. 7.3C). This testis/blood difference is significant (Fisher's exact test, $P = 0.035$). The frequency difference observed in this case might be controlled by intrinsic and/or external factors. For example, tissue-specific in vivo nicking of circular
DNA molecules might be a possible reason for open circles generated only in blood. External factors such as in vitro nicking during DNA preparation were unlikely since DNAs from these two DNA samples were extracted by a very similar approach. However, given insufficient evidence and poor understanding of these putative α1/α2 circular DNA molecules, exact explanations for this topological difference are completely unknown.

7.3.3 Mutants accompanied by deletions
Mutant molecules accompanied by a deletion accounted for 42% of the total number of recombinants recovered from the two samples (Fig. 7.3C). Since the two circular forms of the positive control were only designed for estimating the distribution of 3.8-kb circles, it is not possible to validate the genuineness and to identify the topology of these deleted mutants based only on their distribution. Interestingly, sequencing data revealed that 80% (8 out of 10) of the 5' breakpoints of these ten deleted mutants were clustered in a ~0.3-kb region spanning Y1 and the beginning of Z1 homology blocks; this clustering is significant ($P = 0.0047$ by computer simulation) (Fig. 7.6). Likewise, a similar clustering of breakpoints reported in non-Lepore sperm deletions in the β-globin gene region suggested the possibly existence of a local deletion regulator (Holloway et al. 2006). Although similar arguments could be applied in this situation, the possibility of these deleted mutants being PCR artefacts could not be excluded.
Fig. 7.6. Clustering of 5' breakpoints in deleted mutants. (A) Structure of linearised 3.8-kb extrachromosomal circles as described in Fig. 7.5B, with Y1 plus Z1/Z2 homology blocks and inverse PCR primers (red arrows) highlighted. (B) Distribution of 5' and 3' breakpoints of deleted mutants. 5' and 3' breakpoints of the ten deleted mutants are indicated by blue and pink arrows, respectively. DNA sequence of the 5' breakpoint-clustered region is shown and the 5' breakpoints are marked with blue arrows. The locations of the Y1 and Z1 homology blocks are marked by light purple and light green blocks, respectively.

7.3.4 Possible routes of transmission

SpcDNA derived from chromosomal DNA is usually present in large quantities in mammalian cells (Cohen et al. 1997) and increases with cell stress and aging (Flores et al. 1988; van Loon et al. 1994; Sinclair and Guarente 1997). These circular DNA molecules could be generated by many mechanisms including reverse transcription (Krolewski and Rush 1984), non-homologous recombination (Stanfield and Helinski 1986) and a rolling-circle mechanism (Nosek et al. 2005). Moreover, some of these extrachromosomal circles can carry an autonomously replicating sequence (ARS) derived from chromosomal DNA and can even self-replicate (Cohen et al. 1997). In contrary to these reports, low mutation frequencies of the putative 3.8-kb α1/α2 globin circles hinted that they were unlikely to be replicated by an (unknown) ARS. Despite the evidence of meiotic recombination in this gender reassignment testis sample (section 7.2.1), it is impossible to determine
whether these putative extrachromosomal circles were generated by mitotic recombination or intramolecular exchange during meiosis. In contrast, blood circles, if genuine, must have arisen by an intramolecular recombination pathway during DNA replication. In the absence of a centromere, these mutants would be randomly segregated into daughter cells.

7.4 Conclusions
Intramolecular ectopic exchange, which generates deletions and extrachromosomal circles, has long been proposed to be one of the possible mechanisms giving rise to deletions (Embury et al. 1980; Stankiewicz and Lupski 2002). However, this process has never been proved by the identification of reciprocal deletions and extrachromosomal circles. The work described in this chapter demonstrated how size fractionation in conjunction with inverse PCR strategies could be used to detect extrachromosomal circular DNA molecules arising from intramolecular ectopic exchange in α-globin genes. Although putative mutants were recovered in both testis and blood DNA, it was difficult to prove their authenticity given the low frequencies and dispersed distribution across different size fractions. Besides the identification of these putative 3.8-kb α1/α2 globin circular molecules, a similar portion of putative mutants accompanied by a deletion was also recovered. This indicates that these deleted mutants, if genuine, were most likely generated by other mechanisms such as NHEJ.
Chapter 8
Strategies for Detecting Allelic Recombination Events

8.1 Introduction
Meiotic recombination is a fundamental process that maintains proper chromosome segregation and increases population diversity. It is well documented in yeast and humans that meiotic (allelic) recombination is a non-random event in which DNA strands preferentially exchange at specific regions across the genome (Gerton et al. 2000; Petes 2001; de Massy 2003; Myers et al. 2005). Low-resolution human pedigree studies suggest that recombination frequencies are regionally different between males and females, and that high recombination rates are most likely found in regions with high CpG fraction but low GC content and poly(A)/poly(T) fraction (Broman et al. 1998; Kong et al. 2002). However, recombination hot spots identified so far in human sperm DNA using high-resolution techniques reveal that crossovers cluster within 1–2 kb regions but without predictable DNA sequences (Kauppi et al. 2004; Jeffreys et al. 2005). A recent in silico survey of the HapMap SNP data suggests that there are around 25000–50000 recombination hot spots across the human genome. These hot spots seem to occur near genes but located outside the transcribed domain. Furthermore, this study also reveals that some motifs such as the CCTCCCT motif are significantly over-represented in some recombination hot spots (Myers et al. 2005).

Although the properties of allelic recombination hot spots have been revealed by sperm and population studies, little is known about their influence on ectopic recombination. The first human hot spot characterised by high-resolution sperm DNA analysis showed that the instability of a neighbouring minisatellite MS32 might be controlled by the hot spot, leading to a polarised effect on unequal exchange within the minisatellite repeat array (Jeffreys et al. 1998a). In contrast, the most active autosomal recombination hot spot so far identified, upstream of the β-globin gene in the β-globin gene family, appears not to drive unequal crossover between δ- and β-globin genes since ectopic recombinants have only been detected with an extremely low frequency in sperm, and with breakpoints that tend to avoid the hot spot (Holloway et al. 2006). Previous studies on CMT1A/HNPP diseases have showed that ectopic recombination events between two distant separated CMT1A-REP repeats cluster within a 0.7-kb hot spot (Reiter et al. 1996). More recently, a putative allelic recombination hot spot has been
identified near, but not coincident with, the ectopic hot spot for CMT1A/HNPP rearrangements suggesting that the same hot spot might be responsible for driving both allelic and ectopic exchanges (Lindsay et al. 2006).

Considerable information is now available on processes of α-globin gene deletion and duplication (discussed in Chapters 4 and 6). In contrast, little is known about allelic recombination within the α-globin gene cluster. A total of nine polymorphic markers have been typed in an extensive haplotype study of a ~50-kb region covering the entire α-globin gene cluster. The limited combinations of these markers identified in nine populations suggested that strong linkage disequilibrium exists across the α-globin gene cluster (Higgs et al. 1986). Since only three SNPs across the 12.1-kb α-globin gene region analysed for ectopic recombination have been validated in the International HapMap Project (Gibbs et al. 2003), it is not informative to perform high-resolution LD analysis within this region using this public database. Although strong LD suggests rare historical recombination in the α-globin gene cluster, it does not necessarily mean that recombination hot spots are absent since sperm analysis has revealed hidden hot spots in regions of strong LD in a region of chromosome 1 containing minisatellite MS32 (Jeffreys et al. 2005). Computer simulations suggested that very low recombination rates (<10^{-4} per sperm) can on occasion fail to generate detectable recombinant haplotypes in populations, giving a complete LD block. If allelic crossover rates are similar to meiotic unequal exchange frequencies in the α-globin gene cluster, which are as low as ~10^{-5} per sperm, the existence of allelic recombinant haplotypes will not necessarily be revealed by LD breakdown.

The work in this chapter describes some preliminary genotyping data which will be used for fine-scale analysis of allelic recombination in the α-globin gene cluster, plus some strategic plans for high-resolution sperm analyses. The ultimate goal is to compare allelic and ectopic recombination profiles across the α-globin gene cluster, to determine the relationship, if any, between these two recombination pathways.

### 8.2 Results

#### 8.2.1 LD analysis of the α-globin gene cluster

To determine the level of association between markers located within a DNA region, LD analyses using SNP markers are commonly used. A total of 414 putative SNPs located
within the 12.1-kb α-globin gene cluster region carrying \(\psi \alpha_1\), \(\alpha_2\) and \(\alpha_1\) genes have been reported in the NCBI SNP database. However, only three of them have been validated by genotyping and reported in the International HapMap Project (Gibbs et al. 2003). Since most of these putative SNPs are clustered into the Z homology blocks of the \(\alpha_2\) and \(\alpha_1\) genes, they are unlikely to be genuine SNPs but PSVs instead. As described in Chapter 4 section 4.2.2, 22 genuine SNPs have been identified by semen donor resequencing and genotyped by ASO hybridisation across a panel of 32 semen donors. These data are much more informative than those available in the public database; LD analysis was therefore performed based on these data using a pairwise LD programme written in Perl by A. Webb, Department of Genetics, University of Leicester.

The LD programme measures the strength of association between alleles at pairs of loci using the normalised measure \(D'\) (Lewontin 1988). A simple measure of LD between two loci (A and B) is given as \(D = x_{AIB1} - p_{AI}q_{B1}\), where \(x_{AIB1}\) is the observed frequency of haplotype \(A_1B_1\), and \(p_{AI}\) and \(q_{B1}\) are the observed frequency of alleles \(A_1\) and \(B_1\), respectively. Since \(D\) is heavily dependent on allele frequencies, it is therefore normalised by comparing with the largest value of \(D\) possible given the allele frequencies. \(D'\) is defined as \(D' = D/D_{\text{max}}\), where \(D_{\text{max}} = \min[p_{AI}q_{B1}, (1 - p_{AI})(1 - q_{B1})]\) when \(D < 0\), or \(D_{\text{max}} = \min[p_{AI}(1 - q_{B1}), (1 - p_{AI})q_{B1}]\) when \(D > 0\). Since \(D'\) can be positive or negative depending on the arbitrary labelling of alleles, the absolute \(|D'|\) is used instead. If \(|D'| = 1\), it indicates complete LD between the two loci, with no evidence for historical recombination. If \(|D'| = 0\), it implies free association between the two loci and thus significant historical recombination. The significance of association is estimated from associated likelihood ratio (LR). LR is defined as the ratio of the probability of obtaining the observed genotypes at the maximum likelihood haplotype frequencies to the probability of obtaining the genotypes if markers are in free association. For example, if \(|D'| = 1\) and LR > 10000, it gives excellent evidence in favour of LD rather than linkage equilibrium, whereas if \(|D'| = 1\) and LR < 20, it is not possible to conclude that the corresponding region is in strong LD.

Since demographic factors can significantly influence LD patterns (described in Chapter 1, section 1.2.2) and pooling data from multiple populations can give misleading information on LD profiles, only 20 donors of northern European origin in the 32-donor panel were selected for the LD analysis; leaving only 17 informative SNPs (Fig. 8.1A). The first attempt using all markers revealed a strong LD block across the α-globin gene cluster (Fig. 8.1B). Another analysis by removing markers with low minor allele
frequency (MAF) indicated again very strong LD and with all marker pairs showing a highly significant LR (Fig. 8.1C). However, there is some evidence for localised breakdown, especially for markers in the α2 globin gene region. This incomplete LD breakdown implies the existence of gene conversion and possibly allelic recombination, suggesting that this α-globin gene region might not be completely silent for allelic recombination events.
Fig. 8.1. Pairwise analysis of LD across the α-globin gene cluster. (A) Genotyping data across the α-globin gene cluster from men of northern European origin. A total of 17 loci across the 12.1-kb α-globin gene region were genotyped over 20 men. Homozygous loci are denoted with the corresponding allelic base, ‘A’, ‘C’, ‘G’ or ‘T’, while heterozygous SNPs are indicated by ‘H’. (B) Pairwise analysis of LD across the α-globin gene cluster using 17 SNP markers, marked with blue lines on the sides of the LD graph. The corresponding locations of SNP markers are highlighted with vertical lines along the α-globin gene cluster shown below. The bottom right of the graph represents |D'|, while likelihood ratios in favour of free association are indicated in the top left portion. (C) LD analysis using markers with MAF ≥ 0.2. Only eight markers were left in this analysis. LD breakdown is indicated by an arrow.
8.2.2 Informative SNP searching by resequencing

Earlier haplotype studies (Higgs et al. 1986) and this preliminary LD analysis both indicate that the α-globin gene cluster is located in a strong LD region, but with evidence for some historical allelic recombination suggested in the latter. The ultimate goal is therefore to detect any allelic recombinants generated by haplotype exchange using high-resolution sperm crossover analysis (Jeffreys et al. 1998a). Sufficient numbers of heterozygous SNPs across the target region are essential in this assay not only for designing 5' and 3' allele-specific primers needed for selectively amplifying crossover molecules, but also for fine-resolution mapping of exchange points. The next goal was therefore to identify more informative SNPs across the α-globin gene cluster and to identify informative semen donors for crossover analysis.

8.2.2.1 Screening for α-globin gene rearrangements

Based on the availability of both semen and blood samples, a total of 19 DNA donors with different origins including northern European, Indian and Chinese were chosen for SNP identification by resequencing. These donors were first checked for any α-globin gene rearrangements prior to further characterisation. PCR amplification was performed using nested primers A13.6F plus A25.7R (1° PCR) and A13.6F2 plus A25.6R2 (2° PCR) to detect αα chromosomes and −α3,7, −α4,2 deletion chromosomes. Inverse PCR using primers A22.4F and A21.8R3 designed for detecting the fused α1/α2 gene (3.3 kb) on αααanti3,7 duplication molecules was used to search for ααα chromosomes. Results showed that man 10 was an αα/ααα heterozygote while man 6 was either a −α/−α homozygote or −α/αα heterozygote (Fig. 8.2). These two men were therefore excluded for further analysis. Donors were also tested for the presence of αααanti4,2 duplication by PCR amplification. All of them showed negative results (data not shown). However, since no αααanti4,2 duplication carrier has been identified, this test was performed without a positive control, and thus it is possible that these results were due to PCR failure rather than the absence of αααanti4,2 chromosomes.
**Fig. 8.2.** Screening for α-globin gene rearrangements. (A) Detection of normal or deleted α-globin gene haplotypes by PCR amplification. (i) Structures of progenitor, $-\alpha^{37}$ and $-\alpha^{42}$ deletion molecules, as described in Fig. 4.4. Nested PCR primers are marked with arrows. (ii) Gel electrophoresis of second PCR products visualised by staining with ethidium bromide. PCR amplification on man 25 (marked with an asterisk) failed as shown, but succeeded in a second attempt. (B) Detection of $\alpha_{\alpha\alpha}$, $\alpha_{\alpha\alpha}$ duplication molecules. (i) Structures of progenitor and $\alpha_{\alpha\alpha}$, $\alpha_{\alpha\alpha}$ duplication molecules as described in Fig. 6.3. The divergent primers (arrows) on progenitors become convergent in PCR amplification of duplication molecules. (ii) Gel electrophoretic analysis of PCR amplified duplications visualised by staining with ethidium bromide. A known $\alpha_{\alpha\alpha}$ chromosome carrier (denoted as ' + ') was analysed in parallel as a positive control. M, λ DNA × HindIII.
8.2.2.2 Additional SNPs identified in Z homology blocks

Since the SNP density in both 3' ends of Z homology blocks was low, these two regions were therefore PCR amplified and searched for extra SNPs by resequencing (Fig. 8.3). Four more SNPs were detected in these two regions but each showed low MAF, and were only found in one or two heterozygous men (Fig. 8.3B). In total, there were 21 polymorphic loci along the 12.0-kb region. However, only a minority were heterozygous in any given man. For example, based on both ASO hybridisation data (shown previously in Fig. 4.2) and resequencing data of 3' regions of Z homology blocks, the most informative man (no. 25) carried 12 heterozygous SNPs, while only four heterozygous SNPs were identified in man 21 (Fig. 8.35). In addition, there were very few heterozygous SNPs in the 3' end of the cluster, limiting the choice for the design of allele-specific primers for crossover analysis. Since flanking markers are essential for crossover analysis (Jeffreys et al. 1998a), the focus then moved to a search for more informative SNPs further downstream from the α-globin genes.
Fig. 8.3. SNP detection by resequencing. (A) Regions amplified for resequencing using nested PCR. The structure of the progenitor as described in Fig. 3.1 with Z homology blocks indicated. PCR primers are marked with arrows. Direct sequencing was performed on the selected regions of Z homology blocks, marked by brackets (B) SNP location on the 12.0-kb progenitor, marked by red lines. Genotypes below are a combination of ASO hybridisation genotyping data (Fig. 4.2) and resequencing data. Heterozygous SNPs are highlighted in red, while unknown data on the new donors (men 33-37), who were only genotyped by resequencing of the selected regions on Z homology blocks, are marked with ‘?’. New SNP loci identified by resequencing are marked by rectangles. K = G + T, M = A + C, R = A + G, S = C + G, Y = C + T.

8.2.3 Additional SNPs detected in the 01 gene region

Only those men who had at least seven heterozygous SNP sites across the 12.0-kb progenitor were further analysed. To obtain more informative 3’ flanking markers, a 4.2-kb region including the 01 gene was amplified (Fig. 8.4). SNP detection was performed by resequencing this amplicon from 11 men including man 1 and man 2 who have already been characterised for de novo deletions and duplications (see Chapters 4 and 6). Promisingly, four more variable loci were found; the first two SNPs have not yet been reported, while the last two were dbSNPs (Fig. 8.4). The last three SNPs were clustered into a 124-bp region in the first exon of the 01 gene. This region is GC-rich and is good for designing allele-specific PCR primers for crossover analysis. Based on heterozygous
SNP abundance and the availability of both 5' and 3' flanking heterozygous SNPs, man 2 and man 25 seem to be the best candidates for crossover analysis.

Fig. 8.4. SNP identification in the 3' extended region. (A) The structure of the α-globin gene cluster including ψα1, α2, α1 and θ1 genes. The 3' extended region for SNP searching was PCR amplified using primers A24.7F and A28.9R. The four newly identified SNPs are marked with white circles. (B) Heterozygous SNP distribution across the α-globin gene cluster in 11 men. SNP locations, and heterozygosities in each man are marked by black and red lines, respectively.

8.3 Discussion

8.3.1 Strategies for high-resolution sperm analysis
To date, 27 human recombination hot spots have been characterised using fine-scale strategies including LD analysis and PCR-based crossover analysis by this laboratory (details in Chapter 1, section 1.2.5). Crossover assays guided by historical recombination information from in silico analyses of genotype data using LD mapping and coalescent analysis were designed to detect allelic recombinants across the location of a putative hot spot. Crossover breakpoints can be mapped, depending on the density of markers, to a high-resolution scale, for instance less than 0.5 kb (Kauppi et al. 2004). Although the α-globin gene cluster is obviously in a strong LD region, allelic recombination might occur within the cluster given the evidence that hidden hot spots have been detected in LD blocks (Jeffreys et al. 2005). Since genotyping gave no clues about the possible location
of any hot spots in the α-globin gene cluster, the strategy planned for crossover analysis was to run across both α-globin genes (α2 and α1) using allele-specific primers directed to markers located in the ψα1 region as well as in the θ1 gene (Fig. 8.5). Although the α-globin gene cluster with two GC-peaks (12.1 kb) has been efficiently amplified using the optimised PCR conditions (described in Chapter 3), it is not necessarily the case that high amplification efficiency will also be achieved in the allele-specific crossover assay. As described in Chapter 3, the yield of a 3.4-kb amplicon was greatly improved when extension time increased from five minutes to 15 minutes (Fig. 3.4B). However, the longer the PCR extension time, the lower is the allele specificity (Arnheim et al. 2003). It might therefore be difficult to amplify the targeted GC-rich region (11.0 or 11.8 kb) in crossover analysis with high efficiency and the high level of allele specificity needed to selectively amplify crossover molecules.

Alternatively, these potential problems can be overcome by dividing the entire target into two overlapping amplicons (Fig. 8.5B). With shorter amplicons and the presence of only one GC-peak, allele-specific PCR with high specificity should be much more easily achieved.
Fig. 8.5. Strategies for crossover analysis. (A) Crossover assay for the α-globin gene region. The structure of the α-globin gene cluster with ψα1, α2, α1 and θ1 genes as described in Fig. 8.4. SNPs on haplotypes A and B are indicated by red and blue lines, respectively. Both types of crossover recombinants (AB- and BA-types) can be selectively recovered by two rounds of allele-specific PCRs. Allele-specific primers are shown as red and blue arrows. (B) Breaking up amplicons for alternative crossover analysis. Recombinants can be recovered by amplifying two overlapping amplicons using allele-specific primers. However, recombinants exchanged outside the target region would be missed, but the overall crossover profile can be obtained by combining data on both test intervals.

8.3.2 Strong LD block and low density of heterozygous SNPs in α-globin genes

Haplotype analysis has showed that the α-globin gene cluster is a polymorphic locus where nine polymorphic markers have been found within the 50-kb region (Higgs et al. 1986). However, the few haplotypes identified in different populations indicated that these markers are in strong association with each other. In addition, the only two markers located between the ψα1 and α1 genes were highly associated with each other on most haplotypes, suggesting the rarity of historical recombination within this region. This finding is supported by pairwise analysis of LD using more polymorphic markers, as described in section 8.2.1.

The major obstacle for identifying genuine SNPs for association studies was the miscalling of PSVs as SNPs in the highly similar homology blocks within the α-globin gene region. Sequence differences between homologous regions are believed to be the
evolutionary mark of the divergence of two α-globin genes from one common ancestral
gene, and are minimised by concerted evolution (Michelson and Orkin 1983; Higgs et al.
1984). Although these processes have generated differences between the α-globin genes,
they appear to be functionally normal although more α2 transcripts than α1 transcripts are
detected in normal individuals (Orkin and Goff 1981). Interestingly, the density of SNPs
within the two α-globin genes is very low, with only one SNP in each gene based on
genotyping 37 individuals with different origins; the α2 gene SNP is located within an
intron whereas the second SNP in the α1 gene causes a synonymous codon mutation in
the third exon (codon 93) (Fig. 4.2). This rarity sharply contrasts with the SNP cluster in
the first exon of the θ1 gene; the first two SNPs locate in the 5′ untranslated region and
the last one is a synonymous codon mutation (Fig. 8.4). Since the function of the θ1 gene
is unknown and heterozygous or homozygous deletion of this gene shows no apparent
abnormality (Clegg 1987; Hsu et al. 1988), it raises a possibility that the rarity of SNPs in
the α-globin genes is due to purifying selection, eliminating variants within these
important genes.

8.3.3 Recombination motif analysis
The global analysis of recombination rates and hot spots suggested that a 7-nucleotide
oligomer CCTCCCT is more frequent in hot spots with THE1A/B elements than in other
THE1A/B elements distributed elsewhere in the human genome (Myers et al. 2005). This
finding was strikingly coincident with the polymorphic regulatory activity of the same
sequence motif proposed in the hot spot DNA2, in which a transition from T to C in the
third position of the motif suppresses recombinational activity (Jeffreys and Neumann
2002). Interestingly, a 7-mer sequence CTCCTCC, which is not associated with any
THE1B element but related to the proposed motif, has been identified in both proximal
and distal CMT1A-REPs within the newly located allelic/non-allelic recombination hot
spot (Lindsay et al. 2006). More recently, an extended 13-bp hot spot motif
CCTCCCTNNCCAC, which determines the location of at least 40% human crossover hot
spots in both sexes, has been proposed (Myers et al., unpublished data). Sequence
analysis of the α-globin gene cluster indicates that the first two motifs but not the
extended one commonly occur within the cluster and overlap mostly with each other
(Fig. 8.6). However, although these motifs are frequently found within the target region,
ye give no clues about the presence of allelic recombination hot spots or their potential
locations. These issues can only be addressed by crossover analysis in men with informative markers.

![Diagram of α-globin gene cluster with proposed recombination motifs](image)

**Fig. 8.6.** Distribution of proposed recombination motifs across the α-globin gene cluster. The locations of the motif CCTCCCT and the related motif CTCCTCC are marked with red and blue lines, respectively below the structure of the α-globin gene cluster.

### 8.4 Conclusions

Despite substantial studies on ectopic recombination between the two highly similar α-globin gene regions, allelic recombination within this locus has never been analysed in depth. The LD analysis described in this chapter agrees with earlier observations of limited haplotype diversity and shows that the α-globin gene cluster is embedded in a strong LD block. Given the evidence of hidden hot spots in strong LD regions, preliminary steps have been taken for designing a high-resolution crossover assay including selecting donors with normal α-globin genes, discovering more SNPs and choosing informative men. With two informative men selected, issues including the existence of allelic recombinants in human sperm and the location of possible recombination hot spots are ready to be addressed.
Chapter 9
Final Discussion

9.1 Introduction
DNA instability in the human genome caused by processes like point mutation, transposition, and homologous and non-homologous recombination creates variation between individuals. Although every individual shares roughly 99.9% sequence identity (Reich et al. 2002; Feuk et al. 2006), phenotypical differences are obviously observed. Recently, increasing evidence suggests that individual differences are not solely governed by the local difference of DNA sequences but also by structural variation and post-transcriptional factors (Sebat et al. 2004; Mata et al. 2005; Tuzun et al. 2005; Feuk et al. 2006). Surveys of copy number variation have revealed that the number of deletions/duplications between individuals can be variable at different loci (Sharp et al. 2005; McCarroll et al. 2006), while some of these variations can lead to genetic disorders. However, mechanisms for controlling these variations are largely unknown.

In this thesis, the α-globin gene cluster has been chosen as a representative for the study of DNA stability between locally repeated DNA sequences in the human genome. This cluster is a classic study model for unequal crossover between local DNA repeats driven by the presence of highly similar homology blocks. In previous chapters, the instability of this locus has been analysed by investigating the dynamics and processes of ectopic recombination. In addition, the remarkable contrast between de novo germ-line instabilities and the population incidence of α-globin gene rearrangements implies that selective constraints have acted in maintaining normal αα chromosomes in populations. In this chapter, the focus will be on discussing prominent features of instability in the human α-globin gene cluster and of the abnormally rearranged α-globin gene recombinants reported in previous chapters. In addition, future directions for exploring other recombination events occurred within this region will also be discussed.

9.2 Relationship between GC-content and recombination
Studies of the yeast S. cerevisiae suggest that meiotic recombination hot spots can be categorised into three types, namely α-, β- and γ-hot spots (Petes 2001) (see Chapter 1, section 1.1.6.3 for details). Among these hot spots, γ-hot spots are classified in relation to high GC-content. In addition, global mapping of meiotic recombination hot spots and
cold spots in *S. cerevisiae* showed that although no prominent features are shared among all of the hot spots, a strong association exists between hot spots and GC-peaks (Gerton et al. 2000). A similar phenomenon is also observed in the human genome, in that there is a positive though weak relationship between recombination and GC content (Fullerton et al. 2001; Yu et al. 2001). One of the possible explanations for this correlation is that GC-rich regions are prone to be open to recombination machinery, for example by creating open chromatin structures in contrast to the cohesin-bounded AT-rich regions (Gerton et al. 2000). However, this explanation would only be valid if recombination rates are determined by the differences of GC-content across the genome, since the direction of causation can be the reverse, with the genome composition instead controlled by patterns of recombination (Fullerton et al. 2001).

In contrast, the high-resolution recombination map of the human genome established by Kong et al. (2002) indicates that GC content is only positively correlated with recombination rates when it is assessed separately from other parameters; regions with high recombination rates were predicted to be those with high CpG fraction but low GC-content and poly(A)/poly(T) fraction when other parameters were considered. Likewise, there is no significant evidence from human hot spots that a high level of recombination activity requires high local GC-content (de Massy 2003; Kauppi et al. 2004).

As mentioned in Chapter 1 section 1.4.1, the GC-content of the human α-globin gene cluster is higher than that of the genome average and the region is associated with CpG islands. A study of the short arm of human chromosome 16 using CEPH data indicates that the 2Mb region has a higher recombination rate than the genome average (Daniels et al. 2001). However, no detailed information about the recombination frequency within the α-globin complex is available due to limitations of pedigree data and since microsatellite markers used in this analysis were widely distributed beyond the cluster. In contrast, as described in Chapters 4 and 6, most ectopic recombinants exchanged within the longest uninterrupted identical sequences within Z homology blocks in a region where two GC-peaks are located. However, a comparison between this region and other shorter identical sequences within Z homology blocks reveals that ectopic recombinants are in fact distributed fairly randomly along homology blocks. There is no evidence suggesting that the GC-peaks are hot spots for ectopic recombination. Further investigation on this issue would require donors with more informative markers within the homology blocks.
9.3 Reciprocity of deletion and duplication

With completion of the Human Genome Project, studies of DNA variability in the human genome have recently shed much more light on copy number variation and other genome rearrangements (Redon et al. 2006; Wong et al. 2007). Clinical analyses have shown that many genomic disorders associated with these DNA rearrangements are relatively high-frequency recurrent events. For example, a deletion on 22q11.2, causing DiGeorge syndrome/velocardiofacial syndrome (DGS/VCFS), is prevalent at birth of 1.3–2.5/10000 (Goodship et al. 1998; Inoue and Lupski 2002), while the frequency of Smith-Magenis syndrome (SMS) is about 1/25000 (Greenberg et al. 1991). Ectopic recombination/NAHR, giving deletion and duplication, is the most widely recognised mechanism for generating some of these rearrangements. Since not all deletions and duplications give rise to genetic disorders, reciprocal ectopic recombination cannot be proved directly by most clinical studies. CMT1A and HNPP disorders are one of the most credible examples of disease-causing duplications and deletions driven by ectopic recombination (Chance et al. 1994). Although both of these disorders have been found by clinical studies and share a number of common features such as locations of recombination hot spots and exchange breakpoints, it is not necessarily true that they arise from a fully reciprocal process generating both deletions and duplications at equal frequency.

Single α-globin gene deletions are far more commonly identified than ααα duplications in populations. It is not only because some major forms of α⁺-thalassaemia give clinical abnormalities (Wang 2000), biasing the ascertainment of –α chromosomes, but also that –α chromosomes are favoured by malaria selection (Flint et al. 1986a). If deletions and duplications are both generated strictly by ectopic recombination in a single process (an inter- but not intramolecular pathway), full reciprocity of the two de novo recombinants should be observed (assuming absence of any form of germ-line selective force). Analyses of de novo –α3.7 deletions and ααααα3.7 duplications in the male germ-line reported in this thesis have provided the first direct evidence for a very substantial level of reciprocity of ectopic exchange in α-globin genes. Although evidence for mutational mosaicism, which can significantly alter the frequencies of specific classes of recombinants, has been observed in intrachromosomal exchanges in both blood and sperm, strong evidence of full reciprocity with respect to similar recombination frequencies, haplotype symmetry and reciprocal exchange breakpoints have been shown instead for interchromosomai deletions and duplications.
Despite evidence suggesting that interchromosomal deletions and duplications are generated in a reciprocal fashion, the possibility that a portion of these recombinants are generated by other non-reciprocal mechanisms cannot be excluded. As reported in Chapter 7, there is evidence for the existence of putative extrachromosomal circles, suggesting that some de novo deletions might be produced by a non-reciprocal intramolecular ectopic recombination pathway. In addition, as described in Chapter 1, deletion and duplication molecules can be generated by SSA and SDSA mechanisms, respectively. Products arising by these mechanisms are indistinguishable from ectopic recombinants. However, these mechanisms are unlikely to be the predominant players, since it would require SSA and SDSA occurring at similar and high frequencies to generate products with apparently reciprocal exchange breakpoints. In contrast, a small number of mutants accompanied by a deletion, identified in analyses of −α deletions, ααα duplications, and excised α1/α2 fusion circles, might indicate the involvement of other mechanisms. Despite the fact that these mutants might be jumping PCR artefacts generated at the onset of PCR, they could be genuine mutants arising from the joining of two broken DNA molecules by NHEJ (see section 9.7).

9.4 Intra- vs. interchromosomal ectopic exchanges

Studies in S. cerevisiae show that sister chromatids rather than homologous chromosomes are preferred for recombinational repair in mitotic cells (Kadyk and Hartwell 1992). This preference is also seen in mammalian cells, with recombination between sister chromatids being preferred by two to three orders of magnitude over non-sister chromatids (Moynahan and Jasin 1997). Since sister chromatids are identical to each other, this preferential route is thought to be a mechanism for maintaining genome integrity without any loss of heterozygosity during repair of DSBs. In contrast, non-sister chromatids are instead preferred in meiotic recombination (Schwacha and Kleckner 1997).

De novo α-globin gene deletions and duplications reported in this thesis seem to fit well with this sister-chromatid-biased recombinational repair observed in yeast and mammalian cells. As described in Chapters 4 and 6, almost all blood deletions and duplications were intrachromosomal exchanges, while only a small fraction of recombinants with exchange of flanking markers was observed in the germ-line. Although interchromosomal exchanges were almost certainly meiotic in origin, it is impossible to distinguish meiotic intrachromosomal recombinants from pre-meiotic ectopic exchanges in sperm. However, the significant level of putative mutational
mosaicism observed in both sister chromatid exchanges in blood and sperm implies that the majority of intrachromosomal deletions and duplications in the germ-line occur prior to meiosis, and are instead mitotic in origin. In contrast, given the evidence of sex-dependent mechanisms proposed in the studies of CMT1A/HNPP disorders (Lopes et al. 1997; Lopes et al. 1998), it is unknown whether this biased recombination repair is a universal difference between mitosis and meiosis in humans.

9.5 Factors affecting ectopic recombination

Studies in both yeast and mammalian cells agree that a minimal efficient processing segment (MEPS) is essential for efficient ectopic recombination between homologous regions (Rubnitz and Subramani 1984; Waldman and Liskay 1988; Jinks-Robertson et al. 1993). Similar arguments have also been proposed in humans such as in the studies of the human β-globin gene cluster (Metzenberg et al. 1991) and CMT1A/HNPP disorders (Lopes et al. 1998; Reiter et al. 1998) in which exchanges were found to cluster mainly in the longest stretches of identical sequence between misaligned homologous regions. Although most exchange breakpoints in \( -\alpha^{3,7} \) deletions and \( \alpha\alpha\alpha^{anti3,7} \) duplications cluster in the longest uninterrupted identical sequence (855 bp) in Z homology blocks (~1.7 kb), there is no strong evidence to support the existence of a MEPS (or the MEPS is very short). Unlike recombination events reported in mammalian cells (Rubnitz and Subramani 1984; Waldman and Liskay 1988), there is no dramatic drop of interchromosomal recombination frequency in uninterrupted sequences smaller than the 855 bp within the Z homology blocks (Fig. 6.7). Mutants with haplotype-specific SNPs, which acted as PSVs, have shown that interruption of the longest region of identical sequence does not create any drop in the efficiency of interchromosomal unequal crossover (see Chapter 4, section 4.3.4). However, these studies showed only the relationship between ectopic exchange resolution breakpoints and sequence identity, and gave no clues about where recombination is initiated. These issues can be tackled by either recovering recombination intermediates or detecting gene conversion events to see if they cluster into a hot spot.

In contrast, exchanges in the longest uninterrupted identical sequence (338 bp) in X homology blocks (~1.3 kb) in \( -\alpha^{4,2} \) deletions are very rare comparing to those in Z homology blocks in \( -\alpha^{3,7} \) deletions (see Chapter 4, section 4.3.5 and Fig. 4.8). These results indicate that the efficiency of ectopic recombination must be controlled not just by homology block length but also by other factors such as chromatin topology. Studies of mitotic and meiotic ectopic recombination using artificial inserts in yeast have shown that
the physical distance between repeats is important in controlling the efficiency of ectopic recombination (Lichten et al. 1987; Lichten and Haber 1989). The frequency of ectopic recombination between inserts on homologous chromosomes is negatively correlated with the physical distance between insert loci (Goldman and Lichten 1996). Furthermore, artificial inserts on homologous chromosomes can recombine as efficiently as alleles, whereas the efficiency of ectopic recombination between repeats on heterologous chromosomes is significantly lower. These findings are consistent with the suggestion that the smaller nuclear volume occupied by sister chromatids relative to homologues may trigger sister chromatid exchange over recombination between homologous chromosomes in mammalian cells during mitosis (Johnson and Jasin 2000).

9.6 Gene conversion in mitotic and meiotic recombination

According to the DSB repair model (Szostak et al. 1983), gene conversion associated with crossover can be generated by copying a segment of DNA sequences from a donor template during mismatch repair of heteroduplex DNA. However, studies in yeast and mammalian cells (Paques et al. 1998; Johnson and Jasin 2000) have shown that mitotic ectopic gene conversion is not usually associated with crossover, suggesting that the DSB repair model is not the only mechanism for DSB repair. Similarly, an analysis of homologous recombination between CYP21 and CYP21P genes in humans reveals that deletions caused by unequal crossover between these two genes appear exclusively in meiosis, whereas gene conversion unassociated with crossover occur mainly in mitosis (Tusie-Luna and White 1995). This study suggests that gene conversion and unequal crossover operate differently by distinct mechanisms.

Since the goals of the experiments reported in previous chapters were to detect de novo α-globin gene rearrangements, the topic of gene conversion unassociated with unequal crossover was left completely untouched (see section 9.9). The only definite evidence for gene conversion in these studies was the complex events accompanied by unequal exchange. In total, only 19 of these mutants were detected from the entire analyses of deletions, duplications and putative extrachromosomomal circles in both blood and sperm (~1300 mutants in total). Almost all of these complex mutants were intrachromosomal exchanges (18 out of 19) and most conversion sites of these mutants (17 out of 19) were within 300 bp of the site of ectopic exchange. The rarity of these patchy gene conversions (1.5%) is comparable to their incidence during allelic crossover at recombination hot spots (~1%) including the minisatellite MS32-associated hot spot.
(Jeffreys et al. 1998a), the TAP hot spot (Jeffreys et al. 2000) and the MHC hot spots (Jeffreys et al. 2001). These observations suggest that these complex ectopic exchanges were most likely generated by the same pathway as for simple exchanges, but with patchy conversions arising through patchy mismatch repair of heteroduplex DNA during DSB repair.

9.7 Possible mechanisms for exceptional mutants

In the entire analyses of \textit{de novo} deletions, duplications and putative extrachromosomal circles described in previous chapters, a total of 19 exceptional mutants were detected. These mutants were all smaller than expected, and were accompanied by a deletion ranging from 0.3–6.5 kb. Given the fact that most of them were found in duplications and putative α1/α2 fusion circular DNAs, they would be unlikely to be identified in clinical studies due to the fact that these deletions, in principle, would not give any deleterious effects in carriers or that they would be lost if they are in circular DNAs. However, the authenticity of these deletion mutants is in question and some might possibly be PCR artefacts instead.

If these \textit{de novo} mutants are genuine, then simple unequal crossover is obviously not a pathway for explaining their occurrence. In fact, sequencing shows that these mutants can be classified into three groups (Fig. 9.1B). For the first group, either one or both deletion points are located in a non-homologous region, leading to a deletion of part or all of a homology block (Fig. 9.1Bi). Since there is no homology between 5' and 3' breakpoints, the most possible mechanism for generating these mutants is NHEJ in which little or even no homology is needed for joining two broken DNA molecules (Moore and Haber 1996). For the second group, the deletion is embedded within a fused homology block with again no homology identified between 5' and 3' breakpoints (Fig. 9.1Bii). These mutants can possibly be generated by NHEJ (Fig. 9.1C) or by 'aberrant' unequal crossover (Fig. 9.1D). A shift between misaligned homology blocks might promote ectopic exchange in a region with a lower degree of homology (or no homology) and generate a deletion. However, if this 'aberrant' unequal crossover model is feasible, it raises the question of why strand exchange is preferred in regions of little homology instead of in the 'proper' misaligned regions where sequence similarity is maximal. Finally, the last group of mutants show a clear unequal exchange breakpoint with the deletion located within the fused homology block but away from the site of unequal exchange. These mutants can be produced by unequal crossover in conjunction with
hairpin looping within a homology region on a donor template (Fig. 9.1E). The looped region is lost in the newly synthesised DNA strand; if this strand is used for mismatch repair of heteroduplex DNA, thus the deletion mutant is generated. However, this model cannot easily be used to explain the formation of the deleted mutant with a 7-bp insert as described in Chapter 6, section 6.2.6.4.

However, if these exceptional mutants are genuine and explicable by these proposed models, then it is expected that a similar number of ectopic recombinants should be accompanied by a DNA insertion. However, none of them were duplicated mutants, indicating a bias in favour of deletions (19 deleted vs. 0 duplicated mutants, $P < 0.001$). This issue again raises the question of the authenticity of these mutants and, perhaps, of the existence of other mechanisms for generating these non-reciprocal mutants.
Fig. 9.1. Mechanisms for exceptional mutants. (A) Structures of molecules with duplicated regions (progenitors), ectopic deletions and duplications/excised fusion circles. Homology blocks are highlighted in red and blue. (B) Classification of mutants accompanied by deletion as described in Chapters 4, 6 and 7. (i) Mutants with one or both deletion breakpoints located in a non-homologous region. (ii) Mutants with a deletion embedded within a fusion homology block. (iii) Mutants with unambiguous ectopic exchange breakpoints but carrying a deletion in the fused homology block. The mutant accompanied with a deletion and a 7-bp insert, as described in Fig. 6.10, is marked with an asterisk. (C) Joining of two broken DNA molecules by NHEJ. A fusion homology block can be formed by NHEJ if DSBs occurred within homology blocks (red and blue empty boxes). (D) An 'aberrant' unequal crossover model. Deleted mutants can be caused by this model as a result of a shift between misaligned homology blocks. (E) A modified unequal crossover model with a hairpin looping out on the donor template. Deleted mutants can be produced using the newly synthesised template (bottom dotted line) for mismatch repair of heteroduplex DNA.
9.8 Dynamics for maintaining normal \(\alpha\alpha\) chromosomes in populations

As described in previous chapters, the incidence of \(\alpha\) and \(\alpha\alpha\) chromosomes is consistently found to be very low in most malaria-free populations (Higgs et al. 1989; Hill 1992). The rate of generation of \textit{de novo} \(\alpha\) and \(\alpha\alpha\) chromosomes in the germ-line was therefore expected to be low to give a balanced population incidence of rearranged chromosomes. In contrast, studies of \textit{de novo} \(\alpha\) and \(\alpha\alpha\) chromosomes in the male germ-line indicate that these rearrangements are surprisingly frequent, providing evidence that the low incidence of rearranged chromosomes can only be maintained if significant selection acts against both \(\alpha\) and \(\alpha\alpha\) chromosome carriers and/or homozygotes. However, no significant reduction of fitness has yet been reported in these carriers.

With these remarkable frequencies of chromosome rearrangement, it is expected that the copy number of \(\alpha\)-globin genes on each chromosome would be highly unstable in the absence of selection. Assuming each copy of an \(\alpha\)-globin gene can recombine with a pairing partner at random within an \(\alpha\) array, it is expected that \(\alpha\) chromosomes would act as an absorption boundary leading eventually to fixation in a population since copy number variation can only appear if a \(\alpha\) chromosome interacts with another chromosome carrying more than two copies of \(\alpha\)-globin genes. Computer simulations using an effective population size \((N_e)\) of 10000 across 40000 generations (~1 million years) and with the mean observed deletion/duplication rates starting from \(\alpha\alpha/\alpha\alpha\) homozygotes showed that the copy number of \(\alpha\)-globin genes is chaotic across generations, without fixation of \(\alpha\) chromosomes within that period of time (Fig. 9.2A). A more constrained stepwise mutation model (Fig. 9.2B) involving only adjacent recombination partners (for example, constrained misalignment controlled by flanking sequence which generates \textit{de novo} \(\alpha\alpha\alpha\alpha\) plus \(\alpha\alpha\) instead of \(\alpha\alpha\alpha\alpha\alpha\) plus \(\alpha\) chromosomes in \(\alpha\alpha/\alpha\alpha\alpha\alpha\) homozygotes) gives similar results as the unconstrained one. These results again support the argument that selective constraints must have acted to maintain the steady and very high incidence of normal \(\alpha\alpha\) chromosomes in most populations, with an accompanying very low incidence of other rearranged chromosomes.
Fig. 9.2. Computer simulations of copy number variation of α-globin genes in populations. The simulation programme was written in True BASIC 4.1 by A. Jeffreys, assuming intra- and interchromosomal ectopic frequencies as $3 \times 10^{-5}$ and $1 \times 10^{-5}$ per sperm, respectively, and an effective population size ($N_e$) as 10000. (A) Simulation results for the unconstrained unequal crossover model, with three simulations. The α-globin genes are assumed to recombine with pairing partners at random within an αn array in this model. Simulations across 40000 generations are shown only. (B) Three examples of constrained simulations. The α-globin genes are only allowed to misalign with a shift of one repeat.
9.9 Future directions

The dynamics and processes of ectopic recombination within the α-globin gene cluster have been revealed by the study of specific classes of *de novo* recombinant, especially deletions, duplications and putative extrachromosomal circles. However, the relationship between ectopic and allelic recombination is still unclear since very few high-resolution studies have been done on a locus where both ectopic and allelic recombination occur. The next goal is therefore to investigate allelic recombination in the α-globin gene cluster. As described in Chapter 8, preliminary data have shown that it should be possible to perform direct analyses using high-resolution crossover assays for detecting any potential allelic recombinants in sperm (Jeffreys et al. 1998a). Recombinants will be characterised using ASO hybridisation and DNA sequencing to map, if present, the location of any recombination hot spot and to correlate allelic exchange distributions with the exchange breakpoints of ectopic recombinants.

The investigation of gene conversion will be another direction for studying DNA instability in this gene cluster. Gene conversion can occur, in principle, in both allelic and ectopic recombination (Fig. 9.3). Allelic gene conversion can be detected using DNA enrichment by allele-specific hybridisation (DEASH) (Jeffreys and May 2003; Jeffreys and May 2004). Similarly, a modified method could be developed using biotinylated paralogue-specific oligonucleotides (PSOs) instead of ASOs for detecting ectopic gene conversion events not associated with unequal crossover (Fig. 9.3B).

Finally, the screening of −α deletions and αααα3.7 duplications in the 19-donor panel (Chapter 8 section 8.2.2.1) indicates that two donors are −α or ααα chromosome carriers. If these donors proved to be informative for extensive analyses, it would be very interesting to study the interactions between normal αα and rearranged −α or ααα chromosomes (Fig. 9.4). The dynamics and processes can be studied by characterising *de novo* −α/αα/ααα mutants to see if there are any differences from those reported in previous chapters. More importantly, the dynamics for controlling copy number variation within this gene family can be better elucidated by studying *de novo* −α/αααα chromosomes generated in αα/ααα heterozygotes (Fig. 9.4Bii).
Fig. 9.3. Strategies for detecting allelic and ectopic gene conversion events. (A) DEASH for detecting allelic gene conversion events. (i) Two haplotypes of a man with two homology regions (red and blue blocks) and SNP heterozygosities (white and black circles) are shown. (ii) Gene conversion events can be recovered by two rounds of nested PCR using allele-specific primers (grey arrows) and universal primers (black arrows) after selection by hybridisation with a biotinylated ASO (bio-ASO) specific to the target SNP. (B) Modified strategy for detecting ectopic gene conversion events. (i) Homology block misalignment of two molecules. The target homology block is cleaved out by restriction enzymes (green arrows) prior to further analysis. (ii) Biotinylated PSO (bio-PSO) specific to the target PSV is used for selecting gene conversion events, which can be recovered by nested PCR using universal primers (black arrows).

Fig. 9.4. Ectopic recombination between normal $\alpha\alpha$ and rearranged $-\alpha/\alpha\alpha\alpha$ chromosomes. (A) Possible recombinants generated by ectopic recombination between $\alpha\alpha$ and $-\alpha$ chromosomes. The $\alpha_2$ and $\alpha_1$ genes are represented by red and blue blocks, respectively. A fusion gene is marked as a mixed colour block. (B) Possible recombinants from ectopic recombination between $\alpha\alpha$ and $\alpha\alpha\alpha\alpha$ chromosomes. (i) Mechanisms producing $\alpha\alpha$ and $\alpha\alpha\alpha$ recombinants. (ii) $-\alpha$ and $\alpha\alpha\alpha\alpha$ chromosomes can be produced in an $\alpha\alpha/\alpha\alpha\alpha\alpha$ heterozygote, giving copy number variation.
## Appendix I

### Oligonucleotides for PCR amplification and/or sequencing:

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence</th>
<th>No. of GC/Primer length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A13.6F</td>
<td>5' TTTCCGCTGAGACCTTGCCG' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A13.6F2</td>
<td>5' CCGCGAAATGTCGAGCCCG' 3'</td>
<td>13/18</td>
</tr>
<tr>
<td>A13.7F</td>
<td>5' AGTGCTTGGCTGCTTGC' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A15.0R</td>
<td>5' TCTGGCAATCAGGGGCTGG' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A15.5R</td>
<td>5' AGAGCGACGAGCCCTTGCG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A15.6F</td>
<td>5' CTCTGAGAGCTGGTCTCGG' 3'</td>
<td>11/19</td>
</tr>
<tr>
<td>A15.7F</td>
<td>5' AGGGCACTTGGCTCACACCG' 3'</td>
<td>12/18</td>
</tr>
<tr>
<td>ASPA15.9F</td>
<td>5' TCTCGGAAGAAGTCGCGCA' 3'</td>
<td>9/18</td>
</tr>
<tr>
<td>ASPA15.9GF</td>
<td>5' TCTGGCAATCAGGGGCTGG' 3'</td>
<td>10/18</td>
</tr>
<tr>
<td>ASPA16.2F</td>
<td>5' GCCACATCGAGAAGCGACTTG' 3'</td>
<td>11/21</td>
</tr>
<tr>
<td>ASPA16.6GF</td>
<td>5' CGCAAGATGTCGACACACCG' 3'</td>
<td>11/18</td>
</tr>
<tr>
<td>ASPA16.8TF</td>
<td>5' CCGGGAATTGTCGCAACAT' 3'</td>
<td>10/20</td>
</tr>
<tr>
<td>ASPA17.0F</td>
<td>5' AACTGGCGCTCTCAACAGG' 3'</td>
<td>9/18</td>
</tr>
<tr>
<td>ASPA17.0GF</td>
<td>5' TCTGCAATCAGGGGCTGG' 3'</td>
<td>10/18</td>
</tr>
<tr>
<td>A17.3F</td>
<td>5' AGTCTGAGATGACGTGTGG' 3'</td>
<td>11/20</td>
</tr>
<tr>
<td>A18.1F</td>
<td>5' AGAGCGACGAGCCCTTGCG' 3'</td>
<td>11/20</td>
</tr>
<tr>
<td>A18.1R</td>
<td>5' CCGGTTCCTGGCTCAAGTC' 3'</td>
<td>11/19</td>
</tr>
<tr>
<td>A18.2F</td>
<td>5' ACACTGCACTGCACTACGCG' 3'</td>
<td>11/20</td>
</tr>
<tr>
<td>A18.2R</td>
<td>5' GAGACCTGATATGAATCTGCG' 3'</td>
<td>11/20</td>
</tr>
<tr>
<td>A18.6R</td>
<td>5' GATGCGACGAGATAGGTGTC' 3'</td>
<td>11/19</td>
</tr>
<tr>
<td>A18.9F</td>
<td>5' AGGGCACTTGGCTCACACCG' 3'</td>
<td>10/18</td>
</tr>
<tr>
<td>A19.1F</td>
<td>5' CTCTGCACTAAGTCGACCAACCG' 3'</td>
<td>11/18</td>
</tr>
<tr>
<td>ASPA19.3CF</td>
<td>5' ACTGGCGCTCTCAACAGG' 3'</td>
<td>11/18</td>
</tr>
<tr>
<td>ASPA19.3TF</td>
<td>5' CCGGGAATTGTCGCAACAT' 3'</td>
<td>10/20</td>
</tr>
<tr>
<td>ASPA19.4AF</td>
<td>5' TCTGCAATCAGGGGCTGG' 3'</td>
<td>10/18</td>
</tr>
<tr>
<td>ASPA19.4GF</td>
<td>5' AACTGGCGCTCTCAACAGG' 3'</td>
<td>9/18</td>
</tr>
<tr>
<td>ASPA19.4AR</td>
<td>5' AGGGCACTTGGCTCACACCG' 3'</td>
<td>11/18</td>
</tr>
<tr>
<td>ASPA19.4GR</td>
<td>5' AGGGGAATGAGACGAGGTGACCG' 3'</td>
<td>11/17</td>
</tr>
<tr>
<td>A19.7F</td>
<td>5' GCCGCTAGCTTCTCCCTCGG' 3'</td>
<td>11/18</td>
</tr>
<tr>
<td>A20.3F</td>
<td>5' CACCCCAAGACCTACCTCCCG' 3'</td>
<td>11/20</td>
</tr>
<tr>
<td>ASPA20.6CF</td>
<td>5' CTGGGCGCTCTGGCCTCCGG' 3'</td>
<td>9/15</td>
</tr>
<tr>
<td>ASPA20.6GF</td>
<td>5' CAGCGGCAAAGACCTTCCCGG' 3'</td>
<td>13/15</td>
</tr>
<tr>
<td>A21.0R</td>
<td>5' TCAGACCTTGGTTCACCAGG' 3'</td>
<td>10/20</td>
</tr>
<tr>
<td>A21.84R</td>
<td>5' CTCTGCAATCAGGGGCTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A21.8R3</td>
<td>5' TGAGACGAGAGAGGACGAGGAC' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A21.8R2</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>13/19</td>
</tr>
<tr>
<td>A21.8R</td>
<td>5' ACAGCTGAGATGACGTGTGG' 3'</td>
<td>13/19</td>
</tr>
<tr>
<td>A22.0R</td>
<td>5' ACAGCTGAGATGACGTGTGG' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A22.3F</td>
<td>5' CTCTGCAATCAGGGGCTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.4F</td>
<td>5' CTCTGCAATCAGGGGCTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.4F2</td>
<td>5' TGAGACGAGAGAGGACGAGGAC' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.5F</td>
<td>5' GCCGCTAGCTTCTCCCTCGG' 3'</td>
<td>9/15</td>
</tr>
<tr>
<td>A22.5F2</td>
<td>5' GCCGCTAGCTTCTCCCTCGG' 3'</td>
<td>9/15</td>
</tr>
<tr>
<td>A22.5R</td>
<td>5' CACCCCAAGACCTTCCCGG' 3'</td>
<td>10/18</td>
</tr>
<tr>
<td>A22.5R2</td>
<td>5' AGTCTGAGATGACGTGTGG' 3'</td>
<td>11/18</td>
</tr>
<tr>
<td>A22.8R3</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>13/19</td>
</tr>
<tr>
<td>A22.8R2</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.8R</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.9R</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.9R2</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.9F</td>
<td>5' TGAGACGAGAGAGGACGAGGAC' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A22.9F2</td>
<td>5' TGAGACGAGAGAGGACGAGGAC' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A22.9R3</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>13/19</td>
</tr>
<tr>
<td>A22.9R2</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.9R</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.9F2</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.9R3</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>13/19</td>
</tr>
<tr>
<td>A22.9R2</td>
<td>5' CTGCTGAGATGACGTGTGG' 3'</td>
<td>12/19</td>
</tr>
<tr>
<td>A22.9F</td>
<td>5' TGAGACGAGAGAGGACGAGGAC' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A22.9F2</td>
<td>5' TGAGACGAGAGAGGACGAGGAC' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A22.9R</td>
<td>5' TGAGACGAGAGAGGACGAGGAC' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>A22.9F2</td>
<td>5' TGAGACGAGAGAGGACGAGGAC' 3'</td>
<td>12/20</td>
</tr>
<tr>
<td>Oligo name</td>
<td>Oligo sequence</td>
<td>No. of GC/Oligo length</td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>A13 9A</td>
<td>5' TTTGCGT TCTAGCTAAA 3'</td>
<td>6/18</td>
</tr>
<tr>
<td>A13 9C</td>
<td>5' TTTGCGT TCTAGCTAAA 3'</td>
<td>7/18</td>
</tr>
<tr>
<td>A14 0A</td>
<td>5' ACCAAA GAGACCGCTT 3'</td>
<td>9/18</td>
</tr>
<tr>
<td>A14 0G</td>
<td>5' ACCAAA GAGACCGCTT 3'</td>
<td>10/18</td>
</tr>
<tr>
<td>A14 2A</td>
<td>5' GAGGAGT CCTAGCTGCG 3'</td>
<td>11/18</td>
</tr>
<tr>
<td>A14 2G</td>
<td>5' GAGGAGT CCTAGCTGCG 3'</td>
<td>12/18</td>
</tr>
<tr>
<td>A14 4A</td>
<td>5' AAAAACAGCGTCCAACTC 3'</td>
<td>13/18</td>
</tr>
<tr>
<td>A14 4G</td>
<td>5' AAAAACAGCGTCCAACTC 3'</td>
<td>14/18</td>
</tr>
<tr>
<td>A14 6A</td>
<td>5' TCTGTGC TGTCAATAAA 3'</td>
<td>6/18</td>
</tr>
<tr>
<td>A14 6C</td>
<td>5' TCTGTGC TGTCAATAAA 3'</td>
<td>7/18</td>
</tr>
<tr>
<td>A14 2T</td>
<td>5' AGTGGCT AGRAGCTGGG 3'</td>
<td>8/18</td>
</tr>
<tr>
<td>A14 2G</td>
<td>5' AGTGGCT AGRAGCTGGG 3'</td>
<td>9/18</td>
</tr>
<tr>
<td>A15 3A</td>
<td>5' CTAMTTT TGTCTATTGTTT 3'</td>
<td>10/18</td>
</tr>
<tr>
<td>A15 3T</td>
<td>5' CTAMTTT TGTCTATTGTTT 3'</td>
<td>11/18</td>
</tr>
<tr>
<td>A15 7C</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>12/18</td>
</tr>
<tr>
<td>A15 7T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>13/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' AGTGGCT AGRAGCTGGG 3'</td>
<td>10-11/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>11-12/18</td>
</tr>
<tr>
<td>A15 9C</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>12-13/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>13-14/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' AGTGGCT AGRAGCTGGG 3'</td>
<td>14-15/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>15-16/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>16-17/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>17-18/18</td>
</tr>
<tr>
<td>A15 7C</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>18-19/18</td>
</tr>
<tr>
<td>A15 7T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>19-20/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>20-21/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>21-22/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>22-23/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>23-24/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>24-25/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>25-26/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>26-27/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>27-28/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>28-29/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>29-30/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>30-31/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>31-32/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>32-33/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>33-34/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>34-35/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>35-36/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>36-37/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>37-38/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>38-39/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>39-40/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>40-41/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>41-42/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>42-43/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>43-44/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>44-45/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>45-46/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>46-47/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>47-48/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>48-49/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>49-50/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>50-51/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>51-52/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>52-53/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>53-54/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>54-55/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>55-56/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>56-57/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>57-58/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>58-59/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>59-60/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>60-61/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>61-62/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>62-63/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>63-64/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>64-65/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>65-66/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>66-67/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>67-68/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>68-69/18</td>
</tr>
<tr>
<td>A15 9G</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>69-70/18</td>
</tr>
<tr>
<td>A15 9T</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>70-71/18</td>
</tr>
<tr>
<td>A15 9A</td>
<td>5' ATCTCT TCTCTATTGTTT 3'</td>
<td>71-72/18</td>
</tr>
</tbody>
</table>
Appendix II

DNA sequence of the α-globin gene cluster

(Key restriction sites, heterozygous SNPs, PCR primers and ASOs are marked, and with homology blocks underlined with different colours)

12890  AAACAGATGCCCTAAAGTCTCAGATGAAGGTGCCACGCCGCAACGCCGCTGGGTCATGCG

BamHI

12950  CAGAAACCTCTGGGATCGCGGCCTCCACAGAGCGGGGACTCCGCGCGTGGTTA
13010  CGTCTCGGCTGGGACGAGGTCGGTGCTCGGTCTCCGCGGCCTGGGATGATTACGG
13070  GGGGCGGCGGTGCCGAGCTCTCTAAGAGGCGCAGGGCAGGGGGGACCCCGCAGACCA
13130  GTCAAGCCCCGGCATGGCAAGGCAGGCGCCAAAATCGCAACTGCTGGGACCT
13190  GTATTGCGGCGCAAGGGCAGCATTGCCCGGGGAGACTGCTGGGATGAGACGCCG
13250  GCTCTCGGAGGTCAGGGGAGGATGAGGGGGGTTTTGGGCTGGGGCGGCAACACCGCA

BamHI

13310  TACGAATGGCTCTCTGGGTGCTGGGATCCTGCCGGTGCGCTAGTGAAGCC

RsaI

13370  CAGGCAGCGCCGCGCTCTCCGGTCATGCTCAGGGGCTTCTCAGGCTAGGTGACCT
13430  CCAGCCAAGTCTACTTCCCACGCCACCTCGACGGCCTGAGGCCAGCGCCGCAGGGTGTAGG

Sphi

13490  GCACACGGCGACGCGCTGCTGGGCTGGGACGTCAGGGGCTGGGATGAGACCCG
13550  GCCGCGCGCTGAGCCCGCGGGACCGACGTCAGGGGCTGGGATGAGACCCG
13610  ACTTCCCAGGGCCTGGGTCTCTGCCGGGCAATGTCAGGGCGCCGGGCTGGGATGAGG

----- A13.6F --->  ---- A13.6F2 --->

PvuII

13670  GCTCTGGGGGTCTCTGCTCCAGGGCGCGGACCCGGCTCTACGCGCAGTTGTGGATGAGACGCGCC
13730  AAATCTGAGGGTTCTCAGTGGCTGCGCTGGCTCCTCCACACTGAGGAAGTTCTCCAGGGCAAT

----- A13.7F --->  <----- A13.8R ---->

13790  GCACACGGCGGCTGGGAGAACTTCCCTGACTGTGTGTCGGCCTGGGACGCGCGACGACCTGCTCG

----- A13.9C/A --->

13850  CTTGCATCGCCTGTGCAGGGCCAGGCTCTGGTGCTGGGCTGGGATGAGACGCGCC
13910  TCTCGGCGCCCTGCGCTGCTGGGCTGGGATGAGACGCGCC

----- A14.0A/G --->

13970  GTGACAGCTCGAGGGCTCAGAAAGGACCGCTCTAGAACTCAATGGGGCGGCCACCCCAG
14030  GGTTGTCGCTCCTCGCGTCCGATGGCGGAGTGGTGCTGGGACGCGCC
14090  GTGTAGGGGATGTCGGGATGGCTCTGGACACTCAACCCTGGAAATCCGCGCGGCAGCACGAC
14150  AAACAGCTCGGCTGGACAGCTGATTAACCTCACTCTGCTGGGACTTCTCCGAGCTGGGACGCGCC
14210  GACAGCGGCAGCTCGAGGGGACGCGGAGTGGTGCTGGGATGAGACGCGCC
14270  GAGGGCCATGTCGACACCTAATTGTCCTGGCCAGAGCAGGGCCGGGCGCGCTGGGACGCGCGG
14330  TTCCCCCTTCCCCCTTCCCCACGAGCTCCCACTGACGCTGCAACGCTGGGACGCGCGG

----- A14.4T/A --->

14390  CCTCTCCGAGCTACGTCGGCGCTGGGATGGGATGAGGTGGGAGGAGGCTGGGACGCGCGCC
14450  TTGGCCAGAGCAGAGCTGATTTAAGTCTCTGCAGCTGCTGAGCGCACTCC

----- A14

14510  GGGCGTGCCGACAGCTCCGCGCTTACCGCTTACCTATCGAGAAGAAAGGACGCTCGAGCG

.52C/T --->  ---- A14.6T/A

14570  GAGAATTTTTGGAATTAAATTTTTTGGAAATCATAACAGAGGGTGGGCTGGGCAGGCTGCT

----- PvuII

14630  CAGCCCAACCTCCTCAGGGGAGGCCGGCCGCGCCTGCTGCTCCTAGCTGGGACGCGCGC
14690  AACGAGCTCCCTTCACAGCCTGCTTCTTCTCCTCCCTCCTCCTTCAGACTGGGAGGGAAATCTTA
14750  GGCGTCACCCCGAGCGCTTTTCTACGACAAAGAGTTAGGTCGAGAGAAGAGGTCGGACCCTTG

192
25670  GGTAGAAATATGTCAAATTGTAACGCAGAGGCTGTTGGAGTGCCACGGCTTTTTTACAAATT
<----- A25.7R ------>

25730  AATTTGATCAAGACCCGTATTAAATTATATCATTTCTCTCACTCTCTGCTGTCAGTTG

25790  ACTAAAGCTTCTAGAGAAAAATAATTTAATTATG6CGGCGCGGCGTGCTCCACCTGTGAATTTG

25850  CAGCCTCTTTCAGCCTGCTCAGGGCTTGAATCCCTCAACAGGCGCTTGGAGGACCCAGCT

25910  AGGACACCATAGTGAACACCCCTGTCCTCTACTAAAGACAAATTTGCACGTTGATGACGCT

25970  TCATGTCCTGTAACCTCGGCACTTTGCGGAACTGTTGTAGTGATGCTCAAGATCCGGCACA

26030  TGCAGCTTCCATGTTGGCCACAAAGAAGGCCGAAACTTCTGCTGTCATAAAAAATTTATCTCATAT

DraI  AluSp ->

26090  TTAATTTTTATTAAATTTATTACGCAGG

26150  GCACCTTTTGGGAGGCAAGTTGGGAGATCATGCACAAGGTACAGGCTGACCCAGCT

26210  AATTAGGGAAAACCCCTTCTCTCATATAATAATAAAAAATTACCGGGCTGTTGGGACGGCA

26270  CGCCTGTAATACCCCTAACCTGGGAATCTGCAACATTGAAATCTGAGAAGGACGGGCACA

26330  AGAGGTTGACGTACCCGAGATACCAAACTTCATGATGCTTGGCCTGGGACACAAGC

v.diverged LimBS ->

26390  CTCCCATCTCAAAAAAAAAAAAATTTAAAAATATACATCAGTACGATGATAGGGTAAGAAGATAT

26500  GTGGTAGCTTGGCTGCAAGAATGGGAAAGTCTTATATGGTGGGAGACCCAGCAAAAGGCCCAC

26510  GGAATGATAGCTGCCACCCTGCAAAATATTTGCCAAAGATGAAAAGAAAAGCGCCGACAAAGAGTACC

26570  AAACTCGCATGACCCCCATCATATAGTGAACATGCCCTGCTACAGCCCAAACTATGGAGACCAAAA

RsaI  AluSp ->

26630  GCGGATTAGGCTGTTCGGCCCGGGTTAGTACGTATGTTAATGAGAACGAGGCTGCTGCTGCT

26690  ATGATGAAAA AACCTCCCTGACCCCTAGATGTGCAGTGTGAGTCAGCTACAATAGATCTTAAA

26750  TATCAGCAAAATTTTATACCTGGAATAATGCTCAAAATGTGAAATATTGTCTATTTTATAC

RsaI

26810  CTTTTTTTAAAAACAAAAAGATATATAAGGGTACGACAGATTGAGTGTGCATATGACATTAT

RsaI

26870  CTATTATTCTTGGTGTTACATCCCTCAGCTACATCAAATAATTGATGAAAGAACAAGGAC

26930  CTGCTACGAGTCTGACCCTCACCCTACAGTGTATCTCCGTAAATCTCCCAAATCTGAGAAGAA

26990  ACTGCTACCCCTCATATTTTTATAATAGTAATAAGAATAAGACAGATCCGGCCCAAGTTCACAT

TCTACAAATTTTATACCTGGAATAATGCTCAAAATGTGAAATATTGTCTATTTTATAC

<----- A27.1R ------>

27050  AATCAGAGCCGAGATAAAAATACGACTCTCCCAAATGCCAAATCCATGCGGAAAATTCGCTATA

RsaI  AluSpX ->

27110  AATAATAGTAAATGGGGCGGGCGGGTGGTCACCGCCTOTAATCCCAAGACTTTTTGGGGAGCCT

27170  GAGGGCGGAGGAGGCATACCTGTAGGCGGAGTTTTGAGATAGCTACACTGCCAGAAAACTT

SpH1

27230  TCTCTACTAAAAATAACACAATGGGCATGTTGCGCATGACCCCTGATTCAGCCACTCCG

27290  AGGCTGAGGGCGAGGCTCTAGCTGATGACAGGCGGAGGCGGAGGGTGAGTGAACCGCGAGTG

<-- AluX

27350  CGCCACTGCTACTCCGACCCCGGGCAATTTTTTGTGTGTGTGTAGAGACTAAATACCAT

BamHI

27410  ATAGTGAAACACCTAAAGACGGGGGGCCCTTGACACAGGGCGATTACGAGGGCCCGTGCTGG

HBQ1 exon 1 ->

27470  AGCCTGTCGGAGATTGACGGCGCCGCGCTCCCCGGGATCTCCGACAGGGCCCCTTGAGACCC

27530  GCGCGCAGACGTCGGCGCCGCGCCCTGGGACACCGGGGGCCTGGGAGCCCAGAGGCGCCGCA

ini

27590  GCGCGCAGACGTCGGCGCCGCGCCCTGGGACACCGGGGGCCTGGGAGCCCAGAGGCGCCG

T G p=0.63

27650  GACCCGGGCGCTGGCGGCGCCCGCTGGGAGAAAGCTGCGGCGAAACTGCGGCGCAGCTACA

27710  ACAGAGGGCGCCGTAGACGCGTGGCGGGCGCCCCCCCAGGGGAGGCGCGGCGCCTCCCTC

G

200
HBQ1 exon 2 ->

27770  CCAAGCCCCGGACGCGCCTACACCAGGTGCCTCTTGGGCACTTTGCTGGCTTCTCCC
27830  GCCAGGAAGACTCTCTCTCCACCTGGACCTAGGCCGCCGGCCTCTCACAAAGTACAGGAC
27890  CACGGCAGAGCTGGCAGGCGGCTGCTGGGCGCGGCTGCTGGGCGCGGCTGCTGGGCGCGG
27950  CACGGCAGCTGCTGGGCGCGGCTGCTGGGCGCGGCTGCTGGGCGCGGCTGCTGGGCGCGG

---- A28.0F ---->

28010  TTCCAGGTAGCGCTGGCCGGCTGCTGGGCGCGGCTGCTGGGCGCGGCTGCTGGGCGCGG

HBQ1 exon 3 ->

28070  TGCCGGGGCTGGCCGGCCGGGGCTGCTGGGCGCGGCTGCTGGGCGCGGCTGCTGGGCGCGG

<----- A28.1R -----

28130  GGGGACATGCCTGCTGGTCTACCCCTCCCGCCGACCTACCCCGGAGACTTCAGCCCCGCGCT
28190  GCAGGCCTGCTGGGACAGGTCTCTCTGGACCTGCGGCGCTGGGTTTCCGAGTACCGG

ter  BamHI

28250  CTGAACTGTGGGTGGGTGGCCGCGGGATCCCCAGGCGACCTTCCCCGTGTTTGAGTAAAG
28310  CTTCTCCCAGGACGACGGCTTCTGGCCGTCCTCTCTGGACGGCGAGAGGAGGCG
28370  GGGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
28430  CGGGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
28490  GGGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
28550  GGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
28610  GGGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
28670  GGGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
28730  GGGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
28790  GGGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
28850  GGGCGGGGCTCCCCAGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG

<----- A28.9R -----

201
References


Arnheim N, Calabrese P, Nordborg M (2003) Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. Am J Hum Genet 73:5-16


Barnes WM (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc Natl Acad Sci U S A 91:2216-2220


Carpenter AT (1979) Synaptonemal complex and recombination nodules in wild-type Drosophila melanogaster females. Genetics 92:511-541


Ferguson DO, Holloman WK (1996) Recombinational repair of gaps in DNA is asymmetric in Ustilago maydis and can be explained by a migrating D-loop model. Proc Natl Acad Sci U S A 93:5419-5424


Fullerton SM, Bernardo Carvalho A, Clark AG (2001) Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol 18:1139-1142


Gonda DK, Radding CM (1983) By searching processively RecA protein pairs DNA molecules that share a limited stretch of homology. Cell 34:647-654


205


Keeney S, Kleckner N (1996) Communication between homologous chromosomes: genetic alterations at a nuclease-hypersensitive site can alter mitotic chromatin structure at that site both in cis and in trans. Genes Cells 1:475-489


Lindsay SJ, Khajavi M, Lupski JR, Hurles ME (2006) A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. Am J Hum Genet 79:890-902


Mieczkowski PA, Lemoine FJ, Petes TD (2006) Recombination between retrotransposons as a source of chromosome rearrangements in the yeast Saccharomyces cerevisiae. DNA Repair (Amst) 5:1010-1020


210


Sung P (1997b) Yeast Rad55 and Rad57 proteins form a heterodimer that functions with replication protein A to promote DNA strand exchange by Rad51 recombinase. Genes Dev 11:1111-1121


White MA, Dominiska M, Petes TD (1993) Transcription factors are required for the meiotic recombination hotspot at the HIS4 locus in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 90:6621-6625


