

**Analysis of DNA diversity and meiotic recombination in
the human MHC class II region**

Liisa Kauppi
Department of Genetics
University of Leicester

Thesis submitted for the degree of Doctor of Philosophy

November 2003



**University of
Leicester**

UMI Number: U203706

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U203706

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

CONTENTS

ABBREVIATIONS

ACKNOWLEDGEMENTS

ABSTRACT

1. INTRODUCTION	1
1.1. MEIOTIC DIVISIONS	1
1.1.1. The first meiotic division	2
1.2. PHYSICAL SIGNATURE MARKS OF RECOMBINATION	3
1.2.1. DNA features	3
1.2.2. The synaptonemal complex	4
1.2.3. Recombination nodules and chiasmata	4
1.2.3.1. Mlh1 foci	5
1.3. MEIOTIC RECOMBINATION IN YEAST	6
1.3.1. Tetrad analysis	6
1.3.2. Dealing with a double-strand break in meiosis	6
1.3.3. Yeast recombination hotspots	9
1.4. MEIOTIC RECOMBINATION IN MAIZE	12
1.5. MOUSE MEIOTIC RECOMBINATION HOTSPOTS	13
1.5.1. Mouse hotspots defined by sperm analysis	15
1.6. MEIOTIC RECOMBINATION IN HUMANS	15
1.6.1. Relating physical distance to genetic distance	16
1.6.2. Crossover interference	16
1.6.3. Linkage disequilibrium analysis	17
1.6.4. Recombination hotspots defined by LD and pedigree analysis	17
1.6.5. Detecting recombinants directly from human sperm	18
1.6.5.1. MS32 hotspot	19

1.6.5.2. <i>TAP2</i> hotspot	20
1.6.5.3. The <i>DOA/RING3</i> hotspots	21
1.6.5.3.1. <i>Crossover asymmetry and meiotic drive</i>	21
1.6.5.4. <i>SHOX</i> hotspot	22
1.6.5.5. The β globin hotspot	22
1.7. SEQUENCE MOTIFS IMPLICATED IN MEIOTIC RECOMBINATION	22
1.8. THE HUMAN MAJOR HISTOCOMPATIBILITY COMPLEX CLASS II REGION	23
1.8.1. “Polymorphic frozen blocks”	24
1.8.2. Clustering of familial crossovers	25
1.8.3. Hot recombination domains identified through single-sperm analysis	25
This work	26
2. MATERIALS AND METHODS	28
2.1. MATERIALS	28
2.1.1. Chemical reagents and equipment	28
2.1.2. Oligonucleotides	28
2.1.3. Enzymes	28
2.1.4. Genomic DNA	28
2.1.5. Standard solutions	29
2.2. METHODS	29
2.2.1. DNA extractions	29
2.2.1.1. DNA extraction from semen	29
2.2.1.2. DNA extraction from blood	29
2.2.2. DNA purification by electroelution	30
2.2.3. DNA amplification by PCR	30
2.2.3.1. PCR buffer	30
2.2.3.2. General PCR	31
2.2.4. SNP discovery	31
2.2.5. SNP genotyping	32

2.2.5.1. Dotblots	32
2.2.5.2. ASO hybridisations	32
2.2.6. LD analyses	33
2.2.7. Recombination assays	33
2.2.7.1. Allele-specific primers	33
2.2.7.2. Linkage phasing of suitable semen donors	33
2.2.7.3. Recovering crossovers from sperm	33
2.2.7.3.1. <i>Digestion and quantification of genomic DNA</i>	33
2.2.7.3.2. <i>Primary PCR and S1 digestion</i>	34
2.2.7.3.3. <i>Secondary PCR</i>	34
2.2.7.3.4. <i>Southern blot transfer and hybridisation to detect recombinants</i>	35
2.2.7.3.5. <i>Tertiary PCR</i>	35
2.2.7.4. Calculation of crossover frequencies	36
2.2.7.5. Crossover breakpoint mapping	37
2.2.7.6. Calculation of crossover activity in each inter-SNP interval	37
3. SNP DIVERSITY IN THE MHC CLASS II REGION	42
3.1. INTRODUCTION	42
3.1.1. Measures of nucleotide diversity	44
3.1.1.1. Nucleotide heterozygosity π	44
3.1.1.1.2. Watterson's theta, θ	44
This work	44
3.2. RESULTS	45
3.2.1. SNPs and their allele frequencies	45
3.2.2. Nucleotide diversity	53
3.2.2.1. SNP incidence	53
3.2.2.2. Nucleotide heterozygosity π in re-sequenced regions	54
3.2.2.3. Watterson's theta (non-recombining DNA segments only)	55
3.2.2.4. Two large polymorphic insertion/deletions downstream the <i>DPA1</i> gene	56
3.3. DISCUSSION	58

4. LINKAGE DISEQUILIBRIUM ANALYSES	59
4.1. INTRODUCTION	62
4.1.1. Measures of LD	63
4.1.1.1. Statistical analysis	64
4.1.1.2. Graphical output to illustrate the extent of LD	64
This work	65
4.2. RESULTS	66
4.2.1. LD in the <i>COL11A2</i>– <i>DOA</i> interval	66
4.2.1.1. LD breakdown in the <i>DPB1</i> gene region	67
4.2.1.2. Localised LD breakdown 3' of the in the <i>DPA1</i> gene region	69
4.2.2. LD in the <i>RING3</i> – <i>TAP2</i> interval	69
4.2.3. Decay of LD with distance within LD blocks	71
4.3. DISCUSSION	72
5. IDENTIFICATION AND CHARACTERISATION OF A SPERM CROSSOVER HOTSPOT NEAR THE <i>DPA1</i> GENE	75
5.1. INTRODUCTION	76
This work	77
5.2. RESULTS	78
5.2.1. Crossover rate	78
5.2.2. Distribution of crossover breakpoints	81
5.3. DISCUSSION	85
6. IDENTIFICATION AND CHARACTERISATION OF THE <i>DMB</i> SPERM CROSSOVER HOTSPOTS	89
6.1. INTRODUCTION	89
This work	89

6.2. RESULTS	89
6.2.1. Crossover rate	91
6.2.2. Crossover distribution	93
6.2.2.1. Two separate crossover hotspots in the <i>DMB</i> assay interval	94
6.2.2.2. Assigning crossovers to <i>DMB1</i> and <i>DMB2</i> hotspots in donors 63 and 87	97
6.2.2.3. Centrepoinets of the <i>DMB1</i> and <i>DMB2</i> hotspots	97
6.3. DISCUSSION	98
7. CROSSOVER ANALYSIS AT THE <i>DPB1</i> GENE	102
7.1. INTRODUCTION	102
This work	103
7.2. RESULTS	103
7.2.1. Recovery of putative crossover positive reactions	104
7.2.2. Rate and distribution of putative crossovers	106
7.2.2.1. PCR artefacts map to end intervals	106
7.2.2.2. Crossovers mapping to internal SNP intervals	107
7.2.2.3. Rate estimates of genuine crossovers	107
7.3. DISCUSSION	111
8. SUMMARY OF LD AND RECOMBINATION HOTSPOTS IN THE MHC CLASS II REGION	113
OVERVIEW	113
8.1. LD BLOCKS IN THE MHC CLASS II REGION	113
8.2. CROSSOVER HOTSPOT ANALYSES ACROSS REGIONS OF LD BREAKDOWN	116
8.2.1. LD breakdown but no crossover hotspot at the <i>DPB1</i> gene	116
8.2.2. Crossover hotspots discovered at regions of localised LD breakdown	116
CONCLUDING REMARKS	118

9. EFFECT OF RECOMBINATION HOTPOTS AND POPULATION HISTORY ON LINKAGE DISEQUILIBRIUM	120
9.1. INTRODUCTION	120
9.1.1. Target LD block between the <i>DNA3</i> and <i>DMB</i> hotspots	121
9.1.2. Populations studied - UK North Europeans, Saami and Zimbabweans	121
This work	122
9.2. RESULTS	123
9.2.1. Population allele frequencies differ dramatically within an LD block	123
9.2.2. The same LD block is observed in all populations	125
9.2.3. Cosmopolitan haplotypes within the extended LD block are uncommon	126
9.3. DISCUSSION	129
10. DISCUSSION	132
10.1. LD AND HAPLOTYPE BLOCKS IN THE HUMAN GENOME	132
10.1.1. The HapMap project	133
10.1.2. Relating LD to crossover hotspots	133
10.2. WHAT ABOUT GENE CONVERSION?	135
10.3. HOW DOES A CELL CHOOSE A RECOMBINATION HOTSPOT?	136
10.4. MOUSE CROSSOVER HOTSPOTS VS. HUMAN HOTSPOTS	138
10.5. CONSEQUENCES OF RECOMBINATION ACTIVITY FOR GENOMIC DIVERSITY	139
10.6. CONCLUSIONS AND FUTURE DIRECTIONS	140
Appendix 1, PCR primer sequences	141
Appendix 2, ASO sequences	145
Appendix 3, PCR primer and ASOs used in population study	154
References	158

ABBREVIATIONS

ASO	allele-specific oligonucleotide
CEPH	Centre d'Etude du Polymorphisme Humain
dNTP	deoxynucleotide triphosphate
DSB	double-strand break
HERV	human endogenous retrovirus
HLA	human leukocyte antigen
LD	linkage disequilibrium
LINE	long interspersed nuclear element
LTR	long terminal repeat
Mb, kb, bp	mega-, kilo- base pair(s)
MHC	major histocompatibility complex
MIR	mammalian-wide interspersed repeat element
PCR	polymerase chain reaction
PFB	polymorphic frozen block
RFLP	restriction fragment length polymorphism
SINE	short interspersed nuclear element
SC	synaptonemal complex
SVA	SINE variant Alu element
SNP	single nucleotide polymorphism
UTR	untranslated region

ACKNOWLEDGEMENTS

Firstly, I would like to thank Alec with whom it has been such a privilege to work with, and whose enthusiasm and deep understanding of biology I will always be in awe of.

I am also grateful to Nancy for helping me out with paperwork numerous times; to Celia and Rita who were always willing to help and who have given me invaluable advice throughout the whole of my PhD work; Yuri for his mad Russian parties and jokes, and advice on all matters population genetic; Carole, who I still miss, for being a great friend; Ruth (and Rich) for being so nice and helpful in and out the lab and for some great dinners; John Stead and Maria for helping me get started in the lab; Kimsel-B_nt and Tim for making the lab a fun place to work; Mark Hills and Richard for their help with anything computer-related; Foxie for all the snacks in the computer room; Ben and Eva for proof-reading some of my chapters, numerous dinners and just being good friends and housemates.

Kiitos isälle ja äidille, jotka aina olivat hengessä mukana ja joille olen kaiken velkaa, ja Lurkille, jonka kanssa puhutut pitkät puhelut ovat viihdyttäneet ja pitäneet minut maan tasalla.

Lastly and most importantly, to George (without who I most certainly would not have found my way to Leicester) who was always there for me, in spirit or otherwise, I am grateful for everything.

For my PhD funding I would like to thank: the Instrumentarium Science Foundation, who had enough faith in me to give me funding right from the start, Osk. Huttunen Foundation who funded a substantial fraction of my PhD work, and the Finnish Cultural Foundation.

Analysis of DNA diversity and meiotic recombination in the human MHC class II region

Liisa Kauppi

ABSTRACT

There is mounting evidence that recombination events are not randomly distributed in the human genome, but tend to cluster into distinct regions, so-called recombination hotspots. Work presented here was focused on developing an understanding of how meiotic recombination events are distributed in the human Major Histocompatibility (MHC) class II region. To this end, a large number of SNPs in this region was identified and genotyped, and high-resolution linkage disequilibrium (LD) patterns were examined for evidence of historical recombination. Three regions of LD breakdown, *i.e.* putative recombination hotspots were localised. For these regions, allele-specific PCR methods were used to selectively amplify recombinant molecules directly from sperm DNA.

This led to the identification of three novel crossover hotspots, within which I was able to demonstrate an extremely localised nature of crossover breakpoints. Furthermore, molecular characterisation of the three hotspots showed that crossover rates can vary dramatically from one hotspot to the next and that LD patterns cannot be used to readily predict these rates.

I also investigated the relative importance of known recombination hotspots *vs.* population history in shaping LD in three human populations (UK North Europeans, Saami and Zimbabweans). At least in this segment of the MHC, haplotype structures directly relate to fine-scale patterns of meiotic recombination, even though a distinct paucity of "universal" haplotypes (haplotypes shared by all three populations) was observed.

1. INTRODUCTION

While mutation creates new variation in the human genome, it is recombination that reshuffles this existing variation. Apart from its role in maintaining genetic diversity, recombination also ensures that homologous chromosomes (two copies of the same chromosome) are aligned properly before the first meiotic divisions, thus ensuring proper chromosome disjunction. On the somatic level, homologous recombination makes it possible for a diploid cell to repair DNA damage on one chromosome by copying the lost genetic information from the intact homolog or sister chromatid. Thus, recombination is a basic mechanism crucial for the viability and genetic stability of an organism. Conservation of the total amount of recombination (genetic map length) is remarkable across eukaryotic taxa as varied as fungi and vertebrates; however, recombination rates, *i.e.* amount of recombination per physical distance, is much higher in unicellular organisms (Awadalla, 2003).

Meiotic recombination creates a haploid product whose allelic combinations differ from either of the two parentals. The two types of interchromosomal recombination are crossover and gene conversion. Crossover involves reciprocal exchange of whole chromosome segments, whereas in conversion, a patch of DNA is copied from one chromosome to the other, with or without the exchange of flanking markers.

It has been established that meiotic recombination occurs more frequently in some regions of the eukaryotic genome than others (e.g. Lichten and Goldman, 1995). At the megabase level, DNA domains are observed where genetic distances are longer than expected from their physical size. When regions of active recombination are highly localised (e.g. less than 10 kb), they are referred to as recombination hotspots. Traditionally, recombination hotspots are defined as genomic regions where recombination frequency is greater than average. What makes certain genomic regions active or inactive in recombination has remained unclear. It has long been suggested that recombination may be more common at, if not exclusive to genes, because in a variety of organisms the overall amount of recombination correlates with gene content (Thuriaux, 1977).

Recombination is not always a perfect process. Failure to undergo crossover can lead to non-disjunction of homologous chromosomes, for example trisomy 21 in Down syndrome; unequal crossover can in rare cases result in deletion or amplification of DNA segments and cause genetic defects such as Charcot-Marie-Tooth disease type I (Lupski *et al.*, 1991) or alpha thalassaemia (Higgs *et al.*, 1989).

1.1. MEIOTIC DIVISIONS

In diploid sexually reproducing organisms, such as humans, meiosis produces haploid gametes, which then unite to form the diploid zygote. Meiosis consists of one round of DNA replication, followed by two successive cell divisions (meiosis I and II), so that the products are haploid germ cells (Figure 1.1). Of the two meiotic divisions, the first (the reductional

division) is long and complex and can take years to complete. The pairing of homologous chromosomes and meiotic recombination between non-sister chromatids occurs during prophase of this first meiotic division, prophase I.

1.1.1. The first meiotic division

During meiosis I, homologous chromosomes come together in a co-ordinated fashion and sister chromatids remain associated throughout. Prophase I consists of five stages: leptotene, zygotene, pachytene, diplotene (see Figure 1.2) and diakinesis. During leptotene, the chromosomes are unpaired and exist as fine threads. The bivalents (pairs of homologs) synapse during zygotene, and during pachytene the chromosomes thicken and crossovers occur. Chiasmata (physical connections at sites of crossover) are visible during diplotene, when homologs start to separate. During diakinesis the bivalents are more contracted. Metaphase I follows, during which the bivalents are aligned on the metaphase plate; they are then segregated to different daughter cells (or nuclei) in anaphase I. The second meiotic division, the equational division, is much simpler and is similar to a mitotic division - the sister chromatids are segregated to different daughter cells.

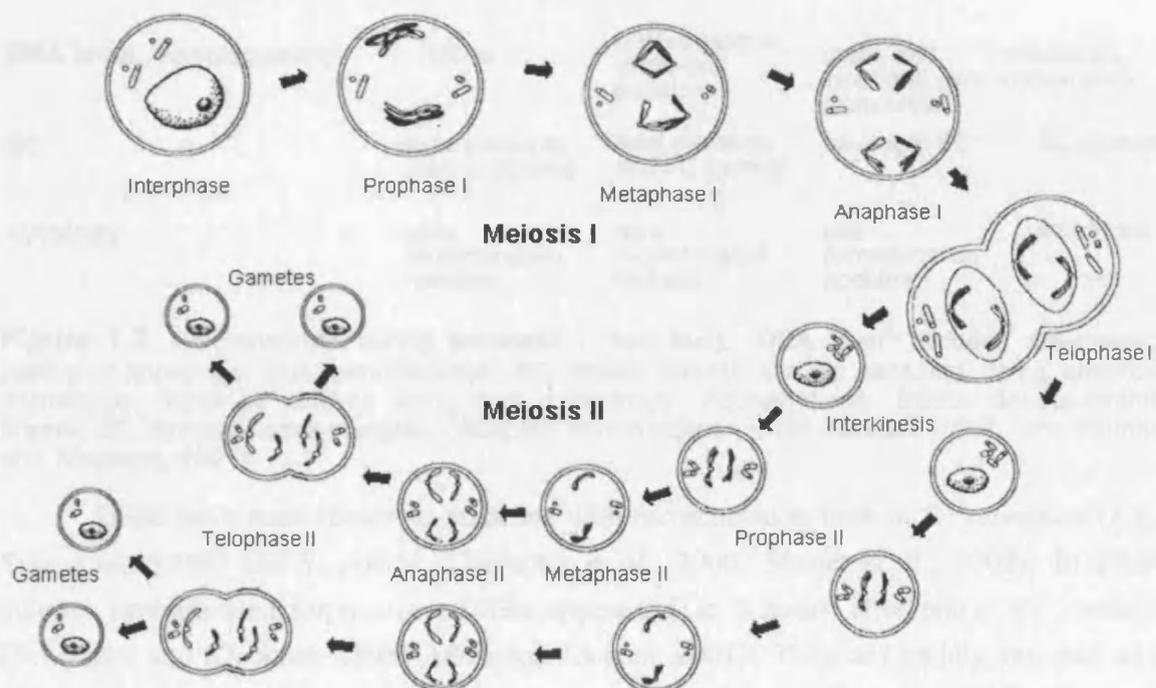


Figure 1.1. Schematic representation of meiotic divisions in a diploid, sexually reproducing organism.

1.2. PHYSICAL SIGNATURE MARKS OF RECOMBINATION

Meiotic recombination is a tightly regulated process, where each stage involves characteristic structures (see Figure 1.2). DNA features (meiotic double-strand breaks and their repair) have been mostly studied in the budding yeast *Saccharomyces cerevisiae*. The synaptonemal

complex, recombination nodules and chiasmata (described below) can be visualised in organisms with sufficiently large chromosomes, and have proven particularly useful in characterising distribution of crossovers in mice and humans.

1.2.1. DNA features

It is now generally accepted that double-strand breaks (DSBs) are the initiating lesions of meiotic recombination. In yeast, DSBs can be monitored on agarose gels after cleavage with restriction enzymes, or *via* pulsed field gel electrophoresis assays. Although DSBs cannot be directly assayed in other organisms, recently it has been shown that an antibody for the phosphorylated histone H2AX (γ -H2AX), which signals DSBs, can be used to investigate timing and distribution of meiotic DSBs in the mouse (Mahedevaiah *et al.*, 2001).

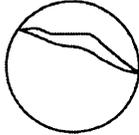
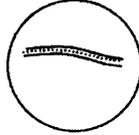
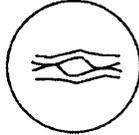
	(pre-leptotene)	leptotene	zygotene	pachytene	diplotene
					
DNA level	homolog pairing	DSBs	DSBs disappear, single-end invasions	single-end invasions, joint molecules	crossovers, conversions
SC		axial elements start to develop	axial elements and SC forming	full length SC	SC disassembled
cytology		early recombination nodules	early recombination nodules	late recombination nodules	chiasmata

Figure 1.2. Chromosomes during prophase I (see text). "DNA level" includes pre-meiotic pairing of homologs, plus recombination. SC related events can be detected using electron microscopy, "cytology" events using light microscopy. Abbreviations: DSBs, double-strand breaks, SC, synaptonemal complex. (Adapted from Kleckner, 1996, Roeder, 1997, and Hunter and Kleckner, 2001).

DSBs have been shown to associate with recombination both in *S. cerevisiae* (*e.g.* Sun *et al.*, 1989) and *S. pombe* (Cervantes *et al.*, 2000, Steiner *et al.*, 2002). In yeast cultures synchronised for meiosis, DSBs appear 2.5 to 3 hours after entry into meiosis (Schwacha and Kleckner, 1995, Allers and Lichten, 2001). They are rapidly resected and remain in the resected form for 15 to 30 minutes (Schwacha and Kleckner, 1995). The next detectable structures are single end-invasions (Hunter and Kleckner, 2001), which are followed by double Holliday junctions, also referred to as joint molecules (Schwacha and Kleckner, 1995). Last, recombinant products (crossovers and gene conversions) appear.

1.2.2. The synaptonemal complex

The meiosis-specific proteinaceous zipper-like structure that holds homologous chromosomes together along their lengths during leptot-, zygo- and pachynema, is called the synaptonemal complex (SC). It consists of axial elements (also called lateral elements in the mature SC) connected by a central region. The SC is fully developed at pachynema, and disassembles in diplotema, around the time when crossovers and conversions appear (Figure 1.2). DSBs precede SC formation in yeast (Roeder, 1995, Kleckner, 1996) and mice (Mahadevaiah *et al.*, 2001, reviewed in Hunter *et al.*, 2001).

Chromatin, which protrudes away from the SC as loops, is anchored to the lateral elements (reviewed by Roeder, 1997). Sister chromatids, in loop form, are located on the same side of the lateral elements. The size of the chromatin loops depends on chromosomal location (loops near telomeres are smaller than interstitial loops), and species (Heng *et al.*, 1996). The more DNA there is per SC length (the bigger the loops), the less crossovers are seen (reviewed in Roeder, 1995). DSBs occur preferentially in the middle of chromatin loops, *i.e.* further away from the axis association sites (Blat *et al.*, 2002).

1.2.3. Recombination nodules and chiasmata

Early recombination nodules (RNs) appear during leptotene and persist until pachytene. There is a much larger number of early RNs per cell than there are crossovers; they may mark the locations of all strand-exchange events. A selected subset of early RNs get converted into late RNs, which are round structures located on the central region of the SC, and can be seen under the electron microscope. There is good correspondence of late RNs to crossovers, and late RNs are generally thought to mark sites of crossover (reviewed in Roeder, 1995).

Chiasmata are physical connections between homologous chromosomes that can be seen under a light microscope during diplotene and diakinesis in organisms with sufficiently large chromosomes (*e.g.* humans, Laurie and Hulten, 1985a and 1985b, and mice, Lawrie *et al.*, 1995). Chiasmata correspond to locations of crossover. Their role in chromosome disjunction is to maintain connections between homologs after the SC has disassembled, until anaphase I (see Figure 1.1), by resisting the pulling forces from the spindle poles. Presumably to ensure correct disjunction, there is at least one chiasma (*i.e.*, crossover) per pair of homologous chromosomes. These are referred to as "obligate" chiasmata; then, depending on chromosome length there may be additional chiasmata. Chiasmata are more common in euchromatin and rare in heterochromatin, and sex-specific differences in chiasma distribution are apparent in humans (see below).

1.2.3.1. Mlh1 foci

Mlh1 foci are a cytogenetic tool for identifying sites of mammalian crossover. Mlh1 is a homolog of MutL, a bacterial mismatch repair protein. In yeast, Mlh1 is required for normal

levels of gene conversion and crossover (Hunter and Borts, 1997). Mutations in mammalian mismatch repair also disrupt meiotic processes: *Mlh1*^{-/-} male mice lack later stage spermatocytes and arrest at meiosis I, while *Mlh*^{-/-} females are infertile (Baker *et al.*, 1996). In spermatocytes of *Mlh1*^{-/-} mice, there is initially normal pairing but chromosomes become separated during diplotene. This suggests that *Mlh1* promotes chiasma formation or stabilisation.

Fluorescent *Mlh1* foci probably mark late recombination nodules, and hence sites of crossovers on mammalian SC spreads. In normal mouse spermatocytes, an average set of autosomal SCs has ~23 *Mlh1* foci (Anderson *et al.*, 1999), in good agreement with chiasma counts (~24 chiasmata, Lawrie *et al.*, 1995). Shorter SCs were found to have one focus, while longer SCs had more. Regardless of SC length, a large proportion of *Mlh1* foci were located near the telomeres, with the highest level of recombination in the subtelomeric interval. The distance between two *Mlh1* foci on the same SC was greater than expected if they were distributed randomly - this is consistent with crossover interference (see below). Most mouse SCs have one or two *Mlh1* foci, but no *Mlh1* foci in heterochromatin at or near the centromere. These findings were confirmed by Froenicke *et al.* (2002) who used *Mlh1* immunolocalisation in conjunction with multicolour fluorescent *in situ* hybridisation, which allowed them to identify each chromosome. They found excellent correspondence between the number of *Mlh1* foci and SC length (rather than DNA content in megabases or mitotic metaphase chromosome length). When numbers of *Mlh1* foci in males of four different mouse strains were assayed, ~15% variation was observed between strains with the lowest and highest number of *Mlh1* foci (Koehler *et al.*, 2002).

On human male autosomal chromosomes (n=44), on average 50.9 *Mlh1* foci were observed (Barlow and Hulten, 1998). Again, this is almost identical to chiasma counts of human male autosomes (50.3 chiasmata, Laurie and Hulten, 1985a). The number of *Mlh1* foci remains constant throughout pachytene, and *Mlh1* foci were preferentially found at subterminal positions, in agreement with the distribution of chiasmata. In human fetal oocytes, *Mlh1* foci first appeared at early zygotene and were present in nearly all of the cells by mid-to-late zygotene and pachytene, persisting until early diplotene, with a mean number of 70.3 foci per cell (Tease *et al.*, 2002). Thus, there are roughly 1.4 times more chiasmata in human females than males. To explain the sex-specific differences, it has been suggested that the strength of chiasma interference may be greater in spermatocytes than oocytes, or that chiasma numbers are directly related to the length of pachytene chromosomes, which in oocytes are roughly double the length of those in spermatocytes (Tease *et al.*, 2002).

Chromosomes 21, 18, 13 and X were examined in more detail in human oocytes; the number of foci increased with the size of the chromosome, chromosome 21 had a mean of 1.23 foci, whereas the X chromosome had a mean of 3.22 foci (Tease *et al.*, 2002). The foci were located interstitially and only rarely close to the centromere or telomere. In humans, as

in mice, it is the SC length, rather than the physical length of the chromosome, that governs the number of Mlh1 foci (Lynn *et al.*, 2002).

1.3. MEIOTIC RECOMBINATION IN YEAST

Eukaryotic recombination mechanisms are best characterised in the budding yeast *S. cerevisiae*. Knowledge of the mechanistics of meiotic recombination is almost exclusively based on yeast studies, though in more recent years, our understanding of recombination in higher eukaryotes has increased. In yeast the two types of meiotic recombination events, crossover and gene conversion, can be distinguished by tetrad analysis.

1.3.1. Tetrad analysis

By mating one haploid strain carrying linked wild-type alleles A and B with another haploid strain carrying mutant alleles a and b, a diploid is constructed. The progeny of this parental diploid that has undergone a crossover between the two loci will show allelic segregation 1AB: 1Ab: 1aB: 1ab, instead of 2AB: 2ab seen in the parental diploid (see Figure 1.3.). The crossover is reciprocal, because the tetrad that has the recombinant spore Ab also contains its reciprocal product aB. However, if the progeny has undergone gene conversion at the B locus, segregation at that locus will be 3B: 1b instead of 2B: 2b, although locus A will segregate normally. The converted tetrad will therefore contain one recombinant spore aB and three spores with parental marker combinations, *i.e.* this recombination event is non-reciprocal. The fact that the four products of meiosis are contained in the tetrad, which can be dissected and the products analysed separately, makes yeast an excellent model organism for studying meiotic recombination.

1.3.2. Dealing with a double-strand break in meiosis

Meiotic recombination is initiated by a DSB on one of the two homologous chromosomes (Sun *et al.*, 1989). DNA replication is required for DSB formation; delaying replication in a genomic region also delays DSB formation in that region (Borde *et al.*, 2000). During the first step of DSB repair, the DNA strands are resected 5' to 3' from the DSB, and single-stranded 3' overhangs are generated (Sun *et al.*, 1991). These single-stranded overhanging DNA tails, which vary in length and can be up to 800 bp long (Sun *et al.*, 1991), invade the (intact) homologous chromosome at the site of homology (Hunter and Kleckner, 2001). This results in the formation of heteroduplex DNA; if sequences on the two homologs differ in this region, a mismatch is generated. The invading strand then primes DNA synthesis.

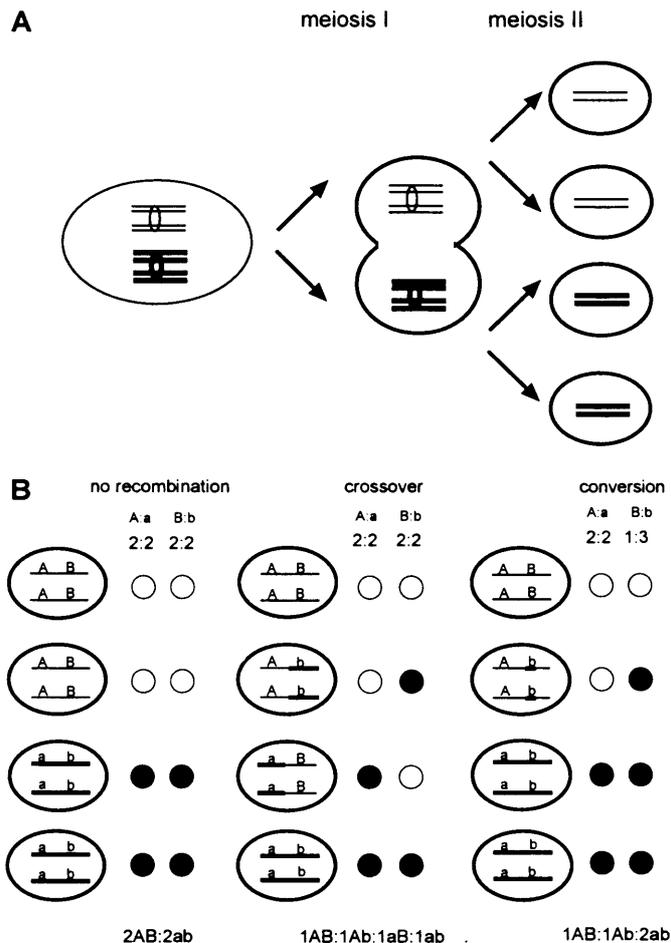


Figure 1.3. Recovery of all four products of meiosis allows tetrad analysis in *S. cerevisiae*. (A) Cell divisions leading to tetrad formation. During the first meiotic division, the two homologous chromosomes (thin and thick lines, drawn here without crossover) are separated. Sister chromatids remain together during the second meiotic division. All divisions take place inside the same physical structure (the ascus), which can then be dissected and the meiotic products analysed. (B) Two loci on the same chromosome, with alleles A and B on one chromosome, and a and b on the other, are followed through meiosis. Both loci carry selectable markers. Three possible meiotic outcomes are shown, no recombination (where parental allelic combinations are retained), crossover and conversion (see text).

It is thought that early on in this process the DSB commits to going down either the crossover or the non-crossover (conversion) route (Hunter and Kleckner, 2001, Allers and Lichten, 2001). For a DSB to be resolved as a crossover, there seems to be some homology requirement, as small sequence differences (9 insertion/deletions in the space of 9 kb) between homologs have been shown to reduce crossover frequency by half (Borts and Haber, 1987). DSB repair and synthesis-dependent strand-annealing (SDSA) seem to be two separate mechanisms that the cell uses to respond to meiotic DSBs (Figure 1.4). They show differential timing, with SDSA intermediates appearing earlier, and genetic separation; in yeast, the SDSA pathway is more common (Allers and Lichten, 2001).

The double-strand break (DSB) repair model (Szostak *et al.*, 1983) explains how crossovers and conversions can arise from the same initiating lesion. When the second 3' overhang is captured, joint molecules (double Holliday junctions) are formed. The plane in

which Holliday junctions are resolved determines whether the product is a non-crossover (conversion) or a crossover (see Figure 1.4, left).

Conversions can also arise *via* a different pathway, as shown by the SDSA model (reviewed in Paques and Haber, 1999). According to the SDSA model, conversions can be generated without capturing the second free end at the DSB. Instead, the invading and subsequently extended 3' end is ejected and re-annealed to its original chromatid. DNA synthesis and ligation followed by mismatch repair results in conversion (see Figure 1.4, right).

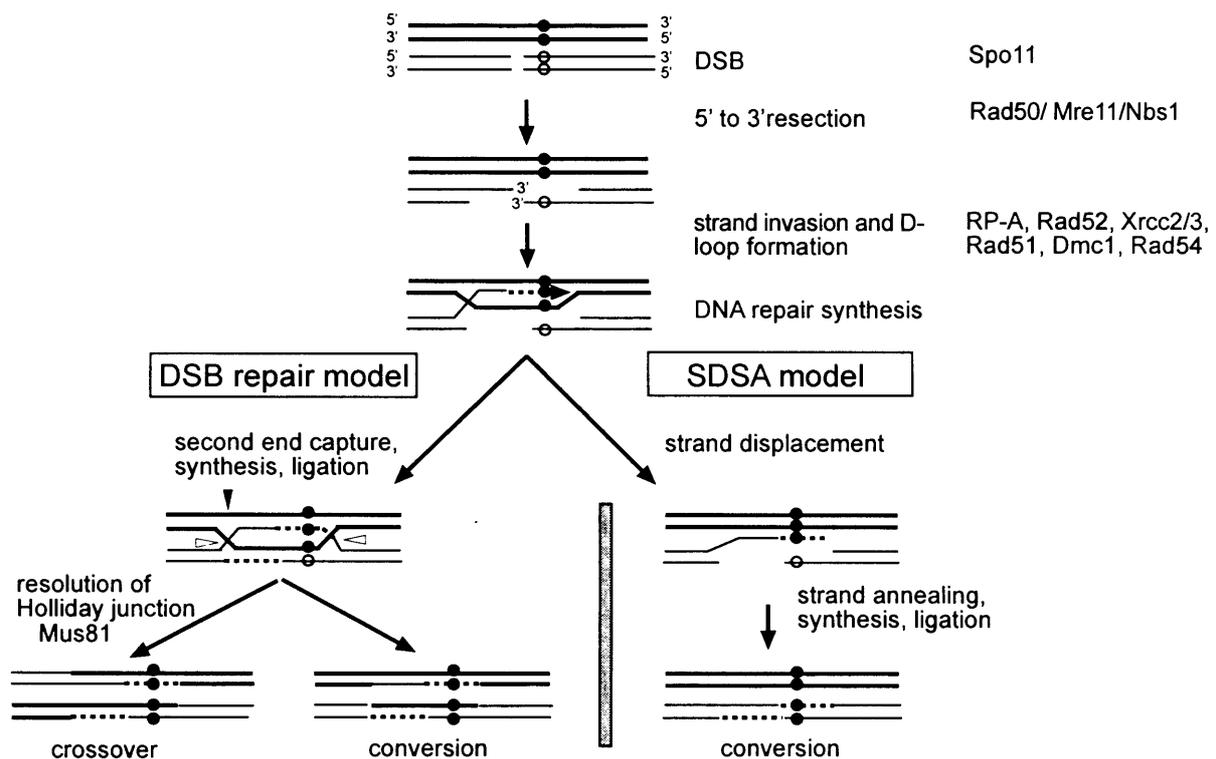


Figure 1.4. Repair of a DSB through recombination. A single-base difference between the two homologs (chromatids shown as thick and thin double lines) is indicated as filled and open circles. Following 5' to 3' resection from the DSB, a single-stranded DNA end invades the intact homolog. In the DSB repair model (left), the second DNA end is captured and double Holliday junctions are formed. Depending on the orientation in which the strands are cut, the Holliday junctions can be resolved as a crossover or non-crossover (shown here as a conversion). If strands are cut in the same plane (two white arrowheads), the result is a non-crossover; cutting in opposite planes (white and black arrowhead) yields a crossover. In the SDSA model, the invading and extended strand is ejected and anneals back with the other 3' overhanging tail. DNA synthesis and ligation follows. For simplicity, the single-base mismatch is shown as being repaired such that a conversion arises in all cases. Alternatively, the mismatch could be repaired using the "white" allele as template, in which case restoration would occur. If the mismatch were left unrepaired, the result would be post-meiotic segregation; such events can be scored as a sectorized colony in yeast tetrad analysis. (Adapted from Allers and Lichten, 2001; meiotic recombination proteins from Roeder, 1997, and Cromie *et al.*, 2001.)

Several recombination proteins have been identified in yeast mutant screens; a gene suspected to have a role in recombination is mutated and its effect on phenotype is

monitored. Some of the most important proteins involved in human meiotic recombination (as reviewed by Cromie *et al.*, 2001) are indicated in Figure 1.4. Briefly, Spo11 presumably catalyses DSB formation and from the break, a Mre11/Rad50/Nbs1 complex resects DNA 5' to 3', leaving a single-stranded 3' tail. The single-strand binding protein RP-A is loaded to these 3' overhangs. Rad52 then allows Rad51 to compete with RP-A for binding of the single-stranded DNA. The 3' overhangs invade the duplex DNA on the homologous chromosome aided by Rad51 and Dmc1. Xrcc2/3 and Rad51B/C/D probably interact with each other and promote recombination in a similar way as Rad51. Rad54 is required for D-loop and heteroduplex formation, and double Holliday junctions are resolved by Mus81 (reviewed in Cromie *et al.*, 2001).

1.3.3. Yeast recombination hotspots

DSBs in yeast are not randomly distributed, but instead tend to cluster in or near promoter regions (Wu and Lichten, 1994, Baudat and Nicolas, 1997, Gerton *et al.*, 2000). This would indicate that most, possibly all, yeast genes contain a recombination initiation site in upstream sequences (Wu and Lichten, 1994, Lichten and Goldman, 1995). DSBs usually occur in DNase I (Wu and Lichten, 1994) or micrococcal nuclease sensitive regions (Ohta *et al.*, 1994); therefore, DSBs appear to occur where nucleosomes are absent or disrupted and the underlying DNA is accessible to recombination machinery. Alternatively, it is possible that open chromatin structure is a consequence of the loading of recombination proteins onto the DSB. Ten or so meiotic recombination hotspots have been analysed in *S. cerevisiae*, of which two have been examined in great detail: the *ARG4* and *HIS4* hotspots.

The *ARG4* gene promoter region contains an initiation site for gene conversion, where DSBs are observed (Sun *et al.*, 1989). In wild-type cells, conversion events initiated at this site result in 7.4% conversion frequency at a restriction site ~450 bp downstream (Nicolas *et al.*, 1989). When several markers across the *ARG4* gene are monitored, a conversion gradient, also referred to as polarity, is seen. Conversion frequencies are highest at the 5' end of the *ARG4* gene (*i.e.* closest to the DSB region), and lowest at the 3' end of the gene (Nicolas *et al.*, 1989, Schultes and Szostak, 1990). The chromosome on which the DSB is initiated is the recipient of genetic information (Nicolas *et al.*, 1989), in accord with the DSB repair model. There is a conversion gradient on both sides of the DSB (Schultes and Szostak, 1990). In some mutants where the *ARG4* promoter and gene regions were inverted, conversion was almost abolished, showing that the DNA sequence itself is not sufficient for stimulating recombination (Rocco *et al.*, 1992). Instead, the authors were able to show that, when mutants had lost a transcription termination signal and consequently showed transcriptional activity across the *ARG4* gene, conversion was reduced. Accordingly, other inversion mutants where a transcription terminator was present showed higher than wild-type conversion frequencies; thus, transcription seems to inhibit recombination (Rocco *et al.*, 1992).

The region upstream of *HIS4* likewise contains a recombination initiation site, where DSBs are spread across a ~50 bp region (Xu and Petes, 1996). In the wild-type, this upstream sequence contains binding sites for four transcription factors (Bas1, Bas2, Rap1 and Gcn4), which are required for wild-type conversion activity (White *et al.*, 1993, Abdullah and Borts, 2001). A hotspot with higher than wild-type activity can be created by replacing the four wild-type binding sites with two Rap1 binding sites, as well as by replacing them with yeast telomeric repeats (White *et al.*, 1993). If inserted at the *ARG4* hotspot, the telomeric repeats also elevate conversion frequency above wild-type, showing the effect is not dependent on sequence context (White *et al.*, 1993). Another artificially constructed hotspot, the *HIS4-LEU2* hotspot, consists of 2.8 kb of DNA containing the *LEU2* region inserted just downstream the *HIS4* locus (Cao *et al.*, 1990). This construct creates two novel DSB sites and has a recombination activity ~6 times higher than the wild-type (Cao *et al.*, 1990). It should be noted that at the *HIS2* locus, there is a ~15% to 5% conversion gradient from 3' to 5' of the gene (Malone *et al.*, 1992). Therefore, not all *S. cerevisiae* hotspots show 5' to 3' gradients at genes.

The *ade6-M26* allele is a well-characterised recombination hotspot in the fission yeast *Schizosaccharomyces pombe*, and unlike the *ARG4* and *HIS4* hotspots, it is intragenic. M26 is a G → T point mutation in the *ade6* gene, and converts ~10 times more frequently than other *ade6* mutations. M26 creates a strong DSB site and a conversion gradient (Steiner *et al.*, 2002). M26 has the heptamer sequence 5'-ATGACGT-3' which is required for activity (reviewed in Davis and Smith, 2001). The *ade6-M26* hotspot is similar to the *HIS4* hotspot in its requirement for transcription factor binding - *ade6-M26* requires the Atf1/Pcr1 heterodimer (Davis and Smith, 2001). Moving the M26 mutation and its surrounding sequence to different genomic locations in most cases does not create a hotspot, demonstrating its dependence on genomic context.

In general, good correlation is found between the frequency of DSBs and the frequency of adjacent gene conversion (de Massy and Nicolas, 1993, Fan *et al.*, 1995, Bullard *et al.*, 1996, Steiner *et al.*, 2002). In all hotspots, DSBs occur in a 100-500 bp long region (Petes, 2001), and mean gene conversion tract lengths at most loci range from 800 to 2000 bp (de Massy, 2003).

More global approaches have generated DSB profiles along whole *S. cerevisiae* (Klein *et al.*, 1996a, Baudat and Nicolas, 1997) and *S. pombe* (Cervantes *et al.*, 2000, Young *et al.*, 2002) chromosomes, and a large number of *S. cerevisiae* chromosomal regions (Gerton *et al.*, 2000). On *S. cerevisiae* chromosomes I and IV, on average one DSB every 25 kb was detected: 10 DSBs on the 230 kb long chromosome I and 12 DSBs on the 275 kb long chromosome IV (Klein *et al.*, 1996a). The authors noted a range in the strength of DSBs on both chromosomes, where the weak DSB signal corresponded to 1% chromosome breakage (the lower detection limit of their assay) and the strong signal to 5 to

10% breakage. A similar DSB density and variation in DSB signal strength was found on yeast artificial chromosomes carrying human DNA (Klein *et al.*, 1996b).

To look at DSB distribution on yeast chromosome III, Baudat and Nicolas (1997) used a more sensitive assay involving a strain homozygous for the DSB-accumulating mutation *rad50S*, with the lower detection limit of 0.2% breakage. Again, there was large variation between the strength of DSB signals (corresponding to 0.2 to 8.8% breakage). DSBs were clustered into two large chromosomal domains, one on each chromosome arm, with a lack of DSBs around the centromere. Within the DSB cluster domains, the vast majority of DSBs was found in intergenic regions containing promoters, and the mean spacing was 2-3 kb (Baudat and Nicolas, 1997). There is a reasonably good correlation between the genetic map and the DSB map on chromosome III.

The first study of DSBs in *S.pombe* chromosomes concluded that DSBs in this organism have very wide spacing, with prominent break sites about 100 to 300 kb apart (Cervantes *et al.*, 2000). Subsequently, the use of a *rad50S* mutant allowed more accurate detection of DSBs, and DSB spacing between prominent break sites was found to be ~25 to 100 kb (Young *et al.*, 2002). Between the prominent sites, breakage occurs at a low (<1%) frequency (Young *et al.*, 2002). It has also been suggested that recombination events initiated at the prominent DSB sites can be resolved at regions remote from the DSB (Cervantes *et al.*, 2000, Young *et al.*, 2002).

Meiotic DSB formation requires Spo11, a topoisomerase II related protein that transiently binds to the 5' ends of the DNA fragments (Keeney *et al.*, 1997). Gerton *et al* (2000) used Spo11 to enrich for DNA that contains DSBs - and therefore recombination hotspots - and used it to probe microarrays containing all 6200 known yeast open reading frames. They found that whereas recombination coldspots associate non-randomly with centromeres and telomeres, hotspots associate non-randomly with regions of high G+C content. Also, a large proportion of hotspots are associated with open reading frames of genes with metabolic functions, particularly amino acid synthesis. The average spacing between hotspots was 54 kb.

Yeast recombination hotspots can be classified into three types: α -, β - and γ -hotspots, which are not mutually exclusive (Petes, 2001). α -hotspots require the binding of a transcription factor; examples are the *HIS4* and *ade6-M26* hotspots. β -hotspots contain nucleosome excluding sequences such as (CCGNN)₁₂, whereas γ -hotspots occur at sites of high GC content. It remains unclear whether this type of hotspot classification is valid in other organisms. No primary sequence determinants for DSB formation have been identified. An open chromatin structure is either required for, or a consequence of, hotspot activity. Open chromatin alone it is not sufficient - other factors can also have an effect, such as the general chromosome environment and the presence of other nearby recombination

hotspots in *cis* or *trans* (Wu and Lichten, 1995, Lichten and Goldman, 1995, Fan *et al.*, 1997). Histone modifications may play a role in activating hotspots (Petes, 2001).

1.4. MEIOTIC RECOMBINATION IN MAIZE

A wealth of information exists on meiotic recombination in maize *Zea mays*. In this organism, meiotic recombinants can, in suitable crosses, be scored as kernels that are phenotypically different from parental kernels. Maize genes are dispersed between large regions of retrotransposon sequences, where recombination rarely occurs; it has been proposed that every gene may contain a recombination hotspot.

The *bronze* (*bz1*) locus contains the most active maize recombination hotspot characterised, with recombination rate peaking at ~500x genome average (Dooner and Martinez-Ferez, 1997). Crossover breakpoints cluster in a <800 bp region, but within this region, are distributed more or less uniformly (Dooner and Martinez-Ferez, 1997). The authors noted, however, that crossovers were preferentially resolved within regions of longest sequence homology. Fu *et al* (2002) examined genetic *vs.* physical distances on either side of the *bz* locus. Proximally to *bz* lies a large (~90 kb long) retrotransposon block, whereas distally to *bz* there is a gene-rich region. The former was found to have a recombination rate close to the maize genome average, while for the latter, the rate was elevated 40 to 80-fold (Fu *et al.*, 2002). This supports the view that genic single-copy sequence is a preferred location for recombination in maize. Fu *et al*'s (2002) data also suggest a role for methylation in recombination - the recombinationally inert retrotransposon portion is heavily methylated, whereas the region distal to *bz* is not methylated.

Distribution of crossover breakpoints at the *waxy* (*wx*) locus resembles that of the *bz* locus. There is a ~2.6 kb long region within the gene across which uniform resolution of crossovers was observed (Okagaki and Weil, 1997) - the interval may in fact be narrower but could not be further refined due to lack of markers. Recombination rate at the *wx* locus is ~100x maize genome average, and the rate was not significantly altered by partial deletion of *wx* promoter sequence (Okagaki and Weil, 1997).

When the 140 kb interval between the *anthocyanin1* (*a1*) and *shrunken2* (*sh2*) genes was examined for physical *vs.* genetic distances, the *a1* gene region was found have a recombination rate roughly an order of magnitude higher than the intergenic region (Civardi *et al.*, 1994). This putative recombination hotspot was subsequently characterised at a higher resolution, and 14 out of 15 crossover breakpoints were found in a very short (377 bp) interval near the 5' end of the *a1* gene (Xu *et al.*, 1995). The 140 kb long *a1-sh2* interval (now known to contain at least four genes: *a1*, *yz1*, *x1* and *sh2*) has recently been found to contain, in addition to the *a1* hotspot, another two recombination hotspots (Yao *et al.*, 2002). One lies in nongenic single copy ("interloop region") sequence, and the other in the *yz1* gene. As the interloop region does not contain known transcribed sequences, and the *x1* gene region has a recombination rate close to the maize genome average, the authors

conclude that not every maize gene contains a recombination hotspot, and *vice versa*, not every hotspot lies in a gene. Depending on hotspot definition, however, the "every maize gene contains a hotspot" idea may still be valid, because *x1* has a recombination rate 30 times higher than its surrounding region, as Yao *et al* (2002) point out.

1.5. MOUSE MEIOTIC RECOMBINATION HOTSPOTS

Recombination activity in mammals is expressed in centiMorgans per megabase (cM/Mb); cM is defined as recombination frequency x 100. For example, one recombinant in 100 meioses observed across 1 Mb equals 1 cM/Mb. The average meiotic crossover activity in the house mouse (*Mus musculus*) genome is 0.5 cM/Mb (Shiroishi *et al.*, 1995). Evidence for non-random clustering of recombination events in the mouse comes from the major histocompatibility complex (MHC) region. The MHC is the best studied area of the mouse genome, and immunogeneticists have collected hundreds of recombinant mouse MHC haplotypes. At least seven recombinationally hot areas in a ~1.5 Mb interval within the mouse MHC locus have been described (see Table 1.1). These hotspot regions are located (centromeric to telomeric) near *Pb*, in the *Lmp2* region, within the *Eb* gene, within the *Ea* gene, in a region between *Ea* and *C4*, in a region between *Tnf* and *H2D* and near the *G7a* gene. While some of these regions - the *Ea-C4* and the *Tnf-H2D* 'hotspots' - could only be located to DNA segments tens or hundreds of kilobases in length, the remaining five (the *Pb*, *Lmp2*, the *Eb*, the *Ea* and the *G7c* hotspots) have been defined and characterised in greater detail.

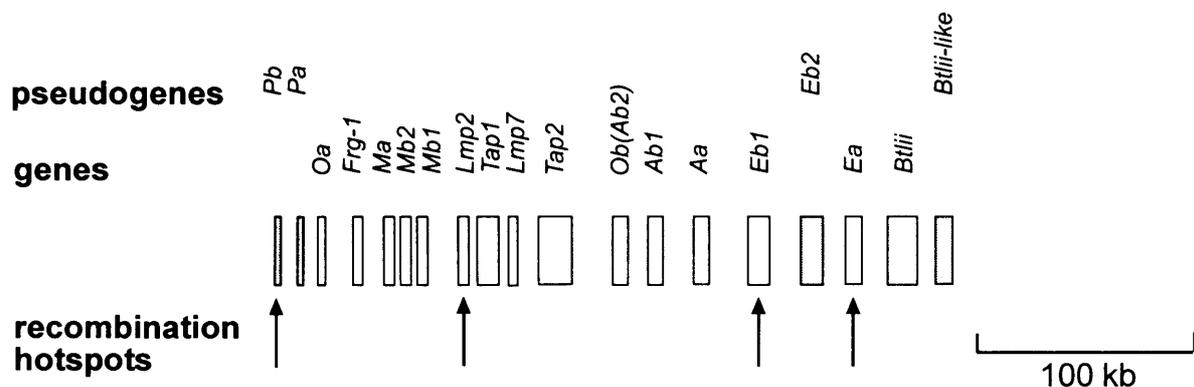


Figure 1.5. The classical mouse MHC class II (H2) region. Genes are shown as open boxes and pseudogenes as gray boxes. Recombination hotspots (indicated by arrows) were identified through pedigree analysis (see text and Table 1.1).

Two important observations have been made from the mouse MHC hotspots: recombination at highly localised hotspots can be enhanced both in a sex-specific and a strain-specific manner. The *Lmp2* hotspot is female-specific in the wild mouse derived *wm7* haplotype, with a recombination rate of 3.2% in *wm7* females and 0.2% in *wm7* males (Shiroishi *et al.*, 1990). In standard laboratory strains, no sex-specific differences in recombination rates were noted, and the average rate was 0.33%. When recombinant

animals were bred, those that had a laboratory strain haplotype upstream and the *wm7* haplotype downstream of the hotspot, showed low recombination frequency, like standard laboratory strains (Shiroishi *et al.*, 1991). Mice that had the *wm7* haplotype upstream of the hotspot, showed high recombination rate, and surprisingly, this effect was seen in both females and males (Shiroishi *et al.*, 1991). The authors concluded that for the *Lmp2* hotspot, the *wm7* haplotype contains (1) a "recombination instigator" in the upstream sequence, and (2) a male-specific recombination suppressor in the downstream sequence, highlighting the importance of *cis* acting factors in recombination. It was also noted that the *wm7* haplotype does not yield reciprocal recombinant offspring at similar frequency (Shiroishi *et al.*, 1990).

The *Eb* hotspot is active in most laboratory haplotypes, but not the *p* haplotype. Instead, the *p* haplotype preferentially recombines at the *Ea* hotspot, which is located ~40 kb telomeric of the *Eb* hotspot (see Figure 1.5. and Table 1.1). The *Lmp2* and *Pb* hotspots also show strain-specificity (Table 1.1). To test for chromatin accessibility at mouse hotspots, Mizuno *et al* (1996) investigated meiotic DNase I hypersensitivity the *Eb* and *Lmp2* hotspots, and found a hypersensitive site near *Eb* but not *Lmp2*; thus, there is no clear correlation.

All of the well-defined hotspots are found at genes (or pseudogenes, in the case of the *Pb* hotspot). As the mouse MHC class II is a gene-dense region, with a gene present roughly every 14 kb (Yuhki *et al.*, 2003), this may be coincidental.

Table 1.1. Mouse MHC recombination hotspots identified by pedigree analyses. (A) Hotspots where location and width (size) has been refined to 15 kb or less. "Activity" refers to how many times above genome average recombination rate is elevated. (B) Hot domains of recombination.

	Hotspot	size	strain/haplotype	activity	location	References
A	<i>Pb</i>	15 kb	cas4	188x	<i>Pb</i> intergenic	Yoshino <i>et al.</i> (1994) Isobe <i>et al.</i> (2002)
	<i>Lmp2</i> (=Psmb9)	2 kb	cas3, cas4, <i>wm7</i>	1880x	3' end of <i>Lmp2</i> gene	Uematsu <i>et al.</i> (1986) Steinmetz <i>et al.</i> (1986) Shiroishi <i>et al.</i> (1990) Mizuno <i>et al.</i> (1996)
	<i>Eb</i>	2.9 kb	most haplotypes, but not <i>p</i>	47x	second intron of <i>Eb</i> gene	Kobori <i>et al.</i> (1986) Lafuse and David (1986) Steinmetz <i>et al.</i> (1986) Zimmerer and Passmore (1991) Bryda <i>et al.</i> (1992)
	<i>Ea</i>	6 kb	<i>p</i> haplotype	not stated	<i>Ea</i> intergenic	Lafuse and David (1986) Lafuse <i>et al.</i> (1989) Khambata <i>et al.</i> (1996)
	<i>G7c</i>	5 kb	most haplotypes	not stated	4 kb telomeric of <i>G7a</i> gene	Snoek <i>et al.</i> (1998)
B	<i>Ea-C4</i>	300 kb	most haplotypes	not stated	not refined	Lafuse <i>et al.</i> (1989)
	<i>Tnf-H2D</i>	70 kb	only <i>p</i> haplotype crosses investigated	not stated	not refined	Heine <i>et al.</i> (1994)

1.5.1. Mouse hotspots defined by sperm analysis

Two of the hotspots previously characterised in pedigrees have been recently subjected to direct crossover analysis from sperm: the *Lmp2* hotspot (Guillon and de Massy, 2002) and the *Eb* hotspot (Yauk *et al.*, 2003). Guillon and de Massy (2002) used an allele-specific PCR assay to enrich for both crossovers and conversions from sperm of *wm7* hybrid mice in the recombination hotspot at the *Lmp2* gene (now re-named *Psemb9*). 69 crossovers and 16 conversions were recovered from 6000 sperm molecules; crossovers were simple and reciprocal, and the crossover frequency was highest in a 210 bp SNP interval. Conversions were assayed at the centre of the hotspot. Most (15/16) involved only a single marker, and were therefore less than 540 bp long. One gene conversion event included this marker plus one adjacent polymorphic site, giving a maximum conversion tract length of 749 bp. The authors also performed temporal analysis of recombinant molecules, and found that crossovers first appear at pachytene stage of meiotic prophase, which agrees with the occurrence of *Mlh1* foci (Baker *et al.*, 1996, Anderson *et al.*, 1999).

Yauk and colleagues (2003) analysed the influence of haplotype on male meiotic recombination at the *Eb* hotspot in great detail. Sperm crossovers were recovered from mice heterozygous for the *s* haplotype and either the *b*, *d*, *k* or *p* haplotype. Different crosses showed crossovers at the same hotspot but significant variation in recombination rates, with the *s* X *k* cross recombining almost 80 times more frequently than the *s* X *p* cross. Notably, the *p* haplotype, which had never been seen to recombine in pedigree studies, still recombined, albeit at a very low rate. Haplotypes showing little sequence divergence with *s* around the centre of the hotspot have higher recombination rates ($k > d > b \gg p$). Also, reciprocal crossover asymmetry (see section 1.7.6.3.1 for fuller description) was observed, consistent with the *s* haplotype being preferentially used for initiation of recombination.

Inter-locus gene conversion events between the MHC class II loci *Ab* and *Eb* have also been analysed in mouse sperm (Högstrand and Böhme, 1994). PCR was performed on sperm and liver DNA from mice heterozygous for the *k* and *d* haplotypes; nested forward primers were allele-specific for the *k* haplotype ("acceptor") at the *Ab* locus and the reverse primer was allele-specific for the *d* haplotype ("donor") at the *Eb* locus. Such inter-allelic conversions were meiosis-specific and rare, occurring at a rate of less than 2×10^{-8} in sperm.

1.6. MEIOTIC RECOMBINATION IN HUMANS

In human males, each cell that enters meiosis produces four sperm; this process takes 64 days to complete and is continuous from the onset of puberty. By contrast, in human females, oocytes undergo the initial stages of the first meiotic division, including synapsis and recombination, in early fetal life. This is followed by cell cycle arrest that lasts until puberty; then one egg per monthly cycle completes meiosis I and is released. Little is known about meiotic recombination in humans, and human gene conversion (apart from minisatellite associated conversion) is virtually unexplored territory.

1.6.1. Relating physical distance to genetic distance

On average, human DNA undergoes crossing-over at a rate of 1.1 cM/Mb (Kong *et al.*, 2002). There is, however, abundant evidence that crossovers are non-randomly distributed in the human genome. Comparisons of linkage maps with physical maps have revealed several Mb long domains of elevated ("hot") and suppressed ("cold") recombination activity (*e.g.* Hattori *et al.*, 2000, Yu *et al.*, 2001), though nothing is known about crossover distribution within the hot domains.

Another feature of human meiotic recombination, revealed by physical *vs.* genetic distance comparisons, is that, for unknown reasons, females and males exhibit differences in recombination rate and crossover distribution. Firstly, females have a higher overall recombination rate higher than males; the female to male genetic map length ratio is estimated to be roughly 1.6 to 1 (Broman *et al.*, 1998, Kong *et al.*, 2002). This is in reasonably good agreement with numbers of Mlh1 foci, discussed in section 1.2.3.2, where the female to male ratio is approximately 1.4 to 1 (Barlow and Hulten, 1998, and Tease *et al.*, 2002). Secondly, again in agreement with Mlh1 foci distribution, female-to-male ratios are highest around centromeres (where male recombination is extremely low) and lowest near telomeres (Mohrenweiser *et al.*, 1998, Broman *et al.*, 1998, Lynn *et al.*, 2000, Kong *et al.*, 2002). It has been suggested that these sex-specific differences result from differential gene expression, and thus differential chromatin structure and accessibility during gametogenesis (Wu and Lichten, 1994).

One peculiar feature of human female recombination is the rate variation between women (Broman *et al.*, 1998, Broman and Weber, 2000). The mean number of crossovers in women was 40, but the range in eight women was 33 - 47 crossovers (Broman *et al.*, 1998). Kong *et al.* (2002) found the same phenomenon, with a range of 3300 - 5700 cM in total map length, and termed it "mother effect"; no significant inter-individual variation was found in men. Female recombination rates may be regulated by a combination of genetic and environmental factors.

1.6.2. Crossover interference

As mentioned in section 1.2.3, chiasmata distribution is regulated such that each chromosome pair obtains at least one chiasma, *i.e.* one crossover. How does a cell decide where and how frequently to make a crossover? In yeast, a large number of DSBs are made, of which only a few mature into crossovers. The phenomenon of crossover interference, *i.e.* crossovers "repelling" each other, has been noted in a variety of organisms, including humans. There seems to be a minimum distance around a crossover within which a second crossover cannot occur (Collins *et al.*, 1996). Evidence for crossover interference comes from cytogenetic (see section 1.2.3, also Laurie and Hulten, 1985b) as well as pedigree studies (Lynn *et al.*, 2000, Broman and Weber, 2000). Observed crossover distributions were different than expected if there was no interference - there are fewer chromosomes with

very low or very high numbers of crossovers. For example, for chromosome 4, most maternal chromosomes had two crossovers, while most paternal chromosomes had one (Broman and Weber, 2000). No fourth chromosomes with more than five crossovers were seen in either sex. The centromere does not pose a barrier to interference (Broman and Weber, 2000).

The mechanism for crossover interference is unclear. It has been suggested that there may be local competition between recombination hotspots for components of the recombination machinery, *i.e.* there would be a limited amount of recombination proteins available at a given genomic location (Wu and Lichten, 1995, Fan *et al.*, 1997). Alternatively, there could be an inhibitory signal from the site of crossover to the surrounding genomic region. The synaptonemal complex is a candidate for transducing such a signal, the recipient could be a recombination enzyme complex sitting on a DSB waiting to be activated (Roeder, 1995 and 1997). Support for the latter is provided by the observation that *zip1* yeast mutants, which cannot form a SC, show random crossover distribution (Sym and Roeder, 1994). Furthermore, *S. pombe* does not have a classical SC and shows no interference (Fox and Smith, 1998).

1.6.3. Linkage disequilibrium analysis

Linkage disequilibrium (LD), *i.e.* the non-random association of alleles at separate loci, can be used to estimate the amount of historical recombination. In LD mapping (also called association mapping), marker associations are examined in unrelated individuals within a population. The logic behind this is that if strong LD is found between two loci, little if any historical recombination has occurred between them. Conversely, if there is little or no LD, this implies historical recombination activity. Thus, LD can be used to infer recombination hotspots (Hedrick, 1988). However, studying patterns of LD is hampered by the fact that LD is not a simple reflection of crossover frequencies along a given genomic region. There are numerous other factors influencing patterns of LD, such as gene conversion, selection of single or linked alleles and demographic history, *e.g.* founder effect and admixture (Laan and Pääbo, 1997, Zavattari *et al.*, 2000). LD is estimated from haplotype frequencies. A haplotype refers to alleles at different loci that are inherited as a "package" from one generation to the next. Haplotypes are usually inferred from diploid genotype data, or can be established directly from pedigree data.

1.6.4. Recombination hotspots defined by LD and pedigree analysis

A classical tool in gene mapping is to look at parent-to-offspring transmission of alleles at several loci. For recombination analysis, marker segregation in families allows direct detection of meiotic recombinants, and is often combined with analyses of LD patterns. Using these approaches, a number of putative crossover hotspots in the human genome have been identified. For example, Smith and co-workers (1998) found three familial crossovers

that fell into a 2 kb region in the replication initiation region of the β -globin gene cluster. In the phosphoglucomutase gene *PGM1* both segregation analysis in CEPH families and LD analysis were used to investigate meiotic recombination (Yip *et al.*, 1999). Two regions of crossover clustering were identified: one between exons 1A and 4 (~40 kb long) and the other in exon 7 (~1 kb long). Cruciani *et al.* (2003) found two maternal crossovers within a ~12 kb region in the adenosine deaminase gene (*ADA*), and subsequent LD analysis showed LD breakdown in this region in all five populations analysed. In the *LRP5* gene region, three LD blocks (blocks 1, 2 and 3 with lengths of 37, 18 and 110 kb, respectively) were identified (Twells *et al.*, 2003). Familial crossovers provided good evidence for the presence of a highly localised hotspot between blocks 1 and 2, in the first intron of the *LRP5* gene (Twells *et al.*, 2003). In the human MHC class II region, three regions show clustering of familial crossovers (Cullen *et al.*, 1995 and 1997); they will be discussed in more detail below, in section 1.8.2.

A well-defined hotspot for unequal crossover causes two autosomal dominant disorders of the nervous system: a 1.5 Mb region, containing the *peripheral myelin protein 22* gene, when duplicated, causes Charcot-Marie-Tooth (CMT) disease type 1A, and when deleted, hereditary neuropathy with liability to pressure palsies (HNPP). The 1.5 Mb region is flanked by two 24 kb sequences, referred to as CMT1A-REPs, which are 98.7% homologous (Lopes *et al.*, 1999). There are sex-specific differences in rearrangements (Lopes *et al.*, 1998): unequal crossovers are mostly of paternal origin, whereas intrachromosomal rearrangements are preferentially of maternal origin. 75% of all unequal crossovers occur in a 1.7 kb segment of the CMT1A-REPs (Lopes *et al.*, 1998). There is also evidence for crossover-associated conversions (Reiter *et al.*, 1998, Lopes *et al.*, 1999).

Neurofibromatosis type 1 (NF1) microdeletions seem to be caused by an identical mechanism to CMT/HNPP. The *NF* gene is flanked by long direct repeats, which are 85 kb long and 1.5 Mb apart. Unequal crossover between these repeats occurs in at highly localised hotspot (2 kb wide), and primarily in maternal meiosis (Lopez-Correa *et al.*, 2001).

1.6.5. Detecting recombinants directly from human sperm

Low marker density and, more importantly, the low frequency at which crossovers occur in human DNA are the main limitations of family studies. The 1.1 cM/Mb average recombination rate means that within a chosen 1 kb interval, a crossover will occur on average in only one gamete in ~90 000; therefore, crossover analysis at this level of resolution is impractical by pedigree analysis. To overcome these problems, single molecule methods have been developed that enable recovery of recombinant DNA molecules directly from sperm DNA. There are two variations on this theme – single sperm analysis, where PCR is performed on haploid genomes of flow-sorted sperm (Hubert *et al.*, 1994), and our method where batches of sperm are screened for recombinant molecules by PCR (henceforth

referred to as batch PCR, Figure 1.6). These PCR methods were initially developed for minisatellites to explore the relationship between tandem repeat DNA instability in the germline and meiotic recombination (Jeffreys *et al.*, 1994, 1998a and 1998b).

As a prelude to the batch PCR approach, SNPs are used as tools to carry out LD analysis (Figure 1.6). LD patterns are used as a guide to point to any putative hotspot region (*i.e.* any localised region of LD collapse). This allows the design of a crossover assay with allele-specific primers targeted to SNP sites flanking the putative hotspot. Using two rounds of allele-specific PCR, it is possible to amplify individual recombinant DNA molecules directly from sperm DNA (Figure 1.6.). The amplified recombinant molecules are dotted onto nylon membranes, and probed with allele-specific oligos (ASOs) targeted towards internal SNP sites. The crossover breakpoint can then be mapped to an interval between two SNP sites.

The obvious drawback of sperm PCR techniques is that they can only be used to analyse meiotic crossovers in males, as eggs cannot be collected like sperm cells. When estimating female recombination rates, we still have to rely on family data.

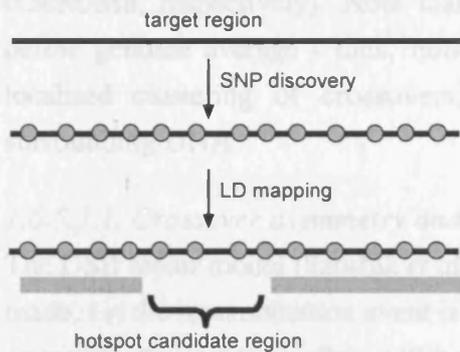
1.6.5.1. MS32 hotspot

Sperm DNA analysis revealed the first ever well-defined crossover hotspot, located just 5' to the minisatellite MS32 on chromosome 1 (Jeffreys *et al.*, 1998a). The overall meiotic crossover rate in and near MS32 is on average 4×10^{-4} per sperm, a rate about 1/20th that of the much more frequent inter-allelic conversion events that occur within the repeat array and lead to minisatellite instability. The MS32 hotspot is highly localised (~1.5 kb wide), and centered 200 bp 5' of the minisatellite array, with crossovers extending into the array. Peak crossover rate is 50 times the male genome average. Array instability and crossover are both suppressed when the C allele is present at a G/C polymorphic site 48 bp upstream of the allele (Monckton *et al.*, 1994, Jeffreys *et al.*, 1998a, Jeffreys *et al.*, 1998b). Recent LD mapping around the MS32 hotspot has shown that there is little evidence for LD breakdown across this hotspot (A. J. Jeffreys, unpublished), highlighting the complexity of LD patterns in the human genome.

1.6.5.2. TAP2 hotspot

It has also been shown that this approach can be extended to other putative recombination hotspots in the human genome, in particular to that in the *TAP2* gene within the class II region of the MHC (Jeffreys *et al.*, 2000). Here, sperm DNA typing has again revealed a highly localised crossover hotspot within the second intron of the *TAP2* gene, about 1.2 kb wide and flanked by DNA that is much less active in recombination. Crossover rate at this hotspot is elevated approximately 8-fold above genome average.

A. LD analysis



B. Crossover analysis

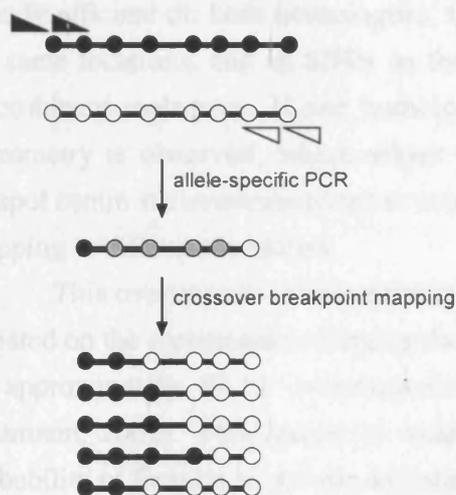


Figure 1.6. Batch PCR approach. (A) The first step is SNP discovery (SNPs shown as grey circles) and LD mapping, allowing the identification of LD blocks (grey boxes) and regions where LD breaks down (gap between grey boxes). Such regions of LD breakdown are putative recombination hotspots. (B) For putative recombination hotspot regions, two rounds of allele-specific PCR are performed on batches of sperm from a man with multiple SNP heterozygosities (black and white circles). Allele-specific primers are shown as black and white arrowheads. This way recombinant sperm DNA molecules are amplified; the location of the crossover breakpoint is mapped to an internal SNP interval (grey circles) by allele-specific oligo hybridisation.

1.6.5.3. The *DOA/RING3* hotspots

Larger scale LD mapping of the MHC class II region revealed a localised region of LD breakdown around the *DOA* (previously called *DNA*) gene region. This region was subjected to crossover analyses, and was found to contain a cluster of three hotspots, which were named the *DNA1*, *DNA2* and *DNA3* hotspots (Jeffreys *et al.*, 2001). These three hotspots are found within the space of less than 15 kb: *DNA2* is located 4 kb downstream of *DNA1*, and *DNA3* is located still another 8 kb downstream. All three hotspots show similar clustering of crossovers within a narrow (1 to 2 kb) region, and are embedded in recombinationally virtually inert DNA. Their recombination activity, however, varies greatly

(*DNA3* >> *DNA2* > *DNA1*, mean recombination rates roughly 110cM/Mb, 3cM/Mb and 0.3cM/Mb, respectively). Note that recombination rate at the *DNA1* hotspot is actually *below* genome average - thus, hotspots should perhaps be defined as regions that show localised clustering of crossovers, and have a higher recombination rate than their surrounding DNA.

1.6.5.3.1. Crossover asymmetry and meiotic drive

The DSB repair model (Szostak *et al.*, 1983) predicts that the homolog on which the DSB is made, *i.e.* the recombination event is initiated, is the recipient of information. It follows that over time, the initiating allele will be lost from the population, leading to the "death" of the hotspot (the "hotspot conversion paradox", Boulton *et al.*, 1996). If DSB initiation is equally efficient on both homologues, then reciprocal crossover breakpoints should map to the same locations, and all SNPs in the hotspot should be transmitted in a 50:50 ratio to recombinant molecules. If one homologue is preferentially used for initiation, crossover asymmetry is observed, where alleles carried on the non-initiating chromosome near the hotspot centre are overtransmitted to crossover molecules, resulting in reciprocal crossovers mapping to different locations.

This overtransmission is evident at the *DNA2* hotspot, where DSBs are preferentially initiated on the chromosome carrying the A allele of a G/A SNP near the hotspot, leading to an approximately 87:13 overtransmission of the non-initiating G allele (Jeffreys and Neumann, 2002). This results in weak but significant meiotic drive, which increases the probability of fixation of the non-initiating G allele.

There is now mounting evidence of crossover asymmetry, or disparity in various organisms. The phenomenon has been noted in yeast (*e.g.* Nicolas *et al.*, 1989) and at a mouse MHC hotspot (Yauk *et al.*, 2003). In humans, at least the *DNA2* and MS32 hotspots show crossover asymmetry (Jeffreys and Neumann, 2002, Jeffreys *et al.*, 1998a). It is possible that this is a feature which is shared by most hotspots, but will go undetected due to lack of markers near the centre of the hotspot, low numbers of reciprocal crossovers recovered, or both.

1.6.5.4. *SHOX* hotspot

The 2.6 Mb pseudoautosomal pairing region (*PAR1*) at the subtelomeric region of Xp/Yp is a site of obligatory crossover in male meiosis. The overall recombination rate across this region is elevated roughly 20-fold compared to the genome average (Lien *et al.*, 2000). High-resolution LD mapping around the *SHOX* gene in the *PAR1* showed no evidence of clear LD block structure (May *et al.*, 2002), consistent with a high uniform rate of crossover. However, a sperm crossover assay that was designed across one of the regions of free association revealed a highly localised hotspot. Recombination activity (140 to 370 cM/Mb in three donors) was higher than the "hottest" well-characterised autosomal hotspots,

DNA3 or *MS32*. The width (1.9 to 2.5 kb), however, is similar to all previously characterised hotspots, as is the contrast of recombination activity between hotspot and flanking DNA (roughly a 10-fold difference in crossover rate, C.A. May, personal communication).

1.6.5.5. The β globin hotspot

From LD and family studies it has long been known that the β globin gene probably contains a recombination hotspot (Chakravarti *et al.*, 1984, Smith *et al.*, 1998). Using the single sperm technique (see section 1.7.6., Hubert *et al.*, 1994), Schneider *et al.* (2002) estimated that the recombination rate in the 11 kb region is elevated roughly 80 times above genome average, but were not able to refine the location further. DNaseI and S1 nuclease hypersensitive sites in yeast artificial chromosomes containing the human β globin gene region (Svetlova *et al.*, 1998) do not show any obvious correlation with the hotspot.

1.7. SEQUENCE MOTIFS IMPLICATED IN MEIOTIC RECOMBINATION

As stated above, open chromatin structure often correlates with DSB sites. But what additional factors may be required to create a recombination hotspot? Numerous attempts have been made to identify specific sequences that act as a recombination signal. In higher eukaryotes, such conclusions are mostly based on “guilt by association”, where a sequence motif found near a recombination breakpoint was postulated to be causative. De Massy (2003) also points out that because the correlation between male and female recombination rates is only $R=0.57$, factors other than primary sequence must play an important role in governing recombination. Summarised below are some of the various sequence features or motifs that have been suggested to promote recombination.

Table 1.2. Sequence features implicated in recombination.

sequence	evidence	References, e.g.
GC rich	Good correlation with DSBs on yeast chromosome III Good correlation with DSBs chromosome-wide in yeast Correlation with recombination rates in humans	Blat <i>et al.</i> (2002) Gerton <i>et al.</i> (2000) Fullerton <i>et al.</i> (2001)
Chi sequence (<i>E. coli</i>)	Stimulates recombination in its vicinity	Reviewed by Eggleston and West (1997)
Translin binding motif	associated with breakpoint junctions in chromosomal rearrangements	Aoki <i>et al.</i> (1995)
GT repeats	Can adopt Z-DNA conformation Long GT repeats associated with high recombination on Mb scale in human males	Wahls <i>et al.</i> (1990) Majewski and Ott (2000), Tapper <i>et al.</i> (2002)
Alu repeats	Can cause unequal crossover, involved e.g. in alpha thalassaemia deletions	Harteveld <i>et al.</i> (1997)
Minisatellite sequences	Found near recombination hotspots, six copies of core sequence stimulates recombination <i>in vitro</i>	Reviewed by Wahls (1998)
O motif GTTTGCAT B motif GGGACTCTCC R motif (CCAG) ₂	a combination of one B motif, plus at least one O and R motif is found at Lmp2, Eb, Ea, Pb and G7c hotspots, and human TAP2 hotspot within the space of <1.2 kb	Isobe <i>et al.</i> (2002)
MT repeat	found at mouse Eb and Lmp2 hotspots	Shiroishi <i>et al.</i> (1990), Bryda <i>et al.</i> (1992)
CoHR (common homology region)	Good correlation with DSB sites in yeast chromosomes I, III and VI	Blumenthal-Perry <i>et al.</i> (2000)
Nucleosome-excluding sequences	(CCGNN) ₁₂ stimulates recombination in yeast	Kirkpatrick <i>et al.</i> (1999)
M26 heptamer (<i>S. pombe</i>)	Creates <i>ade6</i> -M26 recombination hotspot	Reviewed in Davis and Smith (2001)
Palindromic sequences/ hairpins	140 bp palindrome generates DSB site in yeast Inverted repeats stimulate intrachromosomal recombination <i>in vitro</i>	Nag and Kurst (1997) Waldman <i>et al.</i> (1999)

1.8. THE HUMAN MAJOR HISTOCOMPATIBILITY COMPLEX CLASS II REGION

The human major histocompatibility complex (MHC, also called human leukocyte antigen, HLA) class II region, the target region for LD and hotspot analyses in this thesis, was initially chosen for two reasons. Firstly, family studies had indicated that recombination events in this region are to some extent distributed non-randomly (Cullen *et al.*, 1995 and Cullen *et al.*, 1997), and secondly, complete sequence data was available (MHC Sequencing Consortium, 1999).

The MHC region is located on the short arm of chromosome 6 (6p21.31), spans roughly 3.6 Mb and contains more than 200 genes (MHC Sequencing Consortium, 1999). The MHC is a large multigene family whose products control the immune system's capacity to recognise foreign proteins. Not surprisingly, genes in the MHC are implicated in numerous autoimmune disorders, such as psoriasis, inflammatory bowel disease, multiple sclerosis and rheumatoid arthritis. The MHC genes are divided into three classes, centromere to telomere: class II, class III and class I.

The MHC region is the most polymorphic human locus. Overdominant selection (heterozygote advantage) is thought to operate on the MHC, maintaining the level of polymorphism. MHC class II molecules present antigens to T helper cells, which in turn activate B cells. The region is gene-rich, with one gene every 18.1 kb (Yuhki *et al.*, 2003), but coding sequences only account for ~7.5% of total sequence (Beck *et al.*, 1996). Repeat sequences make up 23.5% of total sequence (Beck *et al.*, 1996).

The classical class II region is roughly 0.7 Mb long and contains 38 genes, including pseudogenes. There are three classical class II loci (*HLA-DR*, *HLA-DQ* and *HLA-DP*) and

several class II pseudogenes, such as the *DPA2* and *DPB2* genes (see Figure 1.7). Class II genes have arisen by duplications, and class II molecules contain an alpha and a beta chain, which are encoded by separate genes (*e.g.* *DPA* and *DPB*). Except for *DO*, these genes are located adjacent to each other (Ting and Trowsdale, 2002). The *TAP1* and *TAP2* genes (transporter associated with antigen processing) encode proteins which transport peptides from the cytoplasm to the endoplasmic reticulum for association with class I molecules. Two additional genes, *LMP2* and *LMP7*, encode components of the proteasome, which is involved in protein degradation in the cytoplasm, and yields peptides for transport by the TAP proteins. *RING3* is of unknown, non-immune related function (Beck *et al.*, 1992). MHC class II gene order is remarkably conserved between mice and humans (see Figure 1.7), even though the mouse MHC class II is more compact.

The classical class II region is contained within an L2 isochore, with an average GC content of 40.2% (Stephens *et al.*, 1999). The isochore boundary is located ~220 kb telomeric of *DRA*, and MHC class I and class III regions have higher GC content than class II. The location of the isochore boundary corresponds to a switch in replication timing: class II is a late-replicating, and class I an early-replicating region (Tenzen *et al.*, 1997).

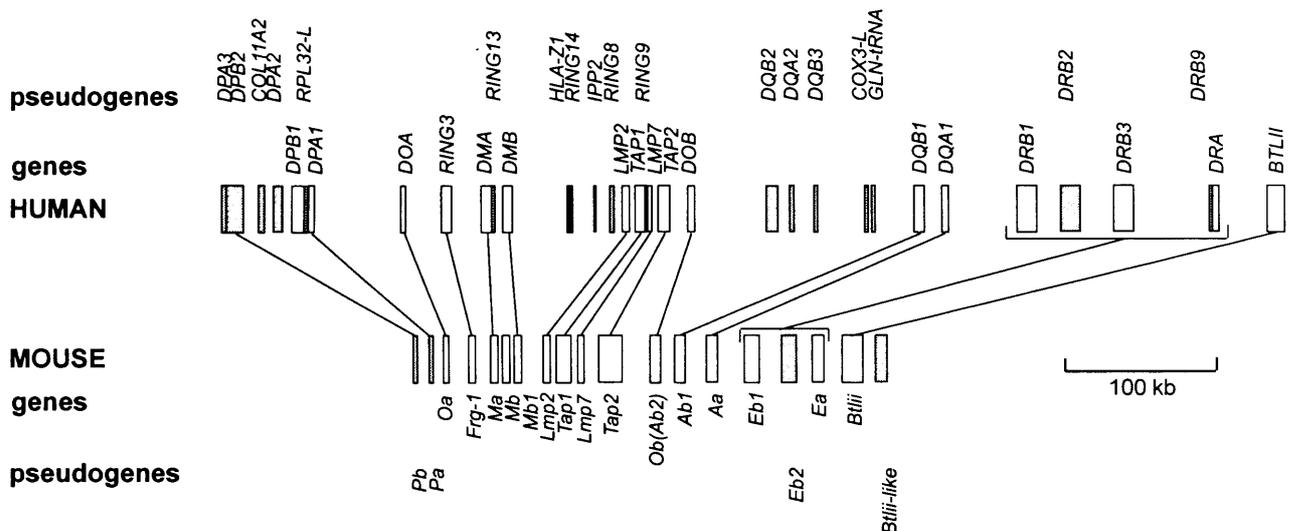


Figure 1.7. The classical human MHC class II region *versus* the classical mouse MHC class II region. Genes are shown as open boxes and pseudogenes as grey boxes; homologous genes are connected with lines. Adapted from Yuhki *et al.* (2003).

1.8.1. “Polymorphic frozen blocks”

It has been proposed that the human MHC consists of so-called polymorphic frozen blocks (PFBs), each 200-300 kb long, where within each block, there is no recombination and alleles thus remain linked (Gaudieri *et al.*, 1997). For example, within the class II region, the *HLA-DR* and *HLA-DQ* are thought to be contained within a PFB, the delta block, sized 60-300kb (Marshall *et al.*, 1993). In families, recombination events have never been

observed in the *DRA-DQA* interval (Trowsdale, 1995). Some PFBs show extreme sequence divergence between individuals.

1.8.2. Clustering of familial crossovers

In the human MHC class II region, familial crossovers tend to cluster into three regions (Cullen *et al.*, 1995 and 1997): a 45kb region between *HLA-DOA* and *RING3*, an 8.8kb region within the *TAP2* gene and a 50kb region between *HLA-DQB3* and *DQB1* (see Figure 1.8A). While there is clear evidence for non-random distribution of recombination events, statistical support for the significance of crossover clustering is poor (with the exception of the *HLA-DOA/RING3* region). Also, the data give no information on intensity, peak activity or a more specific location of the putative hotspots. In a survey of recombination in the human MHC in Sardinian families (Zavattari *et al.*, 2000), a total of 23 recombinant chromosomes were found over the *DPB1-DQB1* interval. Thirteen chromosome breakpoints could fall into the *DOA-RING3* interval (two unequivocally), and ten into the *DQB3-DQB1* interval (three unequivocally). Zavattari *et al.*'s (2000) marker density was lower, and although it supports Cullen *et al.*'s (1995 and 1997) results, it does not rule out the possibility that additional hotspots are located within the MHC class II region.

1.8.3. Hot recombination domains identified through single-sperm analysis

A 3.3 Mb interval, from the start of the classical MHC class II region to the *MOG-CA* marker telomeric of *HLA-F* in the MHC class I region, was subjected to single-sperm recombination analysis (Cullen *et al.*, 2002). 31 microsatellite markers were used, dividing the region into 30 DNA domains for which recombination activity was assayed in 12 sperm donors. In a total of 20,000 sperm, 325 recombinant molecules were identified, giving an average recombination rate of 0.49 cM/Mb across this region. Recombination events were non-randomly distributed - six "hot" domains (ranging from 1.7 to 5.2 times higher than region average rate) and 12 "cold" domains (ranging from 2.3 to 3.2 times lower than region average rate) were identified. 38% of all recombination events occurred within the "hot" domains. Hot and cold domains in the classical MHC class II region are shown in Figure 1.8B. Among the 12 donors, there was one set of HLA identical siblings and two sets of monozygotic twins. In all these pairs, total recombination rate was very similar, whereas for some comparisons between donors, significant (up to 6-fold) rate variation was observed. As family recombination data exists on the MHC class II region (Cullen *et al.*, 1997), sperm data could be compared with it. They were found to be in reasonably good agreement, suggesting these large-scale recombination patterns are similar in both sexes.

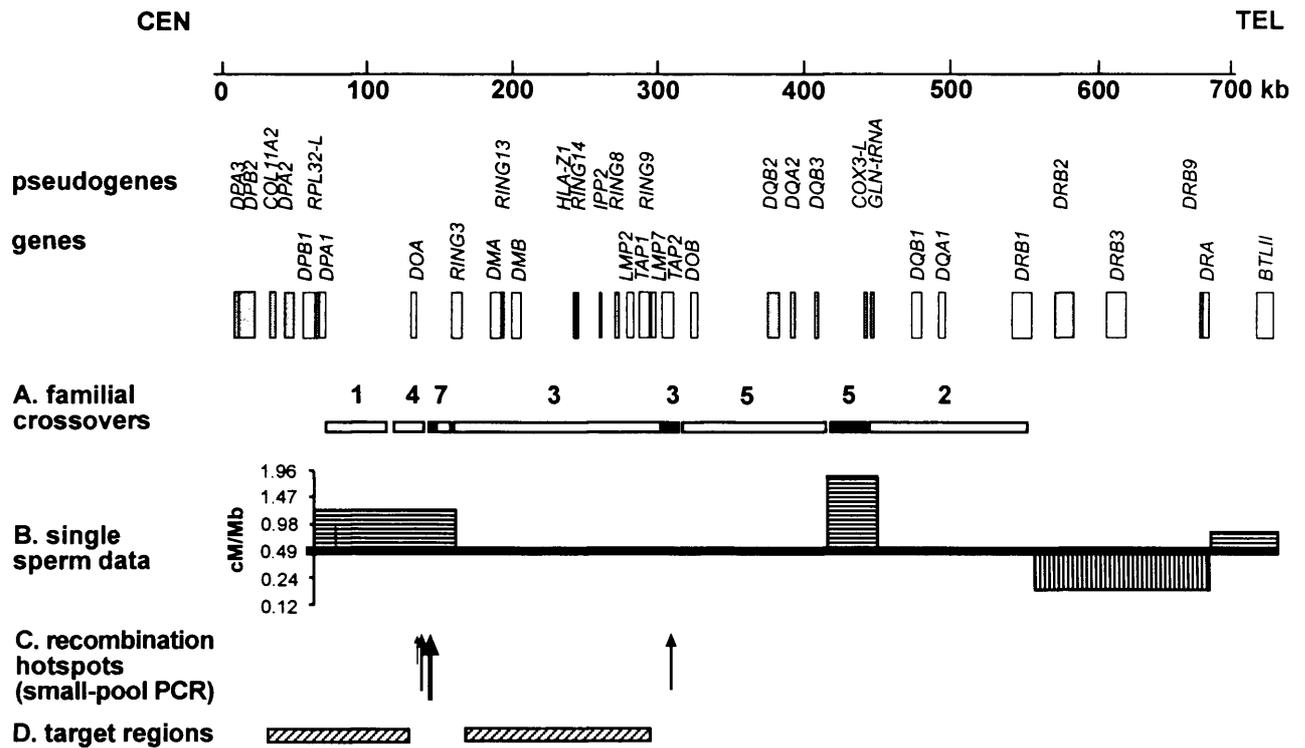


Figure 1.8. Recombination studies in the classical human MHC class II region. (A) Crossovers identified in pedigrees (Cullen *et al.*, 1997). Numbers of crossovers are shown above each domain. Domains in black show good evidence for highly localised recombination activity. (B) Recombinational properties of large DNA segments, characterised in sperm DNA (Cullen *et al.*, 2002). Average male recombination rate, 0.49 cM/Mb, is shown as thick line. The classical human MHC class II region contains three "hot" domains (horizontally hatched boxes) and one "cold" domain (vertically hatched box). The rest of the region had a recombination rate close to average (C) Location of recombination hotspots identified through batch PCR (see section 1.7.6.3). (D) Target regions for LD and crossover mapping, presented in this thesis.

This work

Studying human meiotic recombination at very high resolution will help to define the stability of the human genome and the role of recombination in generating haplotype diversity within functional genes and non-coding sequence. In addition, efforts to map complex common diseases would greatly benefit from a more detailed understanding of meiotic recombination (in particular the prevalence and nature of recombination hotspots), as the number of markers required for a whole-genome scan is dictated by the abundance, distribution and intensity of recombination hotspots (Gray *et al.*, 2000). When the work described in this thesis was started, little was known on how SNPs in the human genome are arranged into haplotypes, although a few examples of regions of high LD had been reported. The analysis of human crossover hotspots was in its infancy, and two localised hotspots had been analysed at high resolution (*MS32* and *TAP2*, see above). The relationship between LD and localised crossover hotspots had not been explored, apart from local LD surveys in the *TAP2* and *DOA* gene regions.

My project extends the LD and crossover analyses in the MHC class II region to cover altogether approximately 200 kb, upstream of the *DOA/RING3* hotspots and between

the *DOA/RING3* and *TAP2* hotspots (Figure 1.8D). My aim was, by using LD patterns as a guide to pinpoint regions active in recombination, to determine to what extent crossovers within these intervals cluster into hotspots, and to characterise any hotspots so discovered. New examples of hotspots would help to define rules that govern hotspot activity. Specific interesting questions to be answered include how LD patterns are influenced by hotspots, how variable hotspot width and recombination activity is, and how the location of human MHC hotspots relates to those in the mouse MHC.

Chapter 3 describes the results of SNP discovery and genotyping as a prelude to LD and crossover analyses. Great variability in both SNP density and diversity within this region is demonstrated.

Chapter 4 summarises my LD analyses. The block-like structure of LD allowed me to identify three regions of LD breakdown, with various degrees of marker associations across them. They are located in 5' to 3' order, in the *DPB1* gene, 3' of the *DPA1* gene and at the *DMB* gene.

Chapters 5 to 7 show the results of crossover assays, designed to span the regions of clear LD breakdown downstream the *DPA1* gene and near the *DMB* gene, and more equivocal LD breakdown at the *DPB1* gene. The *DPA1* and *DMB* regions were found to contain highly localised sperm crossover hotspots (in the case of *DMB*, two adjacent hotspots).

Chapter 8 summarises all data accumulated to date on LD patterns and crossover hotspots in the MHC class II region, contributed not only by the work described in this thesis, but also by A.J. Jeffreys and R. Neumann.

Chapter 9 describes the survey of LD structure and haplotype diversity across a 80 kb region, which consists of an LD block flanked by the *DNA* and *DMB* recombination hotspot clusters, in three populations with very different demographic histories (Saami, Zimbabweans and UK North Europeans), testing the relative importance of known recombination hotspots vs. population history on LD block structure.

Work described in chapters 6 and 9 has been published (Jeffreys *et al.*, 2001, Kauppi *et al.*, 2003).

2. MATERIALS AND METHODS

The methods used during the course of this work followed almost exclusively standard molecular biological procedures adequately described elsewhere. Thus, only brief descriptions (with appropriate references) are given of the general methods used; materials and methods outlined below do not contain an exhaustive list of exact protocols. Each results chapter (chapters 3-7 and 9) includes short descriptions of methods relevant to that chapter.

2.1. MATERIALS

2.1.1. Chemical reagents and equipment

All chemical and molecular biology reagents and equipment were supplied by established suppliers (ABgene, Amersham, Advanced Biotechnologies, Bio-Rad, Boehringer Mannheim, Cecil Instruments, Clare Chemical Research, Clontech, Eppendorf Scientific, Fisher Scientific, Fisons, Flowgen, FMC Bioproducts, Gibco-BRL, Hybaid, Invitrogen, MJ Research, New England Biolabs, Nalge Nunc International, New Brunswick Scientific Co., Perkin-Elmer/Applied Biosystems, Pharmacia, Qiagen Ltd, Serva, Shandon Southern, Sigma and UVP Life Sciences).

2.1.2. Oligonucleotides

Oligonucleotides for PCR amplification and ASO hybridisation were supplied by the Protein and Nucleic Acid Chemistry Laboratory, University of Leicester.

2.1.3. Enzymes

Restriction enzymes were supplied by New England Biolabs. T4 polynucleotide kinase, S1 nuclease and *Pfu* polymerase were obtained from Gibco-BRL, and *Taq* polymerase was supplied by ABgene. The Klenow fragment of DNA polymerase I of *E.coli* was obtained from Pharmacia. Proteinase K was supplied by Sigma.

2.1.4. Genomic DNA

Semen samples from anonymous donors of UK North European origin were provided by Jane Blower (Leicester Royal Infirmary). Additional semen samples, also of UK North European origin, were provided by members of the Department of Genetics, University of Leicester. The collection and use of these human samples was approved by the Leicestershire Health Authority Research Ethics Committee. Zimbabwean DNA samples were donated by A.D. Nakomo and S.B. Kanoyangwa (Forensic Science Laboratory, Causeway, Zimbabwe). Saami blood samples were provided by Antti Sajantila at the Department of Forensic Medicine, University of Helsinki, Finland.

2.1.5. Standard solutions

Southern blot solutions (depurinating solution, denaturing solution and neutralising solution), 20x sodium chloride-sodium citrate (SSC) buffer and 10xTris-borate/EDTA (TBE) electrophoresis buffer, were as described by Sambrook *et al.* (1989), and were supplied by the media kitchen, Department of Genetics, University of Leicester.

2.2. METHODS

General methods for ethanol precipitation, handling DNA, gel electrophoresis *etc.* were performed as adequately described by Sambrook *et al.* (1989). DNA restriction enzymes and other DNA modifying enzymes were used according to the manufacturers' instructions with the supplied buffer systems.

2.2.1. DNA extractions

All DNA extractions were carried out in a category II laminar flow hood under conditions that minimise risk of contamination (Jeffreys *et al.*, 1994).

2.2.1.1. DNA extraction from semen

1ml of 1xSSC was added to each 200 μ l aliquot of semen, and cells were pelleted by centrifuging at 13000 rpm in a microcentrifuge for 1 minute. The supernatant was removed and cells were washed again with 1ml 1xSSC. The supernatant was discarded, and 900 μ l 1xSSC/0.2%SDS were added to lyse any non-sperm cells, contents were mixed and centrifuged at 13000 rpm for 2 minutes. The supernatant was discarded, the sample was re-suspended in 1ml 1xSSC and centrifuged as before. The supernatant was discarded and the sample was re-suspended in 450 μ l 0.2xSSC. Sperm heads were lysed by the addition of SDS (1% final concentration) and 2-mercaptoethanol (1M final concentration), and incubated at room temperature for 5 minutes. Freshly prepared proteinase K was added to a final concentration of 200 μ g/ml and the sample was incubated for 1 hour at 37°C with occasional mixing. Proteins were removed by addition of 300 μ l phenol/chloroform with gentle mixing to allow emulsification, and centrifuged for 2 minutes. The organic layer was re-extracted with 1xSSC/1%SDS, and all of the sample collected from the organic layer was subjected to a second round of phenol/chloroform extraction. DNA was ethanol precipitated, rinsed with 80% ethanol, dissolved in ddH₂O and re-precipitated. The pellet was vacuum-dried and dissolved in 50 μ l 5mM Tris/HCl pH 7.5.

2.2.1.2. DNA extraction from blood

Equal volumes of 1xSSC were added to venous blood samples. The samples were stored at -80°C and thawed. 800 μ l 1xSSC were added to each 500 μ l aliquot, contents were mixed and centrifuged at 13000 rpm in a microcentrifuge for 2 minutes. The supernatant was

removed and the pellet was washed twice with 1ml 1xSSC. The pellet was re-suspended in 300µl 1xSSC. 35µl 10% SDS, 20µl 2-mercaptoethanol (1M final concentration) and 5µl freshly prepared proteinase K (20 mg/ml) were added and the sample was incubated 15 to 30 minutes at 37°C, with occasional mixing. 300µl of phenol/chloroform were added to each tube, mixed gently until emulsified and centrifuged 13000 rpm for 1 minute. The organic layer was re-extracted with 1xSSC/0.2% SDS, and all of the sample collected from the organic layer was subjected to a second round of phenol/chloroform extraction. DNA was precipitated with two volumes of absolute ethanol, rinsed with 80% ethanol, dissolved in ddH₂O, and re-precipitated. The pellet was vacuum-dried and dissolved in 30µl 5mM Tris/HCl pH 7.5.

2.2.2. DNA purification by electroelution

Preparative gel electrophoresis was used to selectively recover DNA fragments (PCR products) of interest. Following electrophoresis under normal conditions in the presence of ethidium bromide, the fragment was excised from the agarose gel. The Dark Reader System (Clare Chemical Research) was used throughout to detect the bands to be purified, eliminating UV damage to DNA.

The DNA band was excised from the agarose gel with a scalpel and transferred to a well cut within a second gel of identical agarose concentration and no ethidium bromide, slightly wider than the excised block itself. A piece of dialysis membrane was prepared by boiling in 5mM EDTA for 5 minutes and soaking in the electrophoresis buffer prior to use, and inserted into the gel slot curled under, and folded over the excised band. The gel was run at 150V allowing the DNA to pass out of the gel and become electroeluted onto the membrane. Current was run for 10 to 30 minutes depending on the size of the DNA fragment. With continuous application of the current, the DNA was recovered by swiftly removing the membrane, with the DNA adhered to it, from the gel and placing it into a microcentrifuge tube with a corner of the membrane trapped in the lid. The DNA was centrifuged off the membrane at 15000 rpm in a microcentrifuge for 3 minutes. The DNA was recovered from the eluate by ethanol precipitation.

2.2.3. DNA amplification by PCR

2.2.3.1. PCR buffer

1xPCR buffer (Jeffreys *et al.*, 1990) was prepared by R. Neumann (Department of Genetics, Leicester) as indicated below. dNTPs and BSA were supplied by Pharmacia.

Component	Concentration of Stock Solution	Volume (arbitrary units)	Final Concentration in PCR Reaction
Tris/HCl pH 8.8	2 M	167	45 mM
Ammonium Sulphate	1 M	83	11 mM
MgCl ₂	1 M	33.5	4.5 mM
2-mercaptoethanol	100%	3.6	6.7 mM
EDTA pH 8.0	10 mM	3.4	4.4 μM
dATP	100 mM	75	1 mM
dCTP	100 mM	75	1 mM
dGTP	100 mM	75	1 mM
dTTP	100 mM	75	1 mM
BSA	10 mg/ml	85	113 μg/ml
Total Volume		676	

2.2.3.2. General PCR

10 to 50ng DNA input was PCR amplified (Saiki *et al.*, 1988) in 0.2ml tubes on a MJ Research PTC-225 Tetrad DNA Engine. PCRs were carried out in 10 μl or 20 μl reactions with 11.1xPCR buffer, supplemented with 12mM TrisBase, 0.2 μM each primer, 0.03U/μl *Taq* polymerase and 0.012U/μl *Pfu* polymerase per reaction. To minimise contamination, reagents and materials used for PCR were kept separate from general laboratory chemicals, and PCRs were set up in a category II laminar flow hood. All primers were designed from the current consensus MHC sequence (MHC Sequencing Consortium, 1999) and are listed in Appendix 1.

2.2.4. SNP discovery

The NCBI SNP database (Sherry *et al.*, 2001) was searched for SNPs in the *COL11A2-DOA* region, using the BLAST option. PCR primers for genotyping (see section 2.2.5) were designed around suitably spaced clusters of database SNPs.

For SNP discovery through re-sequencing, PCR primers were designed to amplify 2-3 kb long, suitably spaced DNA segments. Genomic DNA of six to eight semen donors from the UK North European semen donor panel was PCR amplified, and the DNA was purified by electroelution. For re-sequencing, 12.5ng of DNA was used per kilobase of template, plus 0.2 μM sequencing primer and 4 μl BigDye Terminators (ABI PRISM BigDye™ Terminator Cycle Sequencing Ready Reaction Kit, version 1.0) in a total volume of 10 μl, with the following cycling profile: 96°C 20 sec, then 25 cycles of [96°C 10 sec, 50°C 10 sec, 60°C 4 min]. Sequencing products were transferred to 1.5ml tubes, and precipitated by adding 10 μl ddH₂O, 2 μl 3M Na-acetate (pH 5.4) and 55 μl 95% EtOH, and leaving on ice for 7 minutes. After centrifuging for 20 minutes at 13000 rpm, the supernatant was removed. The pellet was rinsed with 450 μl 70% EtOH and vacuum-dried. The DNA

segments were sequenced on an PE Applied Biosystems Model 377 DNA Sequencing System in the Protein and Nucleic Acid Laboratory, University of Leicester. Sequence range was determined by eye and edited using FACTURA. SNPs were identified by eye from sequence traces aligned using ABI AUTOASSEMBLER software.

2.2.5. SNP genotyping

2.2.5.1. Dotblots

PCR was performed in 20 μ l reaction volumes on genomic DNA of 50 North European men in the semen donor panel. 5 μ l all-purpose loading dye (30% glycerol in 0.5xTBE plus enough bromophenol blue to give adequate blue colour) were added, and amplification was checked by electrophoresis of 2.5 μ l of PCR product/loading dye mix on an agarose gel. 35 μ l of denaturing mix (0.5M NaOH, 2M NaCl, 25 mM EDTA) per replica filter were added directly to PCRs and mixed well by pipetting up and down. A vacuum was applied to the assembled dotblot manifold with a sheet of HybondTM-Nfp (Amersham) nylon membrane (pre-soaked in ddH₂O), plus two backing sheets of Whatman 3MM chromatography paper. Using a multichannel pipette, 35 μ l denatured DNA was loaded into each well. Once all samples were loaded, each well was washed with 150 μ l 2xSSC to neutralise the DNA. The dotblots were dried at +80°C for 10 minutes and the DNA was covalently linked to the membrane by UV exposure in the RPN 2500 ultraviolet crosslinker (Amersham).

2.2.5.2. ASO hybridisations

SNPs were typed by ASO hybridisation to dotblots of PCR products using a tetramethylammonium chloride (TMAC) hybridisation protocol (Wood *et al.*, 1985) modified as follows. Dotblots were pre-hybridised at 58°C for 10 minutes in 3M TMAC, 0.6% SDS, 10mM sodium phosphate pH 6.8, 1mM EDTA, 0.1% Ficoll400, 0.1% polyvinylpyrrolidone, 0.1% BSA and 4 μ g/ml yeast RNA, and hybridised at 53°C in the same buffer for 1 hour with 2ng/ml ³²P-labelled ASO (18-mer, with the SNP site located 8 nucleotides from the 5' end) in the presence of 40 ng/ml unlabelled allelic ASO of the opposite allele as competitor, plus 10 μ g/ml single-stranded salmon sperm DNA. Membranes were washed in three changes of 3M TMAC, 0.6% SDS, 10mM sodium phosphate pH 6.8, 1mM EDTA at 53°C for 30 mins, rinsed in 3xSSC at room temperature, and autoradiographed. The probe was removed by several changes of boiling 0.1% SDS, and

the same membrane was re-probed with the ASO corresponding to the alternative allele of same SNP site. ASO sequences are listed in Appendix 2.

2.2.6. LD analyses

Diploid SNP genotypes of unknown linkage phase were used to extract the most likely haplotypes for each pair of SNPs (using maximum-likelihood methods, software written by A.J. Jeffreys). Estimates of LD were derived from the most likely haplotypes.

2.2.7. Recombination assays

2.2.7.1. Allele-specific primers

Various 5' tail extensions and PCR conditions were tested to ensure efficient and specific amplification when using allele-specific primers. The optimal annealing temperature of each primer was determined by amplifying genomic DNA at various annealing temperatures from one individual homozygous for the SNP allele at the 3' end of the allele-specific primer, and from another individual homozygous for the alternative allele. Sequences of allele-specific primers used in the various crossover assays are listed in Table 2.1.

2.2.7.2. Linkage phasing of suitable semen donors

Individuals that were heterozygous for selector SNP sites and suitable internal SNP sites were identified from the panel of 50 semen donors. Their linkage phases for the selector SNP sites, *i.e.* haplotypes for these SNP sites, were determined by allele-specific PCR using allele-specific primer pairs F1-R2, F2-R1 and F2-R2 in all four possible combinations of each allelic primer (see Figure 2.1).

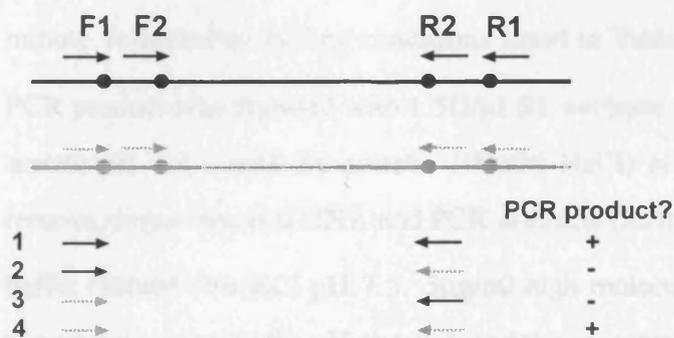


Figure 2.1. Principle of linkage phasing. DNA segments on homologous chromosomes are shown as black and grey lines. Heterozygous selector SNPs are depicted as filled circles, forward allele-specific primers (F1 and F2) and allele-specific reverse primers (R1 and R2) as solid black or dashed grey arrows. To determine the linkage phase of selector SNPs, PCR is performed using all four possible combinations of each allelic forward and reverse primer (with the exception of F1-R1, to avoid generating PCR products that could contaminate crossover assays). An example of amplification using F1-R2 combinations is shown (rows 1-4).

2.2.7.3. Recovering crossovers from sperm

2.2.7.3.1. Digestion and quantification of genomic DNA

Sperm and blood DNA was digested with a restriction endonuclease that cuts immediately outside the recombination assay interval to solubilise the DNA, and to allow more accurate quantification. Suitable restriction enzymes were identified using the MAPLOT option of the Genetics Computer Group (GCG) Sequence Analysis Software Package version 10.0 (Devereux *et al.*, 1984). The same software was used to check that SNPs contained within the assayed interval did not create novel restriction sites for the chosen enzyme. Restriction enzymes used in the crossover assays are listed in Table 2.1. Typically, 20-30 μ g of DNA was digested; digests were carried out in a total volume of 200 μ l with 0.35 U/ μ l of enzyme for 40 minutes at 37°C. Digested DNA was ethanol precipitated, the pellet was vacuum-dried and dissolved in 20 μ l of 5mM Tris/HCl. To quantify the DNA, 0.6, 0.9 and 1.2 μ l aliquots were added to 1ml ddH₂O and the absorbances at 260 nm were measured. The appropriate volume of 5mM Tris/HCl was added to make a working stock DNA concentration (200 to 400ng/ μ l).

2.2.7.3.2. Primary PCR and S1 digestion

To detect recombinant molecules, multiple aliquots of digested sperm DNA, each containing 100 to 30,000 DNA molecules (see Figure 2.2) were amplified by long PCR (Cheng *et al.*, 1994) using allele-specific primers targeted towards outer selector SNP sites in recombinant phase. Blood DNA, at the highest DNA inputs, containing the same selector sites (ideally, from the same donor as the sperm DNA) was used as a negative control throughout. PCRs were carried out in total volumes of 8 μ l; primer sequences used in each recombination assay are listed in Table 2.1. All PCRs were started with an initial denaturation step at 96°C for 1 minute, followed by cycling conditions listed in Table 2.2. A 0.5 μ l aliquot of each primary PCR product was digested with 1.5U/ μ l S1 nuclease in 5 μ l S1 digestion buffer (20mM Na acetate pH 4.9, 1mM Zn acetate, 100mM NaCl) at room temperature for 20 minutes, to remove single-stranded DNA and PCR artefacts (Jeffreys *et al.*, 1998a). 45 μ l of S1 dilution buffer (10mM Tris-HCl pH 7.5, 5 μ g/ml high molecular weight salmon sperm) were added to each tube to bring the pH above 7, and the contents were mixed.

2.2.7.3.3. Secondary PCR

The secondary (nested) PCRs, in total volumes of 8 μ l, were each seeded with 1.6 μ l of S1 digested primary PCR product and carried out using allele-specific primers targeted towards inner SNP selector sites in recombinant phase. As above, all PCRs were started with an initial denaturation step at 96°C for 1 minute, followed by cycling conditions listed in Table

2.2. 3.5µl all-purpose loading dye were added to each tube, and 3µl aliquots were run out on an agarose gel.

If secondary PCR amplification was efficient enough, recombinants were detected simply on an ethidium bromide stained agarose gel. However, the exact number of crossover-positive reactions was usually confirmed by subsequent Southern blot hybridisation (see also Figure 2.2). Occasionally, secondary PCR signals after Southern blot hybridisation were very weak; in these cases tertiary PCR using universal primers was performed on all secondary reactions, and crossover-positive reactions were scored from tertiary PCRs (see below).

2.2.7.3.4. Southern blot transfer and hybridisation to detect recombinants

Southern blot procedures were standard and as described (Jeffreys *et al.*, 1991). For the probe, approximately 10ng of double stranded DNA generated by PCR amplification of the locus of interest were labelled by random hexamer priming (Feinberg and Vogelstein, 1983, 1984) incorporating $\alpha^{32}\text{P}$ -dCTP (Amersham). Southern blot membranes were pre-hybridised for a minimum of 30 minutes at 65°C in 7% SDS, 0.5M sodium phosphate (pH 7.2), 1mM EDTA (modified from Church and Gilbert, 1984). Hybridisation was carried out in the same buffer overnight at 65°C in a Hybaid hybridisation oven. After hybridisation, the membrane was washed at 65°C in two changes of high stringency wash solution (0.1xSSC, 0.01% SDS) and washed briefly in 2xSSC before autoradiography at -80°C.

2.2.7.3.5. Tertiary PCR

Positive secondary PCRs were re-amplified using universal PCR primers located just internal to the secondary PCR primer SNP sites (see Table 2.3 for primer sequences and PCR conditions).

Table 2.1. Restriction enzymes and sequences of allele-specific primers used in crossover assays. Underlined bases in small letters are 5' extensions used to boost GC content, not part of genomic sequence. Adap and Tail are longer extensions, with the following nucleotide sequences: Adap 5' GTCTACGTAGTCAGCTCTGG 3', Tail 5' TGCACATGCCGACCATACGC 3'

	restriction enzyme	forward primers, 5' of the putative hotspot	reverse primers, 3' of the putative hotspot
A. DPB region	<i>Hinc</i> II	-74.16C1 5' GTGTCCACCTTCTCACTCC 3' -74.16A 5' GTGTCCACCTTCTCACTCA 3' -74.02FG 5' CCTGATACTACTCCCTCAG 3' -74.01FC 5' TACTCCCTCAAGTCCCTC 3'	-70.18RG 5' AGGGACAGTGTTGGGAGC 3' -69.56RT 5' CCCTGAGGAATTGCCACAA 3' -70.23RT 5' GGAGCATGTGATGCTGGA 3' -70.11RC 5' GAACAGTGCAGGAGGAG 3'
B. 3' DPA region	<i>Bst</i> Z171	D-50.63FG 5' GATAGATCTCCTATTCCCTG 3' D-50.63FT 5' GATAGATCTCCTATTCCCTT 3' D-50.60FC 5' <u>cc</u> CTTTACTTGAAGCTGTAC 3' D-50.60FT 5' <u>cc</u> CTTTACTTGAAGCTCTAT 3' D-50.34FG 5' <u>Adap</u> ATTTATGAGAAAAGTGTACG 3' D-50.53F+ 5' <u>Tail</u> AAAGGAAAATGGAAAACAGG 3' D-50.53F- 5' <u>Tail</u> AAAGGAAAATGGAAAACAAA 3'	D-46.06RC 5' GCTTATGGCAGTGATCAAGG 3' D-46.06RG 5' GCTTATGGCACTGATCAAGC 3' D-46.03RG 5' CTCACCTGCCTGGCCAC 3' D-46.03RA 5' GCTCTCTGCCTGGCCAT 3'
C. DMB region	<i>Xmn</i> I	JJ6FC* 5' CTGCTGGTGGTGTGGC 3' JJ6FT 5' GCTCTGGTGGTGTGGT 3' JJ7FA 5' <u>ccccg</u> CTTGCTTTGAAATGAGGA 3' JJ7FC 5' <u>cccc</u> CTTGCTTTGAAATGAGGC 3'	J6RA 5' ATAACCCCTCTGTCCCT 3' J6RG 5' ATAACCCCTCTGTCCCC 3' J7RC 5' GCCTAAGAGCAGAGGGAG 3' J7RT 5' GCCTAAGAGCAGAGGGAA 3'

* note regarding nomenclature: JJ6FC stands for Forward primer, targeted towards the C allele of SNP JJ6.

crossover positive reactions recovered, respectively) is used as an example throughout sections 2.2.7.4 and 2.2.7.5.

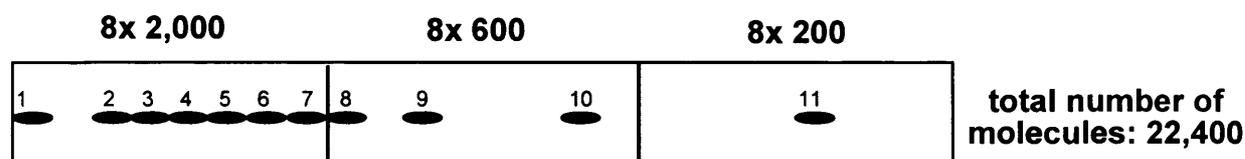


Figure 2.2. Schematic representation of crossover-positive secondary PCRs, after Southern blot hybridisation, at various input DNA pool sizes. For each input pool size of 2000, 600 and 200 DNA molecules, eight PCRs were set up in parallel. Here, eleven reactions (labelled 1-11) are shown as positive for a crossover (for example, 7/8 reactions at 2000 molecule input).

The frequency of crossovers (m) in input DNA was estimated for each input pool by Poisson approximation:

$$m = -\ln(\text{number of negative reactions} / \text{total number of reactions}).$$

Thus, the input pools shown in Figure 2.2 would contain 2.08, 0.47 and 0.13 crossovers per reaction at 2000, 600 and 200 molecule input, respectively. Maximum-likelihood methods (software written by A.J. Jeffreys) were used to estimate the mean frequency of crossovers, combining data from all input pool sizes. The mean crossover frequency calculated for the example data set in Figure 2.2 is 0.89×10^{-3} , with upper and lower 95% confidence intervals of 0.41×10^{-3} and 1.72×10^{-3} .

2.2.7.5. Crossover breakpoint mapping

Denatured PCR products were transferred onto nylon membranes and the status of internal SNP sites was determined by ASO hybridisation as described in section 2.2.5. An example of crossover scoring is given in Figure 2.3 and Table 2.4.

2.2.7.6. Calculation of crossover activity in each inter-SNP interval

Crossover-positive reactions can either be "simple" (see *e.g.* dot number 11 in Figure 2.3B-C), or "mixed", *i.e.* giving a positive signal upon hybridisation with both alleles of a SNP site (see *e.g.* dot number 1 in Figure 2.3B-C). "Simple" ones contain only one class of crossover; they might contain more than one crossover molecule, but all crossovers have their breakpoint in the same inter-SNP interval. "Mixed" reactions contain two or more different crossovers. They were scored as two crossovers, one with a breakpoint 5' to where a succession of one or more mixed SNP sites starts, and the other 3' to where the mixture finishes (Table 2.4B). The estimate of the number of crossovers will always be conservative - simple reactions may contain more than one crossover, but only one will be scored; mixed reactions may contain more than two crossovers, but only two will be scored. The true number of crossovers may therefore be higher, and can be again estimated by Poisson correction.

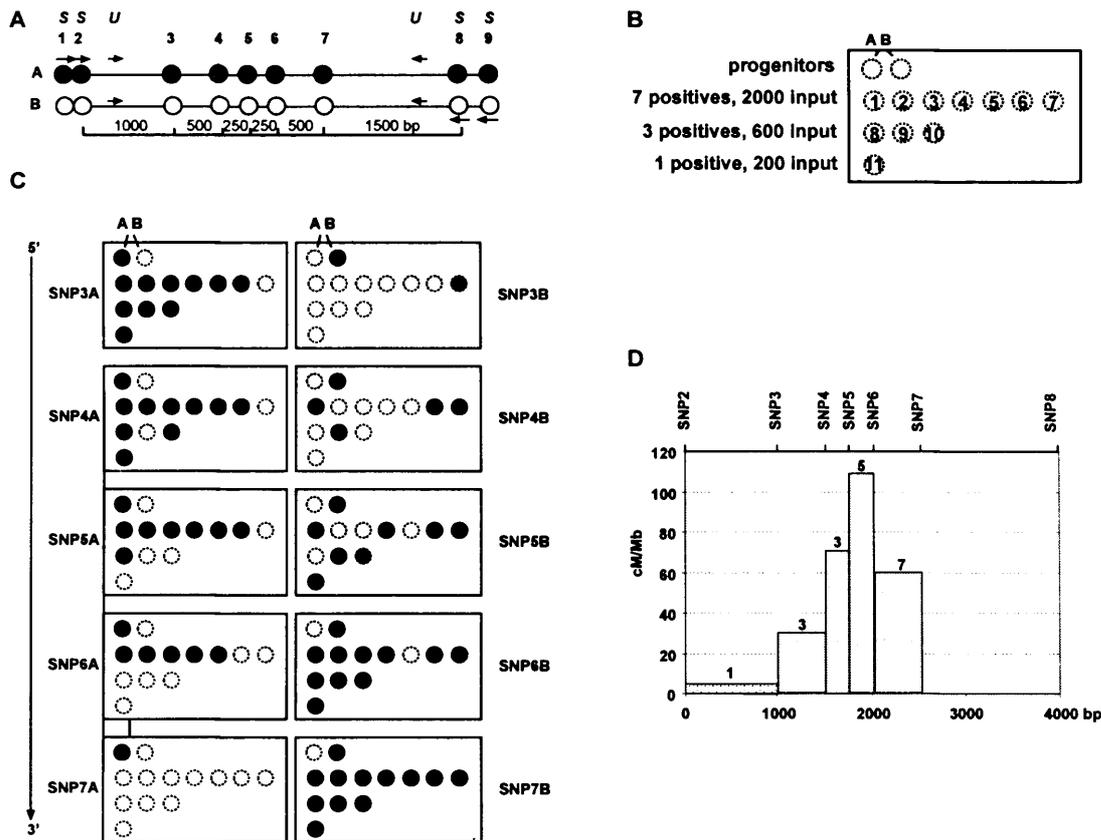


Figure 2.3. Principle of crossover breakpoint mapping, (A) In the crossover assay, if forward primers are allele-specific for the A haplotype (dark grey), and reverse primers for the B haplotype (light grey), "A to B" crossovers are amplified. Crossover positive reactions are re-amplified with universal primers. Five internal heterozygous SNP sites (3-7) and their inter-marker distances are shown. SNPs 1, 2, 8 and 9 are selector SNP sites for allele-specific PCR. S, allele-specific primer, U, universal primer. (B) Dotblot positions of progenitor haplotypes (A and B), plus of the 11 hypothetical crossover positives from the different input pools (1-11) from figure 2.2 are shown. (C) Dotblots were probed with ASOs targeted towards heterozygous SNP sites (5' to 3', top to bottom: SNPs 3-7). Dotblots on the left were probed with ASOs specific for haplotype A alleles, dotblots on the right for haplotype B alleles. Filled circles are positive for that particular haplotype, open dashed circles are negative. See Table 2.4A for crossover scoring. (D) Grey bars indicate recombination activity for each SNP interval (see table 2.5A-B, and below, for calculations). Number of crossovers (rounded to nearest full number) observed in each interval are shown above the bars.

Crossovers were occasionally observed in the end intervals (furthest 5' or 3' SNP interval assayed by ASO hybridisation, see dot 7 in Figure 2.3C). Crossovers of this type could either have their breakpoint in the interval between the inner selector site and the nearest internal heterozygous SNP site, or they could be a PCR artefact, arising from jumping PCR, resulting in the bleed-through amplification of one haplotype. The latter is unlikely if blood PCRs do not show the same pattern. Whenever crossovers mapping to either end interval were found, they were scored but interpreted with caution.

Once mixtures had been scored, crossovers were ordered in the table according to the location of their breakpoints, 5' to 3', and the number of crossovers for each SNP interval was counted (Table 2.4C). For each SNP interval, maximum-likelihood methods (software written by A.J. Jeffreys) were then used to estimate the mean frequency of crossovers in that

interval, taking into account in which input pool the crossovers were observed (Table 2.5A). The number of molecules screened for each SNP interval was adjusted where mixtures were observed: pools that contained crossover mixtures were excluded for all SNP sites that lie between the two intervals where crossovers were scored (see intervals SNP2-SNP3 and SNP3-SNP4 in Table 2.5A). This is because any additional crossovers occurring between the two breakpoints could not be detected. For each SNP interval, the maximum-likelihood Poisson estimate of recombination frequency was then multiplied by the number of molecules screened in that interval, to yield the Poisson-corrected number of crossovers (Table 2.5B). Recombination activity in cM/Mb for each SNP interval was calculated as follows:

(Poisson-corrected number of crossovers/total number of molecules screened) x 100 = recombination fraction in cM,

then:

(recombination fraction in cM/length of interval in bp) x 10⁶ = recombination activity in cM/Mb.

Recombination activity across the example interval in Figure 2.3A is shown in Figure 2.3D and Table 2.5B.

Table 2.4. Example of crossover scoring. (A) In the crossover assay example given in Figure 2.2, 8 pools of 2000, 600 and 200 amplifiable molecules each were screened (24 crossover PCR, 22,400 molecules in total). Eleven out of the 24 PCRs were positive for crossovers (dots 1-11 on blot in Figure 2.3B, and rows labelled 1-11 in table). Progenitor haplotypes (A and B, dark and light grey) are shown at the top. Status of heterozygous SNPs is scored on dotblots for each crossover-positive dot (Figure 2.3C). Five (dots 1-4 and 6) of the eleven positive reactions contain two or more crossovers (a crossover mixture, M). (B) Resolution of crossover mixtures. For example reaction 1 was resolved as 1a and 1b (see text). Note that dot 1 could contain a crossover in intervals 2 to 4 and/or 4 to 5, but they would not be detected. (C) Crossovers ordered according to breakpoint location, 5' to 3', and numbers of crossovers per SNP interval (most 5' interval is allele-specific forward primer to SNP3, most 3' is SNP7 to allele-specific reverse primer). The single crossover observed in the 5' end interval is shown in brackets.

		heterozygous markers						
		(SNP2)	SNP3	SNP4	SNP5	SNP6	SNP7	(SNP8)
Progenitor A			A	A	A	A	A	
Progenitor B			B	B	B	B	B	
	Number on blot	input						
A	1	2000	A	M	M	M	B	
	2	2000	A	A	A	M	B	
	3	2000	A	A	A	M	B	
	4	2000	A	A	M	M	B	
	5	2000	A	A	A	A	B	
	6	2000	A	M	M	B	B	
	7	2000	B	B	B	B	B	
	8	600	A	A	A	B	B	
	9	600	A	B	B	B	B	
	10	600	A	A	B	B	B	
	11	200	A	A	B	B	B	
B	1a	2000	A	A	A	A	B	
	1b	2000	A	B	B	B	B	
	2a	2000	A	A	A	A	B	
	2b	2000	A	A	A	B	B	
	3a	2000	A	A	A	A	B	
	3b	2000	A	A	A	B	B	
	4a	2000	A	A	A	A	B	
	4b	2000	A	A	B	B	B	
	5	2000	A	A	A	A	B	
	6a	2000	A	A	A	B	B	
	6b	2000	A	B	B	B	B	
	7	2000	B	B	B	B	B	
	8	600	A	A	A	B	B	
	9	600	A	B	B	B	B	
	10	600	A	A	B	B	B	
	11	200	A	A	B	B	B	
C	7	2000	B	B	B	B	B	
	1b	2000	A	B	B	B	B	
	6b	2000	A	B	B	B	B	
	9	600	A	B	B	B	B	
	4b	2000	A	A	B	B	B	
	10	600	A	A	B	B	B	
	11	200	A	A	B	B	B	
	2b	2000	A	A	A	B	B	
	3b	2000	A	A	A	B	B	
	6a	2000	A	A	A	B	B	
	8	600	A	A	A	B	B	
	2a	2000	A	A	A	A	B	
	1a	2000	A	A	A	A	B	
	3a	2000	A	A	A	A	B	
	4a	2000	A	A	A	A	B	
	5	2000	A	A	A	A	B	
crossovers per interval		(1)	3	3	4	5	0	total: 16

Table 2.5. Calculating crossover frequencies for each SNP interval, based on the number of molecules screened. (A) Total number of positive plus negative reactions for each input pool, per SNP interval. Crossover-positive pools that contain crossover mixtures spanning more than one interval (in grey italics) have been excluded for all SNP sites that lie between the two SNP intervals where crossover breakpoints were scored. For example, crossover positive number 6 was excluded for the SNP4-SNP5 interval. The total number of molecules screened is lower for the SNP4-SNP5 and SNP5-SNP6 intervals (open box), as follows: 22,400 - (2x2000)=18,400, because two of the 2000 molecule input pools were excluded for these intervals. (B) Recombination activity per interval is calculated from the Poisson-corrected number of crossovers and inter-marker distance (see text). Where only one crossover was observed in an interval (here, see SNP2-SNP3 interval), it was not Poisson-corrected because of the small sample size. End interval crossovers (bracketed) were scored as normal but usually excluded from further analyses (see text).

input pool	number of crossover positive + negative reactions for each SNP interval, in each input pool					
	SNP2 to SNP3	SNP3 to SNP4	SNP4 to SNP5	SNP5 to SNP6	SNP6 to SNP7	SNP7 to SNP8
2000	1+7	2+6	1+5	3+3	5+3	0+8
600	0+8	1+7	1+7	1+7	0+8	0+8
200	0+8	0+8	1+7	0+8	0+8	0+8
molecules screened	22400	22400	18400	18400	22400	22400
			<i>exclude</i>	<i>exclude</i>		
			<i>crossover positives</i>	<i>crossover positives</i>		
			<i>1 and 6</i>	<i>1 and 4</i>		
Poisson-est. crossover frequency	-	0.00015	0.00018	0.00027	0.00030	-

	SNP2 to SNP3	SNP3 to SNP4	SNP4 to SNP5	SNP5 to SNP6	SNP6 to SNP7	SNP7 to SNP8	
uncorrected number of crossovers	(1)	3	3	4	5	0	total: 16
Poisson-corr. number of crossovers	0	3.4	3.3	5.0	6.7	0	total: 19.4
inter-marker distance (bp)	1000	500	250	250	500	1500	
cM/Mb	(4.5)	30	72	108	60	0	

3. SNP DIVERSITY IN THE MHC CLASS II REGION

3.1. INTRODUCTION

Single-nucleotide polymorphisms (SNPs) are the most common type of genetic variant, and are usually defined as single nucleotide differences between individuals, where the rarer allele reaches a frequency of at least 1% in the population. Thus, SNPs are bi-allelic markers; the appearance of a SNP is considered a unique event in the history of a chromosome. It is now estimated that the total number of SNPs in the 3.2 billion basepair long human genome is >10 million, *i.e.* that there is at least one SNP every 300bp (National Center for Biotechnology Information SNP database, <http://www.ncbi.nlm.nih.gov/SNP/>). SNPs with a minor allele frequency >20% are found roughly at a density of one per kilobase. Because the vast majority of the human genome is non-coding (International Human Genome Sequencing Consortium, 2001), most SNPs are expected to be selectively neutral. The human genome contains perhaps less than 400,000 coding SNPs (Cargill *et al.*, 1999).

Variation on the single nucleotide level was already routinely detected in the 1980s in the form of restriction fragment length polymorphisms (RFLPs, Botstein *et al.*, 1980). Of course, these encompass only a fraction of SNPs. Microsatellites and RFLPs were successfully used in the 1990s for the linkage mapping of monogenic (mendelian) diseases in families. Since the late 1990s, SNPs have experienced a renaissance as the marker of choice. This is due to the growing interest in the genetics of "common diseases", such as osteoporosis and diabetes, that are inherited in a non-mendelian fashion (Risch, 2000). These diseases show polygenic inheritance, with each gene contributing modestly to the outcome, and are also called complex or multifactorial diseases. SNPs are abundant in the genome, and can be used for the search for complex disease determinants in two ways: directly, where functional variants at genes of interest are tested, or indirectly, through LD or association mapping that tests if a certain combination of SNP alleles (a haplotype) is more common in patients than control individuals (Collins *et al.*, 1997). SNPs also hold great promise for pharmacogenetics. Here the aim is to identify variants that can be used to predict a patient's response to drugs, and is similar in approach to complex disease mapping. There are now a number of different (often high-throughput) methods to detect SNPs. For example, a hybridisation or primer extension reaction can be performed on the DNA sample, and the product detected by fluorescence or mass spectrometry (reviewed in Syvänen, 2001).

Miller and Kwok (2001) describe the "life cycle" of a SNP as consisting of four stages. Initially, a SNP enters the population as simply a mutational change that occurs on a single chromosome. This new allele then manages to survive, against odds, and subsequently increases in frequency. A SNP's life cycle comes to an end when the allele

becomes fixed, *i.e.* the site is no longer polymorphic. SNPs that have a low minor allele frequency in a given population (rare SNPs) are more likely to be "young" and have entered the population fairly recently. Alternatively, a rare SNP could be reaching the end of its life cycle, and be near fixation. Ancestral states (which allele of a SNP "was there first") can be determined by typing the SNP site in chimps or other great apes. Rare SNPs may be present in certain populations and absent from others. In contrast, SNPs with relatively high minor allele frequency (common SNPs) are often shared between populations and are therefore presumably ancient.

A "newborn" SNP is initially found only on one chromosomal background (haplotype). It is associated with (is in LD with) certain alleles on that haplotype until it is reshuffled onto different haplotypes by recombination. The population frequency of an allele, which itself is selectively neutral, can be influenced by selection if it is linked to an allele that is selected for (genetic hitchhiking) or against (background selection). DNA segments that are recombinationally inert are expected to be more prone to so-called selective sweeps - whether by hitchhiking or background selection - and to therefore show less diversity (reviewed in Nachman, 2002).

Not all sites are equally SNP prone. CpG to TpG or CpA mutations are the most frequent kind in humans, resulting from the deamination of 5-methylcytosine (Cooper and Krawczak, 1989) and accounting for ~25% of all mutations (Wang *et al.*, 1998). SNPs are found less frequently in coding than non-coding sequences, presumably due to selective constraints. SNPs have been systematically identified in 5' and 3' untranslated region, introns and coding sequence of human candidate genes involved in complex common diseases, *e.g.* endocrinological and neuropsychiatric disorders (Cargill *et al.*, 1999) and blood-pressure homeostasis (Halushka *et al.*, 1999). Both these studies, as well as subsequent analysis of SNPs in expressed sequence tags by Sunyaev *et al.* (2000), found that levels of nucleotide diversity were highest at fourfold degenerate coding or synonymous sites, and lowest at non-degenerate/non-synonymous sites.

Across large genomic regions, SNP incidence is not constant, either. At eight X-linked loci SNP density varied from none in 1900 basepairs to one every 190 basepairs (Nachman *et al.*, 1998). On the long arm of the X chromosome, regions 211-kb to 1.66 Mb in length are almost or completely devoid of common SNPs, and were hence termed SNP deserts (Miller *et al.*, 2001). The lack of variation in Caucasians in these chromosomal segments was explained by recent common ancestry. The SNPs were not shared by orangutans and humans, indicating that a SNP's lifespan is shorter than the time of divergence between the two species. This is not surprising, given that for any new SNP allele entering the population the expected time (t) to fixation is $t=4N_e$ x generation time in years, where N_e is effective population size, estimated to be 10 000 in humans (Morton, 1982). If generation time is calculated to be 20 years, the time to fixation would be 800 000 years, a time far shorter than the 5 million years since the ape/human divergence.

The fact that polymorphisms are not shared between humans and apes holds true for most genomic regions, except some MHC alleles, which are presumably ancient, older than these species themselves (reviewed by Li, 1997). An uneven distribution of SNPs has also been noted in the human MHC. When sequences of two HLA haplotypes are compared, nucleotide diversity varies by an order of magnitude within the class II region (Gaudieri *et al.*, 2000).

3.1.1. Measures of nucleotide diversity

3.1.1.1. Nucleotide heterozygosity π

Nucleotide heterozygosity, π (Nei and Li, 1979) is the average pairwise difference between two randomly selected sequences, expressed per site. In other words, it is the probability that two randomly chosen homologous nucleotides from the sample are different. π is therefore dependent on both the number of polymorphisms and their population frequency. Heterozygosity per base pair is estimated according to the following formula:

$$\pi = \frac{1}{L} \frac{n}{n-1} \sum_{p,q} x_p x_q \pi_{pq}$$

where L is the total number of SNPs considered, n is the number of sequences, x_p and x_q are frequencies of variants p and q among n sequences, and π_{pq} is the number of nucleotide differences between alleles (which is always 1 if individual SNPs are considered).

Heterozygosity in humans is on average low, estimated at $\pi=0.75\%$ (Bamshad and Wooding, 2003); this equals one nucleotide difference per 1.3 kb between two randomly sampled individuals. There is apparently a strong positive correlation between nucleotide variability and recombination rate on the megabase level (Nachman, 2001).

3.1.1.1.2. Watterson's theta, θ

SNP diversity can be characterised in terms of K , the observed number of variant sites. K increases with the number of chromosomes (n) analysed and the total sequence length screened (L). Watterson's θ (Watterson, 1975) is a measure of nucleotide variability that does not depend on allele frequency, and is estimated for a sample of non-recombining DNA. The normalised number of variant sites is used to correct for sample size, as follows:

$$\theta = \frac{1}{L} \frac{K}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

This work

When this work was started, previous data generated in our laboratory had verified recombination hotspots in the *DOA-RING3* interval as well as in the *TAP2* gene region (see

Chapter 1). Therefore, to complete the picture of LD and recombination hotspots, the target region of this study consisted of two segments - the 100 kb region upstream of the *DOA* gene, and the 80 kb *RING3-TAP2* interval. As a prelude to LD analyses, SNPs were identified across these two regions. This was done initially through re-sequencing of suitably spaced segments spanning this region, and at later stages, through NCBI SNP database searches, when this resource became available (see Figure 3.1). I demonstrate the presence of highly and moderately polymorphic sub-regions, and a discrepancy between SNP density and heterozygosity, particularly evident in the highly polymorphic DNA segments.

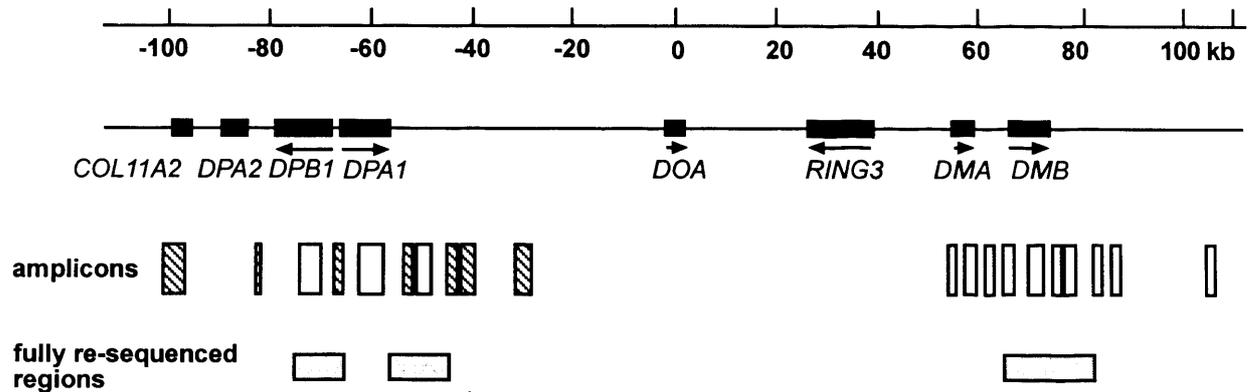


Figure 3.1. Overview of the target region. Amplicons were designed for re-sequencing and/or SNP genotyping in the *COL11A2-DOA* interval and the *RING3-TAP2* interval. Amplicons are shown as white or hatched boxes, hatched boxes indicate amplicons that were designed around database SNPs. Grey boxes are regions that were subsequently fully re-sequenced in six to eight semen donors.

3.2. RESULTS

SNP discovery and subsequent genotyping was undertaken in the two target regions, the *COL11A2* to *DOA* and *RING3* to *TAP2* intervals (Figure 3.1). For historical reasons, position zero (Figure 3.1) is at the *DOA* gene. The location of the starting point of coordinates has no significance, and in principle could be randomly assigned to any position. All positions upstream of point zero (including all amplicons in the *COL11A2-DOA* interval) have a negative sign in front of their genomic location (Figure 3.1).

3.2.1. SNPs and their allele frequencies

Approximately 250 kb of DNA were re-sequenced. 353 SNPs were identified in total. Of these, 40 SNPs were identified by searching the NCBI SNP database (Sherry *et al.*, 2001), and the rest through re-sequencing. Altogether 250 SNPs were genotyped (see Table 3.1). An example of a SNP identified through re-sequencing is shown in Figure 3.2A, and the genotyping of this SNP by ASO hybridisation on dotblots is shown in Figure 3.2B.

For SNP genotyping by ASO hybridisation, long (2-4 kb) DNA segments, each containing multiple SNP sites, were PCR amplified. Dotblots of these PCR products were then probed sequentially with ASOs specific first for one allele of a given SNP, then for the

other allele (Figure 3.2B). This method is simple, robust and accurate; SNP genotypes could be assigned unambiguously >98% of the time. ASO hybridisation on dotblots requires no expensive reagents and with multiplexing, one person can generate up to 800 genotypes a day. This method is particularly well suited for applications that require typing of a large number of SNPs in a relatively small number of individuals.

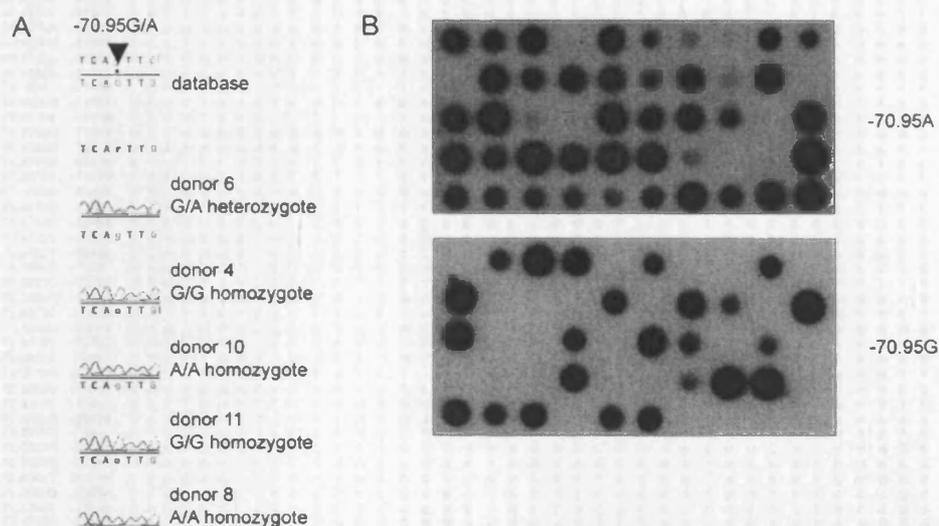


Figure 3.2. Example of SNP identification and genotyping. (A) Sequence traces around SNP -70.95G/A. Donor 6 is a heterozygote and shows a mixed G/A base. Donors 4 and 11 are G/G homozygotes, while donors 10 and 8 are A/A homozygotes. (B) Genotyping of -70.95G/A by ASO hybridisation. Dots are PCR amplified DNA, containing this SNP site from donors 1-50 (in five rows of ten donors each), probed with -70.95A (top), then -70.95G (bottom).

SNPs in the *COL11A2-DOA* interval were named after their approximate position, rounded to the nearest ten basepairs; for example SNP -70.95G/A is located at position -70948. SNPs in the *RING3-TAP2* interval have names beginning with a letter, followed by a number and then alternative alleles, for example SNP J2A/G is located in an amplicon that was called "J", and was the second SNP to be identified in that amplicon.

Table 3.2. Structure of insertion/deletion polymorphisms. Where the polymorphisms were typed by ASO hybridisation, the ASO sequence to detect the insertion (+) and deletion (-) allele is shown. For larger insertion/deletions (SVA1, -53.27 and SVA2) the length is shown. Polymorphic sites in the ASO sequence are shown in bold.

polymorphism	location	allele	ASO structure
-74.08+/-	-74074	+	ATTTTTTCCTCTTTCTA
		-	ATTTTTTCCTCTTTCTAG
-70.69+/-	-70691	+	CATTTTTTTTTGTATATA
		-	CATTTTTTTTTGTATATAA
-68.45+/-	-68454	+	CCATTTTCCCAAAGACA
		-	CCCATTTTCCCAAAGACA
-68.44+/-	-68443	+	CCAAAGACAAGCATACTG
		-	CCAAAGACACTGACCACT
SVA1+/-	-54646	+	(1.8 kb insertion)
		-	(1.8 kb deletion)
-53.58+/-	-53577	+	GAATACAAAAGAAAATT
		-	GAATACAAAAGAAAATTA
-53.42+/-	-53422	+	GGGGGGCGGGCATTTTAC
		-	GGGGGGAGCATTTTACTT
-53.27+/-	-53265	+	(51 bp insertion)
		-	(51 bp deletion)
-53.21-/+	-53218	+	TTTTTCCCTAAKGGCTGT
		-	TTTTTCCCTAAKGGCTGTT
SVA2+/-	-52947	+	(1.8 kb insertion)
		-	(1.8 kb deletion)
-50.75-/+	-50736	+	CAACTGGATAAGATCCAA
		-	ACAACTGGATAAGACCAA
-50.74+/-	-50732	+	CCAATAGTTTGATATAC
		-	CCAATAGTTTGATATAC
-50.66+/-	-50663	+	TTCTGTATAAGGAAGGT
		-	TTTCTGTTAAGGAAGGT
-50.53+/-	-50538	+	AAAACAGGAGTAAAATAG
		-	GAAAACAAAGTAAAATAG
-50.32-/+	-50325	+	TACATGATTGATTCCTT
		-	TACATGATGATTCCTTT
-49.66+/-	-49655	+	CAGTTCTGTTTCATAGTG
		-	CAGTTCTTTTCATAGTGG
-49.62-/+	-49630	+	ATCAGAGTGTACGTGGA
		-	ATCAGAGGTGTACGTGG
Jjk11+/-	75062	+	GTGAGAGAAAGAGAGGAA
		-	GTGAGAGAGAGAGGAAAA
Jjk12+/-	75079	+	AAAAGAGAAGGAGTGGAA
		-	AAAAGAGGAGTGGAAAGAG

Allele frequencies for major and minor alleles at each SNP site were calculated (see Table 3.3). When ordered into allele frequency bins of 10%, most SNPs show minor allele frequencies of 10 to 20% (Figure 3.3). The second most common minor allele-frequency class is 1 to 10%.

Table 3.3. SNPs and their minor allele frequencies (MAF) in the two intervals studied. SNP types: CTi, transition at CpG doublet, Ti, other type of transition, Tv, transversion, ID, insertion/deletion.

COL 11A2-DOA interval								
SNP	type	MAF	SNP	type	MAF	SNP	type	MAF
-100.49G/A	CTi	0.45	-69.23A/T	Tv	0.12	-49.91C/T	Ti	0.14
-99.65G/A	CTi	0.19	-69.02C/T	Ti	0.01	-49.70G/C	Tv	0.15
-97.62C/T	CTi	0.19	-68.68A/G	Ti	0.14	-49.66+/-	ID	0.46
-97.55T/C	CTi	0.27	-68.54C/T	CTi	0.14	-49.62-/+	ID	0.15
-96.42G/A	Ti	0.19	-68.47A/T	Tv	0.14	-49.58T/C	Ti	0.14
-82.49A/G	Ti	0.20	-68.45+/-	ID	0.02	-49.57T/C	Ti	0.15
-82.37G/A	Ti	0.20	-68.44+/-	ID	0.48	-49.54A/G	Ti	0.07
-74.16A/C	Tv	0.10	-68.35G/A	Ti	0.01	-49.49T/C	CTi	0.14
-74.08+/-	ID	0.02	-68.21T/C	CTi	0.15	-49.43G/A	Ti	0.14
-74.02A/G	Ti	0.03	-67.80G/C	Tv	0.19	-49.38A/G	CTi	0.02
-74.01T/A	Tv	0.11	-67.19T/C	Ti	0.14	-49.34A/G	Ti	0.16
-73.85G/C	Tv	0.12	-66.77A/G	Ti	0.14	-49.33A/G	Ti	0.14
-73.83A/C	Tv	0.01	-66.69G/C	CTi	0.14	-49.32A/T	Tv	0.12
-73.57T/C	CTi	0.02	-66.65T/C	Ti	0.14	-49.28G/A	Ti	0.12
-73.56T/A	Tv	0.08	-66.07A/G	CTi	0.18	-49.26G/A	CTi	0.12
-73.49G/A	CTi	0.29	-61.37A/G	Ti	0.15	-49.23A/G	CTi	0.12
-73.48T/C	Ti	0.02	-61.19T/G	Tv	0.15	-49.17G/T	Tv	0.38
-73.47C/T	Ti	0.01	-60.29T/A	Tv	0.15	-49.13G/C	Tv	0.22
-73.41C/T	Ti	0.02	-59.15C/T	CTi	0.15	-49.02C/T	CTi	0.03
-73.40C/T	Ti	0.02	-55.89T/C	CTi	0.15	-48.81G/A	CTi	0.03
-73.36C/A	Tv	0.11	-55.74T/C	CTi	0.15	-48.83T/A	Tv	0.20
-73.29T/C	CTi	0.03	-55.70G/A	CTi	0.10	-48.77T/C	Ti	0.50
-73.22G/C	Tv	0.12	-55.65A/G	Ti	0.15	-48.71T/C	Ti	0.10
-73.20T/C	Ti	0.14	-55.60C/T	CTi	0.15	-48.65C/T	CTi	0.44
-73.04A/G	CTi	0.01	SVA1+/-	ID	0.15	-48.53A/T	Tv	0.40
-72.94C/T	Ti	0.13	-54.64A/G	CTi	0.15	-48.47T/C	Ti	0.39
-72.92T/C	Ti	0.28	-54.43C/T	Ti	0.15	-48.45G/A	CTi	0.38
-72.91T/G	CTi	0.16	-54.17C/G	Tv	0.15	-48.41G/A	Ti	0.35
-72.89C/G	Tv	0.15	-53.94C/T	Ti	0.15	-48.40A/G	Ti	0.36
-72.88G/T	Tv	0.45	-53.81bC/T	Ti	0.15	-48.35G/A	CTi	0.13
-72.83A/G	CTi	0.45	-53.68A/T	Tv	0.15	-48.29A/G	CTi	0.08
-72.82A/T	Tv	0.11	-53.58+/-	ID	0.31	-47.97C/T	Ti	0.19
-72.75C/A	Tv	0.23	-53.55G/A	CTi	0.15	-47.95C/A	Tv	0.19
-72.74A/C	Tv	0.25	-53.42+/-	ID	0.15	-47.93G/A	Ti	0.19
-72.66C/T	Ti	0.11	-53.38T/C	CTi	0.14	-47.89A/G	Ti	0.19
-72.65C/G	Tv	0.02	-53.27+/-	ID	0.18	-47.31G/T	Tv	0.22
-72.63C/T	Ti	0.09	-53.21-/+	ID	0.15	-47.04C/T	Ti	0.20
-72.47G/A	Ti	0.02	-53.17G/A	Ti	0.15	-46.89A/G	Ti	0.20
-72.26C/T	Ti	0.01	-53.11T/C	Ti	0.15	-46.22G/A	Ti	0.35
-72.18G/T	Tv	0.07	-53.09G/A	CTi	0.15	-46.06C/G	Tv	0.20
-72.06A/G	Ti	0.01	-53.08T/G	Tv	0.15	-46.05C/G	Tv	0.21
-71.93A/T	Tv	0.02	-53.03C/T	CTi	0.15	-46.04G/A	Ti	0.20
-71.90G/C	Tv	0.14	-53.02A/C	Tv	0.15	-46.02T/A	Tv	0.20
-71.75A/G	CTi	0.23	SVA2+/-	ID	0.15	-46.01A/G	CTi	0.20
-71.71T/G	Tv	0.28	-51.22A/C	Tv	0.15	-45.98C/T	CTi	0.20
-71.30A/G	Ti	0.26	-51.13T/C	CTi	0.21	-45.89A/G	CTi	0.21
-71.19A/T	Tv	0.14	-51.08C/T	CTi	0.15	-45.71A/T	Tv	0.2
-71.04C/G	Tv	0.12	-50.94G/A	Ti	0.17	-45.67G/A	Ti	0.21
-71.03A/G	Ti	0.22	-50.84G/A	CTi	0.15	-45.66T/A	Tv	0.37
-71.01A/G	Ti	0.14	-50.76T/C	Ti	0.15	-45.12G/A	Ti	0.02
-70.95G/A	Ti	0.31	-50.75-/+	ID	0.15	-44.90C/G	Tv	0.08
-70.69+/-	ID	0.26	-50.74+/-	ID	0.17	-44.50A/T	Tv	0.21
-70.42G/A	CTi	0.07	-50.73C/T	Ti	0.17	-44.21G/A	Ti	0.32
-70.37A/G	Ti	0.31	-50.72aG/A	Ti	0.15	-44.16C/A	Tv	0.44
-70.25T/C	CTi	0.46	-50.72bC/T	Ti	0.14	-43.56A/G	Ti	0.29
-70.23T/C	CTi	0.31	-50.69C/T	CTi	0.15	-43.25T/C	Ti	0.29
-70.21T/C	Ti	0.31	-50.66+/-	ID	0.15	-43.07T/C	Ti	0.38
-70.18G/A	CTi	0.05	-50.63G/T	Tv	0.32	-42.89G/A	Ti	0.43
-70.11C/T	Ti	0.07	-50.60aG/C	Tv	0.14	-42.05C/T	CTi	0.14
-69.98C/T	Ti	0.05	-50.60bC/T	Ti	0.15	-41.27G/A	Ti	0.30
-69.84T/C	Ti	0.37	-50.53+/-	ID	0.15	-40.61G/A	Ti	0.30
-69.79A/G	Ti	0.05	-50.48T/C	Ti	0.23	-40.37G/A	Ti	0.34
-69.56G/T	Tv	0.2	-50.34G/T	Tv	0.21	-40.32A/C	Tv	0.30
-69.54A/G	Ti	0.47	-50.32-/+	ID	0.14	-30.23A/G	Ti	0.22
-69.50G/A	Ti	0.47	-50.27A/G	Ti	0.14	-28.13A/G	Ti	0.20
-69.43A/T	Tv	0.42	-50.24G/A	CTi	0.14			
-69.27C/T	Ti	0.47	-50.19G/C	Tv	0.15			

Table 3.3. (continued)

RING 3-TAP 2 interval			SNP			SNP			SNP		
SNP	type	MAF	SNP	type	MAF	SNP	type	MAF	SNP	type	MAF
H6C/A	Tv	0.10	J4A/G	Ti	0.40	Jjk4G/A	Ti	0.12			
H8C/T	CTi	0.07	J13C/T	CTi	0.17	Jjk18G/C	Tv	0.41			
H9C/T	CTi	0.10	J14G/C	CTi	0.07	Jjk14G/A	Ti	0.41			
H10C/T	CTi	0.07	Jjk1C/T	CTi	0.05	Jjk10C/T	CTi	0.17			
H11C/A	Tv	0.07	Jjk2C/T	CTi	0.17	Jjk11+/-	ID	0.09			
H12G/A	Ti	0.37	Jjk13G/A	Ti	0.17	Jjk12+/-	ID	0.04			
H13C/G	Tv	0.03	Jjk15T/C	CTi	0.20	Jjk6G/A	Ti	0.03			
H22A/G	Ti	0.03	Jjk16G/C	Tv	0.20	Jjk7G/C	Tv	0.01			
H23A/G	Ti	0.03	Jjk3G/A	CTi	0.19	Jjk19T/C	Ti	0.02			
Jjh12G/C	Tv	0.08	Jjk5T/C	CTi	0.37	Jjk8C/G	Tv	0.33			
JJ1G/A	CTi	0.01	Jjk9C/T	CTi	0.43	J5C/T	CTi	0.33			
JJ3T/A	Tv	0.45	Jjk4G/A	Ti	0.12	J6G/A	CTi	0.49			
JJ4T/C	CTi	0.03	Jjk4G/A	Ti	0.12	J7C/T	Ti	0.16			
JJ5C/A	Tv	0.04	Jjk18G/C	Tv	0.41	JK4T/C	CTi	0.49			
JJ6T/C	CTi	0.49	Jjk14G/A	Ti	0.41	Kk3G/C	Tv	0.01			
JJ7A/C	Tv	0.39	Jjk10C/T	CTi	0.17	Kk4A/T	Tv	0.11			
J8A/G	Ti	0.14	Jjk11+/-	ID	0.09	Kk5G/C	Tv	0.48			
J1A/T	Tv	0.07	Jjk12+/-	ID	0.04	K5A/G	CTi	0.01			
J2A/G	Ti	0.22	Jjk6G/A	Ti	0.03	K7C/T	CTi	0.28			
J3A/G	Ti	0.42	Jjk7G/C	Tv	0.01	K8G/A	CTi	0.41			

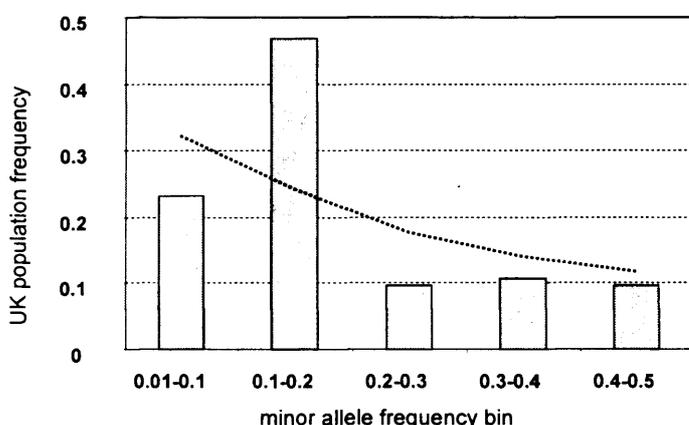


Figure 3.3. SNP allele frequencies in 50 UK North Europeans. Minor allele frequencies sorted by bins of 10% show bias of intermediate frequency (0.1 - 0.2) alleles. Dashed line indicates SNP minor allele-frequency distribution on chromosome 21 (from Patil *et al.*, 2001, see discussion).

Only fourteen SNPs (<6% of all genotyped SNPs) were in coding sequence, nine of these in the highly polymorphic second exon of the *DPB1* gene. Ten of the coding sequence SNPs are at non-synonymous sites: all nine SNPs located in exon 2 of the *DPB1* gene, and one in exon 3 of the *DMB* gene.

Genotype frequencies at 98% of SNPs sites were in Hardy-Weinberg equilibrium. Only five SNP sites deviated from Hardy-Weinberg equilibrium: -72.88, -72.83, -46.22, -44.16 and K8. This deviation was caused by an excess of homozygotes for SNPs -72.88, -72.83, -46.22 and K8, and an excess of heterozygotes for SNP -44.16. Note that SNP sites at -72.88 and -72.83 are only 57 bp apart and in absolute association, and are therefore not independent. The χ^2 values for the deviations varied between 4.0 and 4.9 for SNPs -72.88, -72.83, -44.16 and K8 (P values between 0.05 and 0.025). SNP -46.22 had a higher χ^2 value of 6.7 (P \approx 0.01). However, considering the total number of SNPs tested

(250), it is expected that some SNPs show marginally significant deviation from Hardy-Weinberg equilibrium by chance.

3.2.2. Nucleotide diversity

Nucleotide diversity was calculated only for fully re-sequenced DNA segments where sequence reads were available from six or more individuals. Amplicons designed around clusters of database SNPs were not included in the calculations, because the actual SNP incidence in these segments was not verified through re-sequencing and could therefore be higher.

3.2.2.1. SNP incidence

When SNP incidence (SNP density) is considered, the re-sequenced segments can be divided into two groups: those with higher than human genome average SNP incidence (which is 1 in 300 bp, see introduction) and those with lower than human genome average SNP incidence (Figure 3.4). The former group contains all re-sequenced DNA segments in the *COL11A2-DOA* interval (roughly between positions -74000 to -44000), and the latter group contains all re-sequenced domains in the *RING3-TAP2* interval (roughly between positions 53000 to 106000).

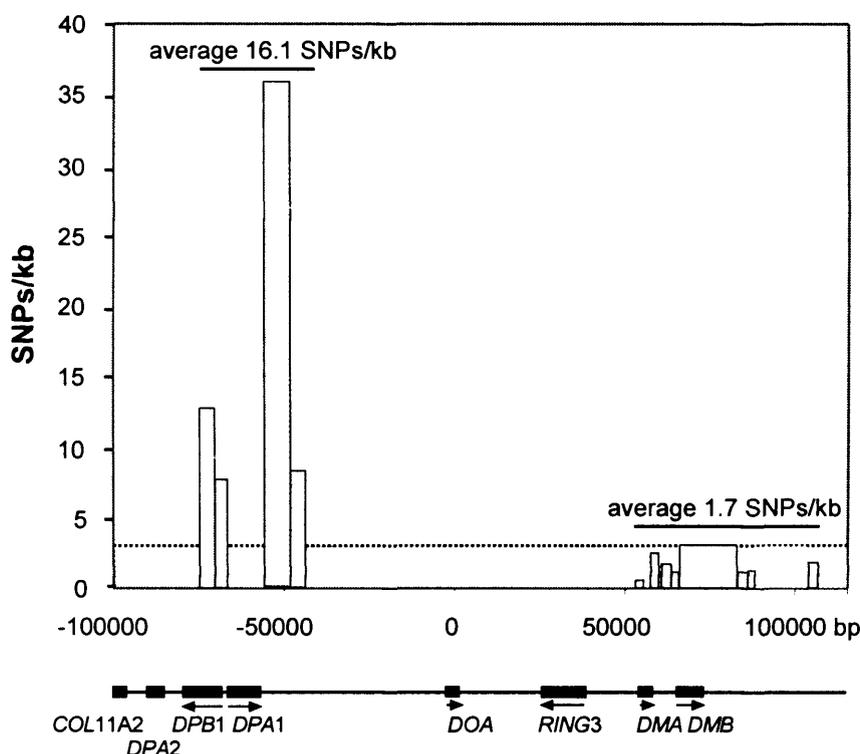


Figure 3.4. SNP density across re-sequenced segments in the two target regions (all known SNPs, even those that were not genotyped, included). Average SNP density in the human genome is shown as a dashed line. Segments that were not re-sequenced, and where SNP density therefore is unknown, are indicated as light grey hatched regions.

When re-sequenced, one of the segments, between primers D-56.0F and D-48.6R, roughly at positions -56000 and -48600, was found to be extremely SNP rich (see Figure 3.5). In addition to the 64 SNPs for which were genotyped (5' to 3': SNPs -55.89T/C to -49.23A/G, Table 3.1), another 93 SNPs were identified in this segment through re-sequencing.

The information content of these 93 SNPs, based on genotypes in the 7 re-sequenced donors, was likely to be identical to other SNPs in this DNA segment. For example, an untyped C/T SNP, at position -53834, had exactly the same genotypes for the seven re-sequenced donors as SNP -53.94C/T 109 bp further downstream (genotypes of donors 2 to 8, for both SNPs: H H C C C C T, see also Table 3.1). Therefore, to minimise time and effort, such SNPs were not genotyped. All known SNPs (genotyped or not) were, however, considered in the SNP density and Watterson's theta calculations (see below).

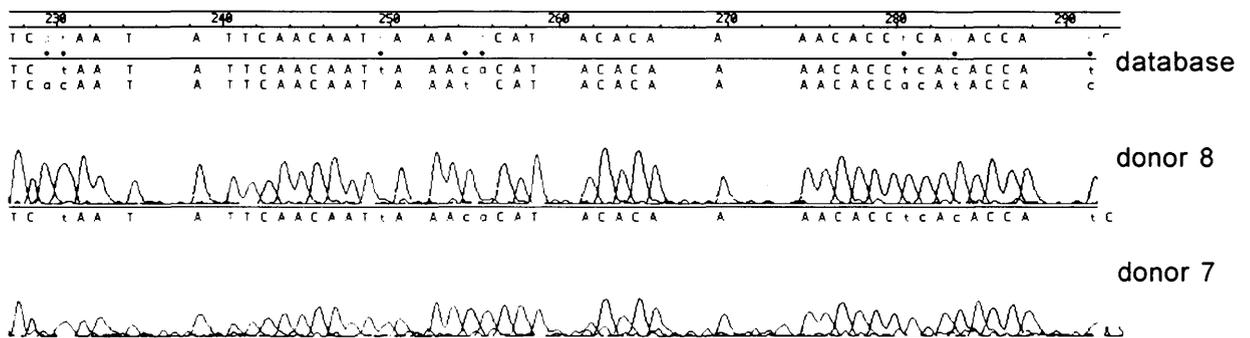


Figure 3.5. An example of the extremely high SNP density observed in the segment between positions -56000 and -48600 (see text). In the 65 bp long DNA segment shown here, 8 SNPs (indicated with dots) are found.

3.2.2.2. Nucleotide heterozygosity π in re-sequenced regions

For the re-sequenced regions (domains I to X in Table 3.4), nucleotide heterozygosity π was calculated, based on the genotypes of 50 UK North European semen donors.

π values range from 0.005% in domain V to 0.248% in domain III. GC content in the re-sequenced segments varied from 33 to 51%. High π values were seen in domains of both relatively high and low GC content (for example, domains I and III). Note that all π values will be a slight underestimate, as they are based on a relatively small number (6-8) of re-sequenced individuals.

When re-sequenced, one of the segments, between primers D-56.0F and D-48.6R, roughly at positions -56000 and -48600, was found to be extremely SNP rich (see Figure 3.5). In addition to the 64 SNPs for which were genotyped (5' to 3': SNPs -55.89T/C to -49.23A/G, Table 3.1), another 93 SNPs were identified in this segment through re-sequencing.

The information content of these 93 SNPs, based on genotypes in the 7 re-sequenced donors, was likely to be identical to other SNPs in this DNA segment. For example, an untyped C/T SNP, at position -53834, had exactly the same genotypes for the seven re-sequenced donors as SNP -53.94C/T 109 bp further downstream (genotypes of donors 2 to 8, for both SNPs: H H C C C C T, see also Table 3.1). Therefore, to minimise time and effort, such SNPs were not genotyped. All known SNPs (genotyped or not) were, however, considered in the SNP density and Watterson's theta calculations (see below).

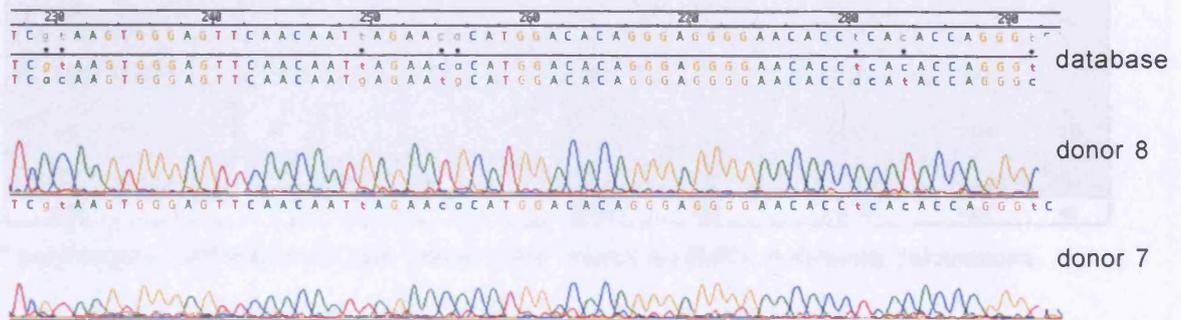


Figure 3.5. An example of the extremely high SNP density observed in the segment between positions -56000 and -48600 (see text). In the 65 bp long DNA segment shown here, 8 SNPs (indicated with dots) are found.

3.2.2.2. Nucleotide heterozygosity π in re-sequenced regions

For the re-sequenced regions (domains I to X in Table 3.4), nucleotide heterozygosity π was calculated, based on the genotypes of 50 UK North European semen donors.

π values range from 0.005% in domain V to 0.248% in domain III. GC content in the re-sequenced segments varied from 33 to 51%. High π values were seen in domains of both relatively high and low GC content (for example, domains I and III). Note that all π values will be a slight underestimate, as they are based on a relatively small number (6-8) of re-sequenced individuals.

Table 3.4. Nucleotide heterozygosity π in each of the re-sequenced DNA segments (I-X). Domains shaded in gray have lower than human genome average SNP incidence (1 SNP in 300 bp) and π values (0.075%, Bamshad and Wooding, 2003). Each domain is defined forward and reverse primers, shown in the first column.

Domain		donors		length(bp)	SNPs	SNP incidence	π (%)	%GC
		re-sequenced/	genotyped					
I	D-74.2F to D-70.4R	7	50	3730	47	1 in 79	0.125	51
II	D-70.7F to D-65.9R	7	50	4624	28	1 in 165	0.091	43
III	D-56.0F to D-48.6R*	7	50	4296	72	1 in 60	0.248	36
IV	D-48.6F to D-44.3R	8	49	3871	28	1 in 138	0.121	42
V	R53.4F to R55.6R	6	50	1779	1	1 in 1779	0.005	45
VI	R56.9F to R59.4R	6	50	1810	4	1 in 453	0.016	49
VII	R60.7F to R63.2R	6	50	2392	4	1 in 598	0.013	40
VIII	R65.4F to R83.4R	6	50	12145	36	1 in 337	0.043	43
IX	R85.2F to R87.5R	6	50	2206	3	1 in 735	0.016	33
X	R104.1F to R106.5R	6	50	1735	3	1 in 578	0.026	39
total				38588	227	1 in 169	0.077	42

* polymorphic SVA elements (see below) were treated as SNPs in diversity calculations.

3.2.2.3. Watterson's theta (non-recombining DNA segments only)

Watterson's theta is a measure of diversity for non-recombining regions. Therefore, theta was calculated on a subset of the DNA domains used for π calculations. While for π calculations, the data set consisted of 50-donor-genotypes of SNPs in re-sequenced segments (domains I-X), for theta, firstly, only non-recombining DNA segments (domains a to j in Table 3.5) were included. Information on which segments are recombining and which are non-recombining was gained from subsequent LD and crossover analyses described in chapters 4 - 7. For this reason, domains V-VII and IX-X in Table 3.4 are identical to domains d-f and i-j in Table 3.5. Secondly, the number of chromosomes for calculating theta was the number of *re-sequenced* chromosomes, rather than the set of 50-donor-genotypes. As mentioned above, the region containing domain b was found to be extremely SNP rich (see Figure 3.5). In addition to the 64 genotyped SNPs, the 93 SNPs identified in this segment through re-sequencing were also included in theta calculations. Similarly, 10 SNPs contained in segment c, identified through re-sequencing, were not genotyped, but were included in Watterson's theta calculations.

Table 3.5. The nucleotide diversity measure Watterson's theta in each non-recombining DNA segment. Outermost primers or SNP sites that define each domain are shown. Domains shaded in grey have lower theta values than human average (average across 22 gene regions is 0.083%, Nachman, 2001). For DNA segments, where genotypes for all SNPs were available, nucleotide diversity value π was calculated for the same data set of 12 chromosomes.

Domain	chromosomes	length(bp)	SNPs	theta (%)	π (%)
a -70.23 to -66.07	14	3414	27	0.243	0.133
b D-56.0F to -49.23	14	4385	157	1.1	-
c -47.97 to -44.50	16	3710	31	0.247	-
d R53.4F to R55.6R	12	1779	1	0.018	0.005
e R56.9F to R59.4R	12	1810	4	0.071	0.021
f R60.7F to R63.2R	12	2392	4	0.053	0.01
g R65.4F to JJ7	12	4485	6	0.052	0.011
h J5 to JK4	12	3161	4	0.041	0.021
i R85.2F to R87.5R	12	2206	3	0.043	0.019
j R104.1F to R106.5R	12	1735	3	0.056	0.031
total	13	29077	240	0.259	-

3.2.2.4. Two large polymorphic insertion/deletions downstream the *DPA1* gene Domain b in Table 3.5, which is defined 5' by forward primer D-56.0F and 3' by SNP -49.23A/G, contains two large polymorphic insertion/deletions. They were discovered to be length polymorphic accidentally by PCR (Figure 3.6), and located by sequencing to positions -54646 and -52947, approximately 2.5 kb and 4.2 kb downstream the *DPA1* gene, respectively (see Tables 3.1 and 3.3).

The insertion located further 3', which is more common in our panel of 50 donors, is contained in the MHC database sequence, while the one located further 5' is not. Sequencing of the insertion junctions of the 5' insertion revealed that both elements involve a similar insertion/deletion of a ~1.7 kb long SINE variant Alu (SVA) element; they were named SVA1 and SVA2 (Figure 3.6).

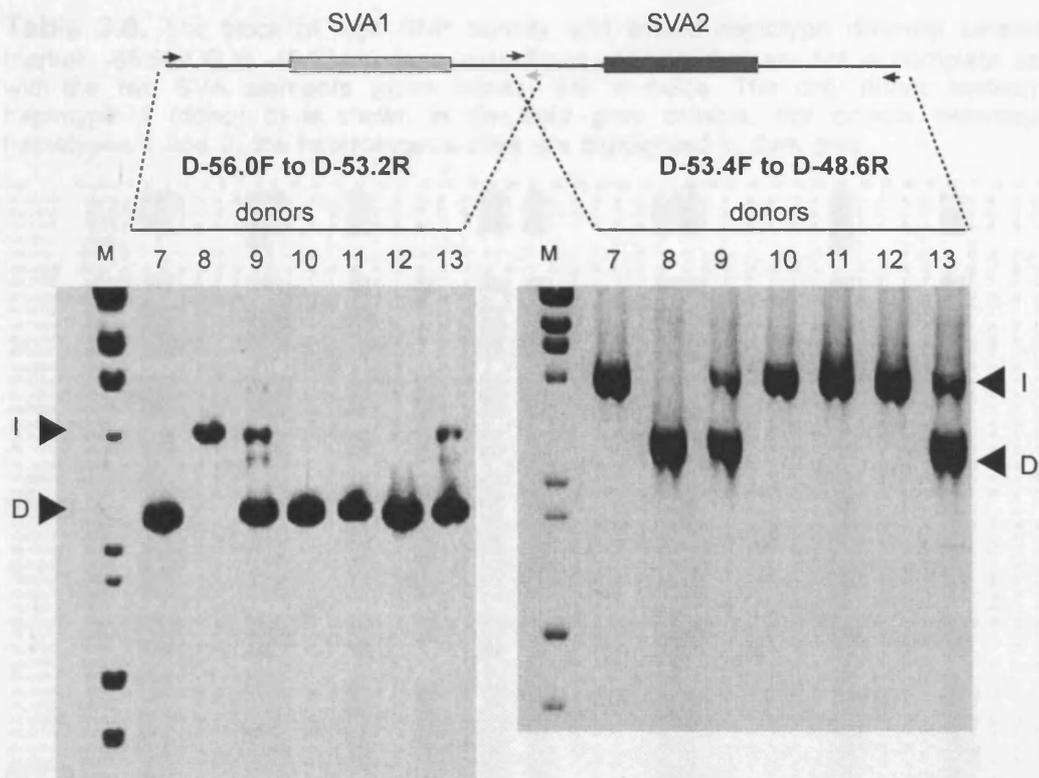


Figure 3.6. PCR amplification products across the two SVA elements in donors 7 - 13. Using the combinations of forward and reverse primers (arrows) shown, the presence or absence of the two SVA elements (light and dark grey boxes) can be scored as a large length polymorphism (black arrowheads, I, insertion, D, deletion). M, marker (*Hind*III digested λ DNA plus *Hae*III digested ϕ x DNA).

Across domain b, there is very limited haplotype diversity as is clear simply by visual inspection of genotypes across this region (Table 3.6). Two major haplotypes can be identified. The polymorphic SVA elements can be used to define haplotypes in segment b as follows: presence of SVA1 = haplotype 1 (0.15 frequency), presence of SVA2 = haplotype 2 (0.85 frequency). Thirty-one of the genotyped SNPs in segment b are in absolute association with the SVA elements, *i.e.* contain precisely the same genotype information. The SVA elements will be discussed in more detail in chapter 5.

These two haplotypes in fact extend beyond the 5' end of segment b, with marker -68.68A/G at position -68674 still being in complete association (see Table 3.6). The last 3' marker in complete association with the SVA elements is -49.23A/G at position -49232, thus making the haplotype block ~19.5 kb in length. Sixty-five of the genotyped SNPs are in complete association with the SVA elements. Additionally, there are 13 genotyped SNPs within this haplotype block that are not in association with the SVA elements (Table 3.6).

identification through re-sequencing. Regions of particular interest would still have to be re-sequenced, but these efforts can be precisely targeted.

The regions examined were found to be highly (*COL11A2-DOA* interval) or moderately (*RING3-TAP2* interval) polymorphic. This is important for subsequent recombination analysis, because LD patterns can only be examined at high-resolution if there is good marker density, and crossover analysis relies on a sufficient number of heterozygous SNP sites in several sperm donors.

The apparent abundance of SNPs with intermediate allele frequencies reflects the SNP discovery strategy used. When only six to eight donors are re-sequenced, low minor allele frequency SNPs - which are in fact more abundant - are likely to go unnoticed, while SNPs with intermediate allele frequency are enriched for. In another study, where database SNPs on chromosome 19 were genotyped, a more uniform allele-frequency distribution was found (Phillips *et al.*, 2003), reflecting the bias towards common SNPs in public databases. Complete SNP ascertainment would result in a very different allele frequency distribution, with most SNPs having a low (<10%) allele frequency. This was shown by Patil *et al.* (2001), who used high-density oligonucleotide arrays to re-sequence the unique sequence portion of chromosome 21 (67% of total chromosomal sequence) of a total of 20 chromosomes, from African, Asian and North European individuals. They identified ~36,000 SNPs, of which a significant proportion (32%) were observed only once (see dashed line in Figure 3.3).

The frequency of different SNP types in the MHC presented here (8% insertion/deletions, 64% A/G, 10% A/C and 18% C/G) agrees well with previous genome-wide (66% A/G, 9% A/C, 16.5% C/G, Zhao and Boerwinkle, 2002) and X chromosomal estimates (8% insertion/deletions, 63% A/G, 17% A/C, 8% C/G, Miller *et al.*, 2001). Thus, the distribution of SNPs into the different mutational types probably reflects the overall mutational dynamics of the human genome. For example, the high prevalence of A/G (= G/A, C/T or T/C) SNPs is the result of the higher mutability of CpG sites, as discussed above. The input rates of different SNP types into the population vary, while genetic drift, influencing the subsequent population frequency, is independent of SNP type. On a very localised level, there seems to be no correlation between nucleotide diversity and GC content.

The difference in SNP incidence between the first and second target regions (Figure 3.4) suggests that large (tens of kilobases long) DNA segments behave in a co-ordinated manner with regard to nucleotide variability - a long segment with high SNP density is followed by a long segment of low SNP density. A similar but more extreme pattern was observed by Horton *et al.* (1998) in the *DQB* region of the MHC class II: a 28 kb long region immediately upstream, and including, the *DQB1* gene had >200 times higher SNP density than the neighbouring 38 kb downstream segment. In our study, the two extremes in SNP diversity observed in this study are DNA domains b (36 SNPs/1000 bp) and d (0.6

SNPs/1000 bp) in Table 3.5, giving a 60-fold difference. These differences are not as pronounced when SNP diversities in the re-sequenced segments in the first (*COL11A2-DOA*) and second (*RING3-TAP2*) target interval are averaged (see Figure 3.4), in which case the difference in SNP density is ~9-fold. The SNP density, as well as π and theta values in the region from -56000 to -49000 are extraordinarily high (elevated approximately 11-fold, 3-fold and 13-fold above human genome average, respectively).

High SNP incidence within the first target region (*COL11A2-DOA* interval) does not necessarily translate into high nucleotide diversity - for example, domains I and III in Table 3.3 have similar SNP densities but nucleotide diversity π in domain III is almost twice as high. The low π value in domain I is due to the low minor allele frequencies of SNPs within this domain. In the second target region (*RING3-TAP2* interval), where SNP incidence is lower than genome average, π values too are below genome average.

For most data on nucleotide variability in the human genome, π and theta values for a given gene region are similar (Nachman, 2001). As seen in Tables 3.4 and 3.5, and Figure 3.7, the SNP data presented here show that π and theta do correlate - regions with high π values also have high theta values (Figure 3.7). However, there is a clear discrepancy between the absolute values of π and theta - theta is consistently higher for each of the segments examined, even when data sets are identical (see last two columns in Table 3.5). In other words, in the MHC DNA segments examined, a relatively high SNP density but a low heterozygosity is observed.

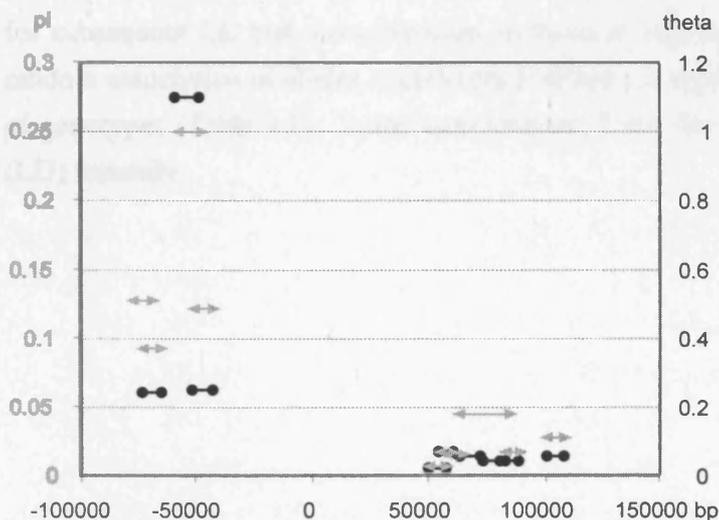


Figure 3.7. Comparison of nucleotide diversity measures π (π , scale on left, gray double arrows) and Watterson's theta (scale on right, black lines with closed circles) vs. distance (data from tables 3.3 and 3.5). Note the different scales on the two measures.

Theta values were also consistently ~1.6 times higher than π values for 292 autosomal gene regions and 21 X-linked genes (Stephens *et al.*, 2001a). The authors interpreted this as evidence for a recent population expansion. An excess of low-frequency

variants can also arise from positive selection (Bamshad and Wooding, 2003). The high SNP density and diversity observed in the MHC in the *COL11A2-RING3* interval suggests that the region is relatively old. Nucleotide heterozygosity at population equilibrium should then be similarly high. Instead, we observe a consistently lower heterozygosity across all segments examined. Several factors could account for the observed SNP behaviour. Firstly, genetic drift can alter allele frequencies in a stochastic manner. Similarly, a population bottleneck or rapid population expansion (which can be viewed as accelerated drift) can result in certain alleles increasing in frequency, which is how Stephens *et al.* (2001a) explain their data. Lastly, both positive and negative selection can shift allele frequencies by favouring or eliminating certain alleles, together with alleles at linked loci.

The lower heterozygosity (in comparison to SNP diversity) in the MHC data set would suggest that selection, if present, favours specific alleles, rather than heterozygotes. The π vs. theta discrepancy is particularly extreme in the DNA segment around the SVA elements (Table 3.5 and Figure 3.7). The only gene located in the SNP-rich haplotype block defined by the SVA elements is *DPA1*. A selective sweep involving this region that would have significantly reduced the number of haplotypes such that the two diverged haplotypes remained in the population, could explain the observed pattern. On the other hand, genetic drift, particularly when population size is small, could also generate the observed SNP pattern. Thus, both selection (which can easily be envisaged acting on the MHC) and drift remain possible explanations for the curious SNP diversity patterns observed.

I have shown that SNP density in the MHC class II is variable and sufficiently high for subsequent LD and recombination analyses at high resolution. In some cases, the non-random association of alleles at different SNP sites is apparent simply by visual examination of genotypes (Table 3.6). In the next chapter, I test the extent of non-random association (LD) formally.

4. LINKAGE DISEQUILIBRIUM ANALYSES

4.1. INTRODUCTION

LD analysis can be used to infer the historical recombination activity of a DNA segment. Although this approach is complicated by population genetic factors (see Hedrick, 1988, and section 1.7.4), it is a useful first step in homing in on recombinationally active regions. LD between any two SNPs should corrode over time, if recombination is occurring in the DNA region separating them.

There is considerable interest in understanding LD in the human genome, mainly to aid localisation of disease loci by disequilibrium mapping. Numerous studies have been conducted to characterise the range of LD over large distances (for example, Laan and Pääbo, 1997, Huttley *et al.*, 1999, Zavattari *et al.*, 2000, Dunning *et al.*, 2000, Taillon-Miller *et al.*, 2000, Service *et al.*, 2001, Abecasis *et al.*, 2001). Earlier LD studies used microsatellites as markers (Peterson *et al.*, 1995, Laan and Pääbo, 1997, Huttley *et al.*, 1999), but more recently, the focus has shifted to SNPs because of their simpler mutational dynamics and greater prevalence in the human genome (Taillon-Miller *et al.*, 2000, Dunning *et al.*, 2000, Abecasis *et al.*, 2001, Stephens *et al.*, 2001a, Reich *et al.*, 2001). LD between microsatellites has been reported to extend over very long distances, from over 1 Mb (Peterson *et al.*, 1995) up to 14 Mb (Laan and Pääbo, 1997) and can vary between populations, as shown by studies of microsatellites in Xq13 (Laan and Pääbo, 1997). Such long-range associations are not seen between pairs of SNP markers, with some rare exceptions (Taillon-Miller *et al.*, 2000). Pritchard and Przeworski (2001) recently termed the two classes of LD as "long-distance LD" and "short-scale LD". The differences probably reflect the different mutational dynamics of the two types of marker – microsatellites are multiallelic, mutate frequently and thus carry young alleles, whereas SNPs are biallelic with relatively ancient alleles. Ancient haplotypes, as defined by SNPs, are more likely to be disrupted by recombination.

For SNPs, an early simulation study assuming random crossover distribution predicted that "useful levels of LD [for the purpose of complex disease gene mapping] are unlikely to extend beyond an average distance of approximately 3 kb in the general population" (Kruglyak, 1999), but this has since proven inconsistent with empirical observations. It has been shown that there is variation both in the extent of LD between genomic regions, and between populations. Dunning *et al.* (2000) found that SNP marker pairs up to 20 kb apart could show significant LD and that levels of LD were similar across four populations of European ancestry. Abecasis *et al.* (2001) examined LD in three genomic regions in UK North Europeans and found that half of the SNP markers were in "useful" LD at a distance up to 50 kb, with occasional associations extending as far as 500 kb. Reich *et al.* (2001) convincingly demonstrated variation in the extent of LD between genomic

regions. They examined LD between high frequency SNPs (minor allele frequency ≥ 0.35) in 19 randomly chosen genomic regions; for each region, one SNP was used a reference point ("core SNP"), and LD was calculated between this SNP and SNPs located 5, 10, 20, 40, 80 and 160 kb away. In one region (the *WASL* gene), no LD decay was observed across 160 kb, whereas in another (the *PCI* gene) LD values dropped by half already at less than 10 kb distance. These data were for US individuals of North European descent in Utah - the authors also compared the extent of LD in two other populations, Swedes and Nigerians. LD decayed at a similar rate in the Utah and Swedish sample, but much more rapidly in Nigerians. As a general rule, LD extends less far in African than non-African populations; this will be discussed in more detail in chapter 9.

4.1.1. Measures of LD

Various LD measures exist for characterising the statistical association between alleles at different loci (see e.g. Hedrick, 1987). A simple measure of LD is

$$D_{ij} = x_{ij} - p_i q_j,$$

where x_{ij} is the observed frequency of gamete or haplotype $A_i B_j$, p_i and q_j are the frequencies of alleles A_i and B_j at loci A and B, respectively, and the expected frequency of gamete $A_i B_j$ is $p_i q_j$ assuming no statistical association between the alleles. The range of this measure is a function of the allele frequencies, and is therefore not ideal.

The normalised measure D' (Lewontin, 1964) is perhaps the most widely used, as its range is the same for all allele frequencies. It does tend to be over-inflated when sample size is small, however (Weiss and Clark, 2002). D' is defined as follows:

$$D'_{ij} = D_{ij}/D_{\max}$$

where $D_{\max} = \min[p_i q_j, (1-p_i)(1-q_j)]$ when $D_{ij} < 0$, or $D_{\max} = \min[p_i(1-q_j), (1-p_i)q_j]$ when $D_{ij} > 0$.

D' can range from -1 to 1; whether the sign is positive or negative depends on the arbitrary labelling of alleles, and hence the absolute value $|D'|$ is mostly used. If $|D'|=1$, this is referred to as complete LD, with at most three haplotypes present in the population, and means there has been no recombination during the history of the sample. Three haplotypes can be explained by mutation, without having to invoke recombination. $|D'|=0$ equals no LD, or free association.

Δ is a measure of absolute association (Hill and Robertson, 1968), and is dependent on allele frequencies; it only shows values of 1 when just two haplotypes per marker pair are observed, in which case both markers must have the same allele frequencies.

$$\Delta = D_{ij}/(p_i p_j q_i q_j)^{1/2}$$

A value of $\Delta=1$ is called perfect or absolute LD. Δ also has a formal relationship to effective population size N_e , recombination rate r per unit distance and inter-marker distance d , where $\Delta^2=1/(1+4N_e rd)$ for a population at recombination/drift equilibrium for selectively-neutral haplotypes (Sved, 1971).

Usually, Δ is lower than D' over a given genomic distance (Weiss and Clark, 2002). Statistically significant high D' (values near 1) is useful for reporting on lack of historical recombination, but intermediate D' values should be interpreted with caution (Ardlie *et al.*, 2002).

4.1.1.1. Statistical analysis

D' and Δ measures of LD and odds of linkage equilibrium, as statistical confirmation of the results, can be calculated from unphased diploid genotype data. The analysis program used (software written by A.J. Jeffreys) first calculates the observed allele frequencies at each locus pair and varies one haplotype frequency (x_{11}) from 0 to the lesser value of p_1 or q_1 . From this, the corresponding values for each of the remaining 3 possible haplotypes are defined.

The probabilities (P) of getting the observed genotype data at the two loci are then computed; the value of x_{11} that gives the maximum probability (P_{\max}) of getting the observed genotype data is determined, to produce maximum-likelihood values of all 4 possible haplotypes x_{11} , x_{12} , x_{21} , x_{22} . From these values, D' and Δ are computed. P is then re-computed for haplotype frequencies at linkage equilibrium, *i.e.* $x_{11} = p_1 q_1$ *etc.*, to give a value of P (P_{eq}) at equilibrium. An odds ratio P_{\max}/P_{eq} is calculated and forms the statistical test of significance of association, showing how many times more likely the observed genotype data are, if the SNP pair is in linkage disequilibrium as opposed to free association.

4.1.1.2. Graphical output to illustrate the extent of LD

For an output that allows visual evaluation of the extent of LD, LD plots are generated. During the course of this work, $|D'|$ rather than $|\Delta|$ measures of LD were used for the graphical output, because D' is more useful for the identification of obligate recombinant haplotypes. The LD plots show pairwise LD values, plus odds ratios, between every SNP with every other SNP in the target region as coloured areas. An example is shown in Figure 4.1.

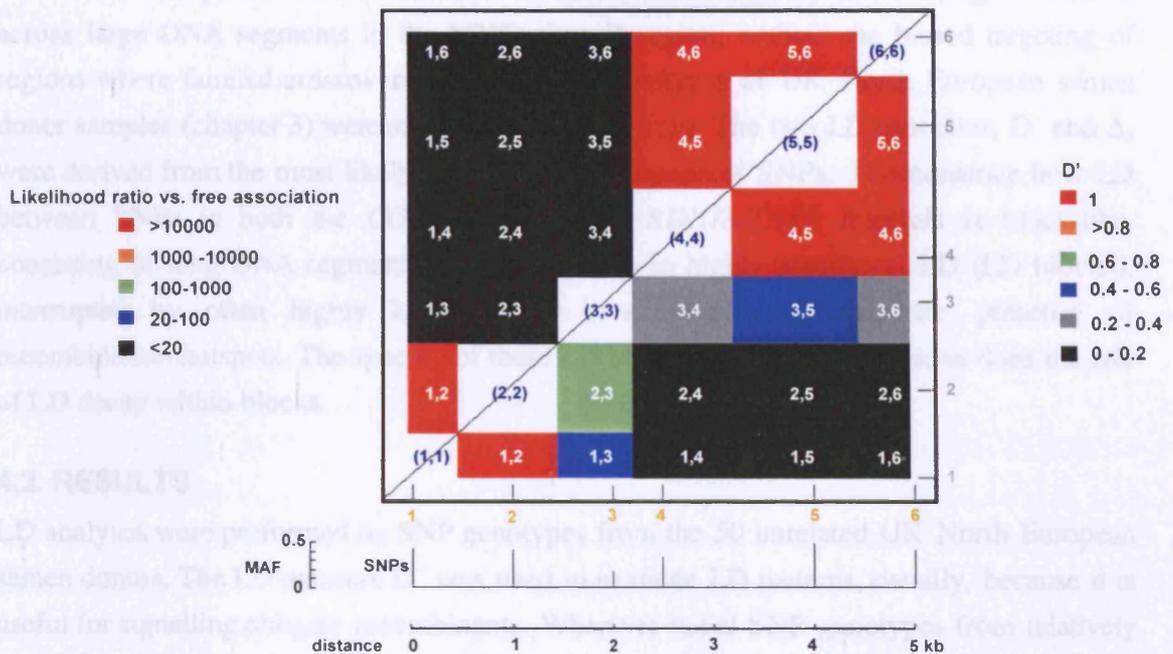


Figure 4.1. Graphical output of $|D'|$ values and their statistical significance. In this example, there are six SNPs, labelled 1-6, shown in yellow below the plot, and in grey to the right of the plot. Also shown below the plot are the minor allele frequencies (MAF) for each SNP (see scale on left). Below the diagonal are $|D'|$ measures of LD between each pair of SNPs. Each point is plotted as a rectangle centered on each SNP, and extending half way to adjacent markers. The white text on each rectangle shows which markers are being compared. White squares (labelled with blue text) are where markers are plotted against themselves, thus having no D' value. A red rectangle is indicative of complete LD and no more than three haplotypes; the other coloured rectangles represent $|D'|$ values as indicated on the right, and imply more than 3 haplotypes, and therefore possible recombinational activity. The statistical significance of each $|D'|$ value can be deduced from the likelihood ratio vs. free association, shown above the diagonal. For example, D' values between markers 3 and 5 can be found by following marker 5 on the x axis (yellow) and marker 3 on the y axis (grey). This blue rectangle indicates that the D' value between these two sites is 0.4-0.6. To test the significance of this observation, marker 3 on the x axis (yellow) can be followed vertically up to marker 5 on the y axis (grey). The black square represents poor odds of linkage disequilibrium (likelihood ratio of LD vs. free association <20). In this case, the intermediate D' value is not supported by significant statistics, and the pair of markers could be in free association. In contrast, between markers 1 and 2, $D'=1$, indicating high levels of LD supported by a very high (>10000) likelihood ratio, indicated by a corresponding red rectangle above the diagonal.

This work

To localise genomic regions that are good candidates for harbouring recombination hotspots, we examine LD patterns for evidence of localised historical recombination, that is, localised areas of LD breakdown. Studies of the previously verified recombination hotspots in the *DOA/RING3* and in the *TAP2* gene regions had involved some very targeted LD analyses immediately around areas of familial crossover clustering. While it had become clear that there was no LD between markers across these two hotspot regions, separated by roughly 160 kb, there was no information on the large-scale patterns of LD across the MHC class II region.

The data presented here constitute the first systematic survey of LD at high-resolution across large DNA segments in the MHC class II region, without the biased targeting of regions where familial crossovers cluster. SNP genotypes of UK North European semen donor samples (chapter 3) were subjected to LD analyses. The two LD estimates, D' and Δ , were derived from the most likely haplotypes for all pairs of SNPs. I demonstrate how LD between SNPs in both the *COL11A2-DOA* and *RING3-TAP2* intervals is block-like, consisting of long DNA segments where SNPs are in highly significant LD (LD blocks), interrupted by often highly localised LD breakdown, suggesting the presence of recombination hotspots. The spacing of these LD breakdown regions varies, as does the rate of LD decay within blocks.

4.2. RESULTS

LD analyses were performed on SNP genotypes from the 50 unrelated UK North European semen donors. The LD measure D' was used to examine LD patterns visually, because it is useful for signalling obligate recombinants. Wherever initial SNP genotypes from relatively widely spaced amplicons suggested LD breakdown, complete re-sequencing of that region was undertaken in six to eight donors. This way, more SNPs in the region of interest were identified, and the densely spaced SNPs allowed further examination and refinement of regions of LD breakdown. A block-like structure of LD was observed, in particular with common (minor allele frequency ≥ 0.15) SNPs. In general, markers with a low minor allele frequency are not very informative for reporting on historical recombination. Firstly, they often show poor statistical power (high D' values, accompanied by low odds ratios). Secondly, they tend to represent relatively recent mutations and have not had time to recombine; therefore, they appear to be in LD with most markers on the haplotype they are travelling on. Such private lineages can give misleadingly high LD values even across recombination hotspots. For this reason, when interpreting LD patterns for subsequent crossover analyses, SNPs with minor allele frequency < 0.15 were excluded (see Figures 4.2Ai-ii and 4.3Ai-ii). Note that the LD blocks would have better-defined boundaries if the minor allele frequency (MAF) filter was set even higher (say, > 0.25). The problem with removing all SNPs with MAF < 0.25 is that a large proportion of markers would be lost, because some DNA segments do not contain high-frequency SNPs.

4.2.1. LD in the *COL11A2-DOA* interval

Within the ~70 kb long *COL11A2-DOA* interval, three LD blocks (blocks 1-3, Figure 4.2Aii) were observed, within which SNPs are in LD. These blocks are not "crisp", in that some marker associations are not always strong within a block, while in other cases, associations run across block boundaries. Filtering out SNPs with a low (< 0.15) minor

allele frequency makes little difference (compare Figure 4.2Ai with 4.2Aii), because in this interval, such SNPs are rare.

Most markers in block 1, which is at least 30 kb long, are in strong association from the first 5' marker to the markers located at around -72000. Some SNPs between approximately positions -74000 to -70000, which have mostly low minor allele frequency, are associated with markers in both block 1 and block 2 (Figure 4.2Ai). Most high D' values are, however, not statistically significant, as expected for low frequency markers, and disappear when only high frequency markers are considered (Figure 4.2Aii). The 3' boundary of block 1 lies at around -70000; markers 3' of this location do not show statistically significant associations with block 1.

Block 2 is ~18 kb long, extending approximately from -68000 to -50000. This block is characterised by a very high number of SNPs, and a limited number of haplotypes (see chapter 3). Most SNPs in block 2 are in complete and highly significant association with each other. There are some markers that "misbehave" within this block - they do not show associations with other markers in block 2, but are instead in strong but not statistically significant LD with markers in neighbouring blocks.

Block 3 starts at around -48000 and extends until the 3' end of the assayed interval. This block again contains two markers that are not in LD with other markers within this block, but instead with markers outside the block. Furthermore, several markers in block 3 show high D' values both with markers in this block as well as block 2. It is mostly the D' values for markers within block 3, however, that have good statistical support.

Thus, within the *COL11A2-DOA* interval, two apparently localised regions of LD breakdown were located. Because these regions, one around the *DPB1* gene and the other less than 10 kb 3' of the *DPA1* gene, are candidates for containing recombination hotspots, their LD patterns were examined in more detail, as described below.

4.2.1.1. LD breakdown in the *DPB1* gene region

The boundary between LD blocks 1 and 2 localises to the first intron of the *DPB1* gene (Figure 4.2B). The precise location of this boundary is not clear - some SNPs at the 3' end of block 1 show strong association with SNPs at the 5' end of block 2, so that blocks 1 and 2 overlap (see shaded grey region in Figure 4.2B). Furthermore, block 2 contains some markers that are in complete LD with a number of SNPs across block 1 (marked with black dots in Figure 4.2B). However, the majority of these high D' values are not statistically significant. In block 1, there is also a cluster of three SNPs (marked with squares in Figure 4.2B) around position -73000 which are in more or less free association with other block 1 markers, but shows peculiar patterns of association with a few markers in block 2.

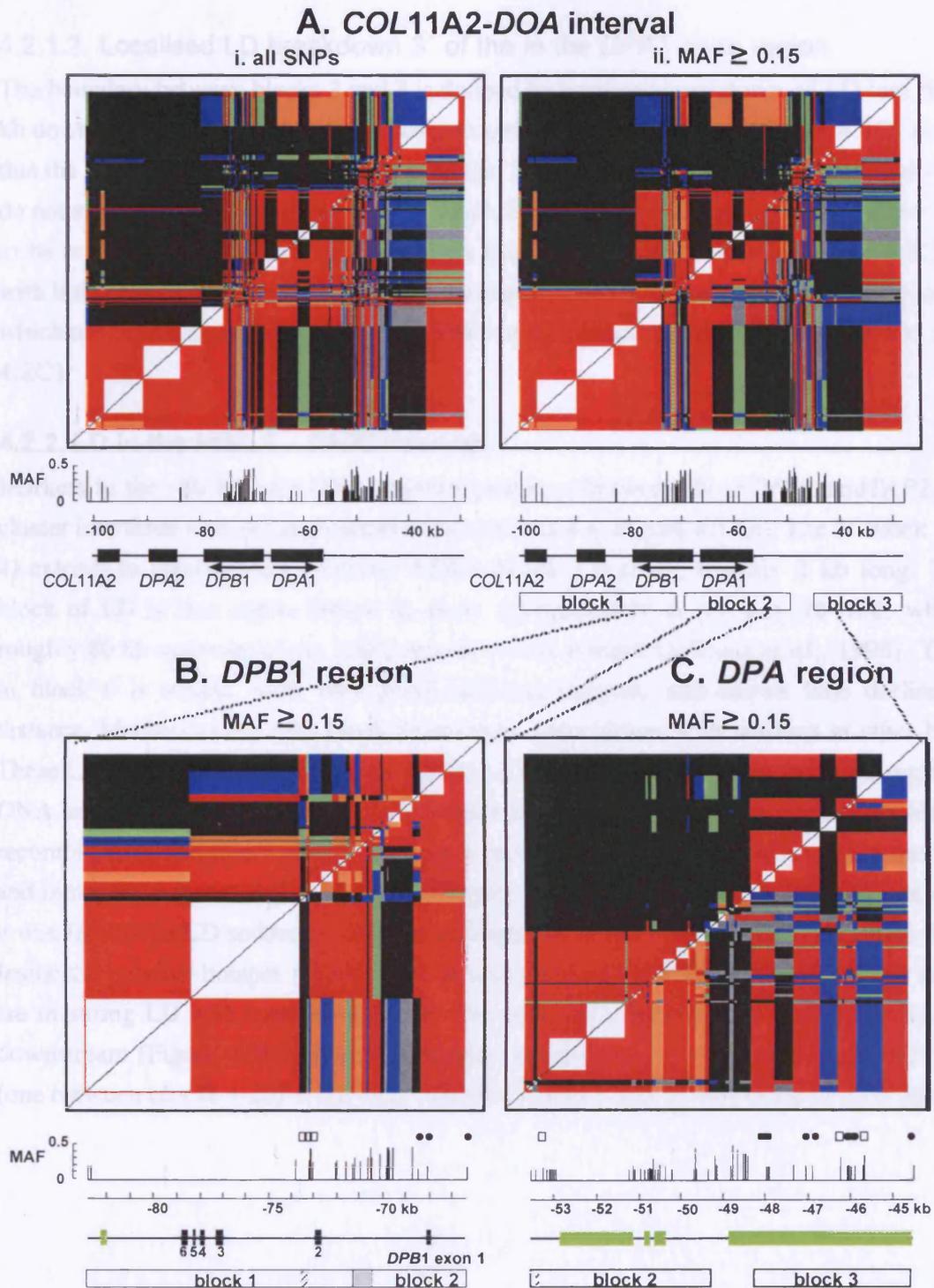


Figure 4.2. LD in the first target region, upstream of the *DOA* recombination hotspots. (A) LD across the ~ 80 kb long *COL11A2-DOA* interval, between all pairs of SNP markers (i) and between SNPs with minor allele frequency (MAF) ≥ 0.15 (ii). LD blocks labelled 1-3 were identified visually as clusters of SNPs in strong association. (B) Region of LD breakdown around the *DPB1* gene. Numbered black boxes are exons. (C) Localised LD breakdown downstream of the *DPA1* gene. In (B) and (C), several markers (marked with black dots) are in complete LD with markers in block 2, but most of these associations are not statistically significant. Some markers (marked with open squares, and small hatched areas inside the LD block) are in free association with other SNPs in the same LD block. Green boxes are repeat elements (LINE, SVA or HERV).

4.2.1.2. Localised LD breakdown 3' of the in the *DPA1* gene region

The boundary between blocks 2 and 3 is defined by localised breakdown of LD less than 10 kb downstream of the *DPA1* gene. Closer examination of this region (Figure 4.2C) revealed that the 3' boundary of block 2 lies at ~49000. SNPs between positions -49000 and -48000 do not associate significantly with either block 2 or 3. Some markers in block 3 were found to be in complete LD with block 2 markers (marked with black dots in Figure 4.2C), but with little or no statistical support. Also, similarly to block 1, three markers were identified which are in free association with other SNPs in their block (marked with squares in Figure 4.2C).

4.2.2. LD in the *RING3* – *TAP2* interval

Markers in the ~60 kb long DNA segment examined between the *RING3* and *TAP2* genes cluster into three well-defined blocks of LD (blocks 4-6, Figure 4.3Aii). The 5' block (block 4) extends to approximately position 72000. Block 5 is small, roughly 2 kb long. The 3' block of LD in this region (block 6) starts approximately at position 76 000, which is roughly 80 kb upstream of the *TAP2* recombination hotspot (Jeffreys *et al.*, 1998). The LD in block 6 is strong, with very good statistical support, and shows little decline with distance. Markers within each block do not show associations with markers in other blocks. These LD patterns indicate that in the *RING3-TAP2* interval, only one localised region (the DNA segment at the 5' end and just downstream the *DMB* gene) has undergone historical recombination, and therefore may harbour a recombination hotspot. When LD at the 5' end and in the downstream region of the *DMB* gene was examined more closely (Figure 4.3B), it was found that LD suddenly collapses downstream of the markers around position 72000. Inside the putative hotspot region, there is a small block of SNPs (block 5) where markers are in strong LD with each other, but in free association with the LD blocks upstream and downstream (Figure 4.3B). This may suggest the existence of two recombination hotspots (one between blocks 4 and 5, the other between blocks 5 and 6) very close to each other.

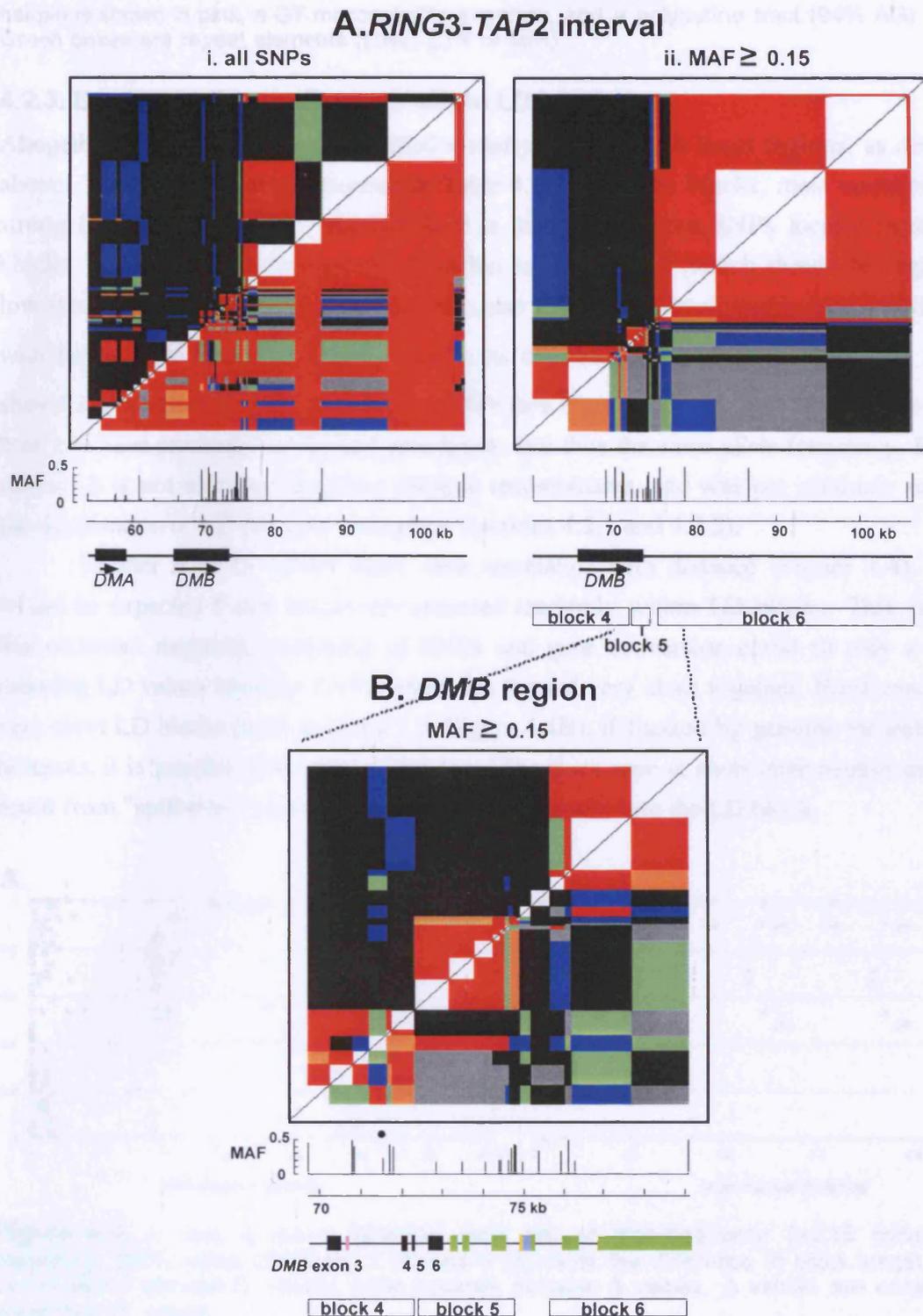


Figure 4.3. LD in the second target region, between the *DOA* and *TAP2* recombination hotspots. (A) LD across the ~50 kb long *RING3-TAP2* interval, between all pairs of SNP markers (i) and between SNPs with minor allele frequency ($MAF \geq 0.15$) (ii). LD blocks labelled 4-6 were identified visually as clusters of SNPs in strong association. (B) Localised LD breakdown around the *DMB* gene. One marker in block 4 (marked with black dot) is in complete LD with several markers in block 5, but these associations are not statistically significant. Numbered black boxes are exons. A ~80 bp long sequence capable of forming a

hairpin is shown in pink, a GT microsatellite in yellow, and a polypurine tract (94% AG) in lilac. Green boxes are repeat elements (LINE, LTR or MIR).

4.2.3. Decay of LD with distance within LD blocks

Altogether six LD blocks were identified visually, three in both target regions, as described above. These blocks are summarised in Table 4.1. Within the blocks, most markers are in strong LD with each other, whereas there is little LD between SNPs located in different blocks. To investigate the decay of LD within each LD block (which should be regions of low recombination activity), Δ values were also calculated for all within-block SNP pairs with minor allele frequency ≥ 0.15 . Due to its dependency on allele frequency, Δ usually shows lower values than D' for a pair of SNPs (see Figure 4.4) - Δ only has a value of 1 if both loci have precisely concordant genotypes, and thus the same allele frequency. For this reason, Δ is not ideal for signalling obligate recombinants, and was not routinely used for the localisation of LD breakdown regions (sections 4.2.1 and 4.2.2).

Neither Δ or D' values show clear correlation with distance (Figure 4.4), which would be expected if rare crossovers occurred randomly within LD blocks. This suggests that recurrent mutation, mistyping of SNPs and gene conversion could all play a role in reducing LD values between SNPs which are located very close together. Furthermore, for very short LD blocks (such as block 5 in Figure 4.4B), if flanked by genuine recombination hotspots, it is possible that some of the low LD values seen at short inter-marker distances result from "spill-over" of crossovers from the hotspot(s) into the LD block.

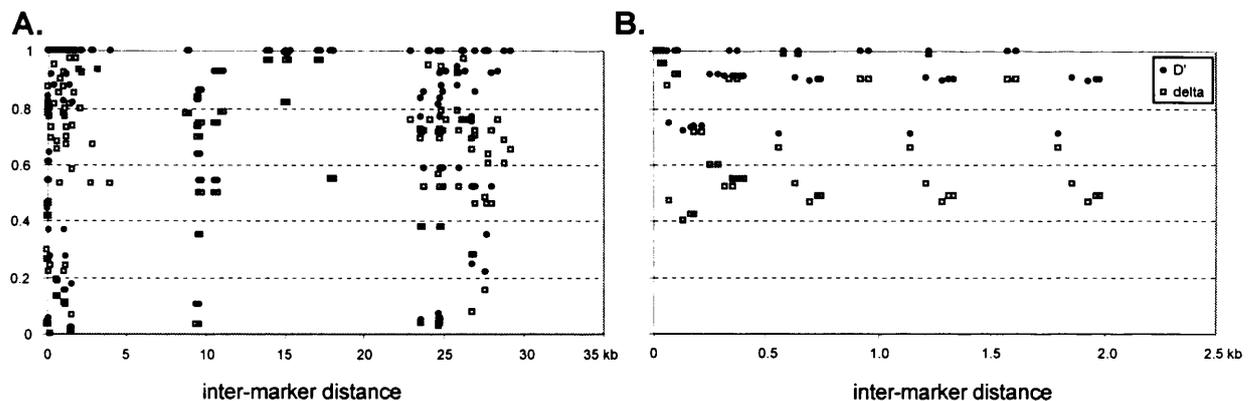


Figure 4.4. D' and Δ values between each pair of high-frequency (≥ 0.15 minor allele frequency) SNPs within LD blocks 1 (A) and 5 (B). Note the difference in block length. Black circles depict pairwise D' values, white squares pairwise Δ values. Δ values are consistently lower than D' values.

Δ is, however, a particularly useful measure of LD because it has a formal relationship to effective population size, recombination rate per unit distance and inter-marker distance (see section 4.1.1). For each SNP pair in the MHC target regions, Δ and inter-marker distance d was known. The equation $\Delta^2 = 1/(1+4N_e r d)$ was then used to calculate least squares - best fit values for the product $N_e r$ (effective population size

multiplied by recombination rate per unit distance) for the observed Δ vs. distance (d) values in each LD block. This value is lowest for block 6, and highest for block 5 (see Table 4.1). Using the least squares - best fit values, the best-fit curve for each LD block was plotted (Figure 4.5).

Table 4.1. The six LD blocks identified in these LD analyses. Three are in the first target region (*COL11A2-DOA* interval) and three in the second target region (*RING3-TAP2* interval). The 5' and 3' SNPs that define each block, along with block size, are shown. The number of high frequency SNPs refers to all SNPs with ≥ 0.15 minor allele frequency. The best-fit values (per Mb) were calculated from the observed pairwise Δ values for SNPs in each block (see text).

		5' marker	3' marker	size	number of high freq SNPs	best fit value for $N_e \times r$	cM/Mb if $N_e=10000$
<i>COL11A2-DOA</i> interval	block 1	-100.49G/A	-71.30A/G	≥ 29.2 kb	18	31	0.31
	block 2	-68.68A/G	-49.23A/G	19.4 kb	52	58	0.58
	block 3	-44.50A/T	-28.13A/G	≥ 16.4 kb	13	186	1.86
<i>RING3-TAP2</i> interval	block 4	H6C/A	JJ7A/C	≥ 17.2 kb	4	148	1.48
	block 5	J13C/T	Jjk14G/A	2.0 kb	10	495	4.95
	block 6	J5C/T	K8G/A	≥ 30.5 kb	7	19	0.19

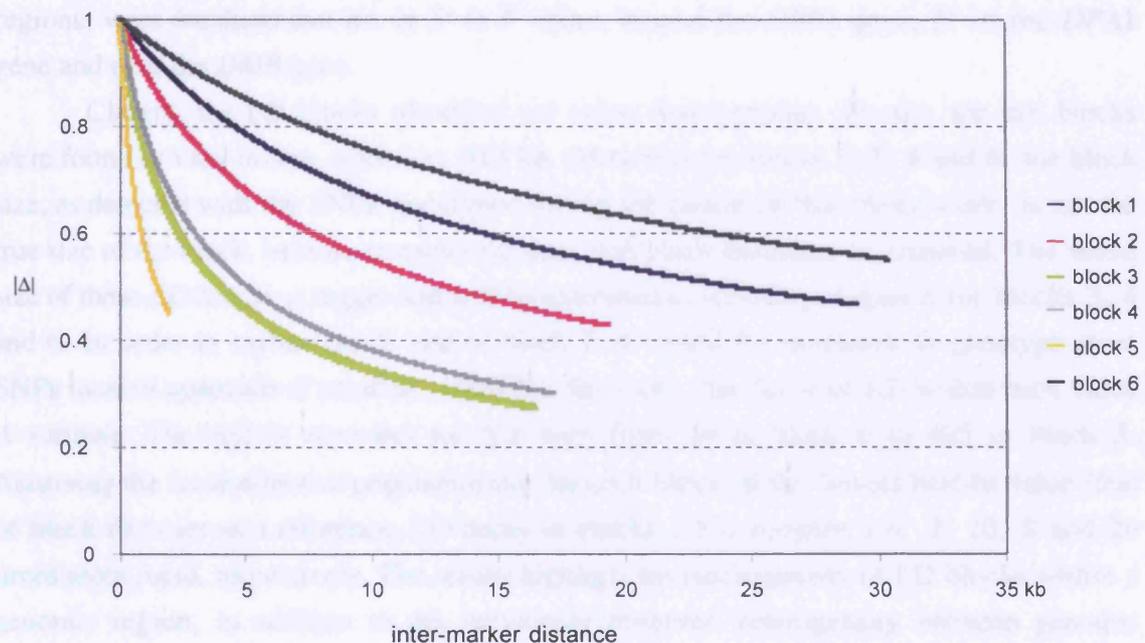


Figure 4.5. LD decays at different rates within LD blocks. Lines show the decay of the $|\Delta|$ measure with distance within the six LD blocks. Lines (colour-coded as shown on the right) show the least squares - best fit curves for the relationship $\Delta^2 = 1/(1+4N_e r d)$ where N_e = effective population size, r = recombination rate per Mb interval and d is the distance in Mb between markers.

4.3. DISCUSSION

The high density SNP map allows examination of LD patterns across both target regions (the *COL11A2-DOA* and *RING3-TAP2* interval) at high resolution. The aim was to localise regions of LD breakdown, and therefore putative recombination hotspots. For both target regions, LD analysis revealed a block-like pattern of marker association. High-frequency SNPs are more suited to reveal these LD blocks, because they are more likely to be older,

and have had time to recombine onto different haplotypes. Younger (low-frequency) SNPs have not necessarily been present in the population for long enough, and have not yet been fully re-shuffled onto different haplotypes by recombination. The difference in LD block detection using all SNPs *vs.* filtering out low-frequency SNPs is evident particularly for the *RING3-TAP2* interval, where there are several low-frequency SNPs in blocks 4 and 6 (Figure 4.3Ai-ii). The *COL11A2-DOA* interval LD plots are affected less by the filtering out low-frequency SNPs (Figure 4.2Ai-ii), simply because LD blocks 1-3 do not contain many SNPs with minor allele frequency <0.15.

In both target intervals, LD did not decline gradually over distance, but instead collapsed abruptly within localised regions. The high density of SNPs allowed refinement of these regions of free association (the location of LD block boundaries) to DNA segments shorter than 5 kb. Such regions are good candidates for containing recombination hotspots, because LD patterns suggest that they have been active in recombination in the history of the UK North European population. Three regions of LD breakdown, *i.e.* candidate hotspot regions, were localised that lie, in 5' to 3' order, around the *DPB1* gene, 3' of the *DPA1* gene and near the *DMB* gene.

Clearly, the LD blocks identified are rather heterogenous. Firstly, the six blocks were found to vary in size, from 2 to 30.5 kb. (Note that for blocks 1, 3, 4 and 6, the block size, as detected with the SNPs genotyped during the course of this thesis work, is not the true size of the block, because in each case only one block boundary is captured. The actual size of these LD blocks is bigger and will be examined in summary chapter 8 for blocks 3, 4 and 6. In order to capture the 5' end of block 1, it would be necessary to genotype more SNPs located upstream of position -100000.) Secondly, the decay of LD within each block is variable. The best-fit estimates for $N_e r$ vary from 19 in block 6 to 495 in block 5. Assuming the same effective population size for each block, if the lowest best-fit value (that of block 6) is set as a reference, LD decay in blocks 1-5 is roughly 1.6, 3, 10, 8 and 26 times more rapid, respectively. The results highlight the heterogeneity of LD blocks *within* a genomic region, in addition to the previously observed heterogeneity between genomic regions (for example, Reich *et al.*, 2001). In other words, more historical recombination, as measured by LD, is detected in some blocks. The assumption that within-block LD heterogeneity is a result of variability in crossover rate inside the blocks, is the most "recombinogenic" explanation for the observed data. The fact that within LD blocks, there is no clear correlation of LD measures with distance (Figure 4.4) suggests that other factors play an important role in lowering LD. Within-block LD differences could either result from variable gene conversion rates (as suggested by Ardlie *et al.*, 2001), or from other genetic factors such as drift or selection. It is also important to consider the possibility of genotyping errors, when single "misbehaving" markers are concerned. Four of the markers that do not show associations with other markers within their block (a subset of SNPs marked with squares in Figures 4.2B-C), deviate from Hardy-Weinberg equilibrium (homozygote excess

in all cases, SNPs -72.88, -72.83, -46.22, -44.16, see section 3.2.1). This suggests that for the genotypes of these markers could be erroneous, resulting in their unusual behaviour within the LD blocks. However, these SNPs represent only a small minority (2%) of the SNP genotype data set. Their inclusion in the Δ -based least squares best-fit calculations may make the decay of within-block LD slightly more rapid for blocks 1, 2 and 6, but it does not affect the conclusions about the block-like LD structure.

A close inverse correlation between nucleotide diversity and LD has been noted (Ardlie *et al.*, 2002). In general, this trend is evident from our data, as shown by the locations and minor allele frequencies of SNPs in Figures 4.2 and 4.3. We observe a clustering of SNPs with a high minor allele frequency around regions of LD breakdown. It is difficult to evaluate this, however, because in our sample, there was a bias of re-sequencing efforts towards these regions. Also, there is one deviation from the high LD - low diversity trend: the segment in the *COL11A2-DOA* interval characterised by extremely high SNP density and relatively high nucleotide diversity (Chapter 3). This entire region falls into (and defines) LD block 2, in disagreement with the inverse correlation between nucleotide diversity and LD.

The analysis of LD patterns is one way of identifying recombinationally active genomic regions. It is important to do this as a first step in recombination analysis, because characterising crossover hotspots at high-resolution using single molecule PCR methods must be targeted to specific locations - long-range PCR is only feasible over distances up to ~10 kb. Therefore, we must first have knowledge of which areas to target for crossover analysis. Regions of LD breakdown have probably undergone relatively high levels of historical recombination, and hence are good candidate regions for containing true meiotic recombination hotspots. Above, I have shown how three such regions were identified, around the *DPB1* gene, 3' of the *DPA1* gene and near the *DMB* gene. In the following three chapters (chapters 5-7), I will describe the results of crossover analyses on each of these regions.

5. IDENTIFICATION AND CHARACTERISATION OF A SPERM CROSSOVER HOTSPOT NEAR THE *DPA1* GENE

As a first step in the identification and characterisation of recombinationally active DNA segments in the human MHC class II, LD patterns were examined for evidence of LD breakdown, as described in Chapter 4. These LD patterns can vary in how well-defined the boundaries of LD blocks are, showing either clear or incomplete LD breakdown, as illustrated by the three LD breakdown regions identified. In two cases (downstream the *DPA1* gene and at the end of the *DMB* gene), I was able precisely locate the block boundaries on either side of the LD breakdown regions (Figure 5.1). The third LD breakdown region, within the *DPB1* gene region, was equivocal and lacked clear block boundaries, showing some overlap of the 5' end of one LD block with the 3' end of another. LD breakdown is an indication of historical recombination activity, and hence, regions where it is observed are putative recombination hotspots.

Once such hotspot candidate regions have been identified, the next step is the recovery of molecules directly from sperm DNA that have undergone crossover in these specific regions. In the following three chapters, I first describe the characterisation of two crossover hotspots where there was good evidence for relatively high recombination activity, based on LD patterns. I present the data on crossovers recovered in the region downstream of the *DPA1* gene (this chapter), and in the *DMB* gene region (Chapter 6), and finally, in Chapter 7, the very challenging crossover assay across the *DPB1* region, where LD patterns suggested relatively weak recombination activity. Chapter 8 aims to summarise all data on recombination (LD, plus both crossover and gene conversion) in the MHC class II region generated in our laboratory to date.

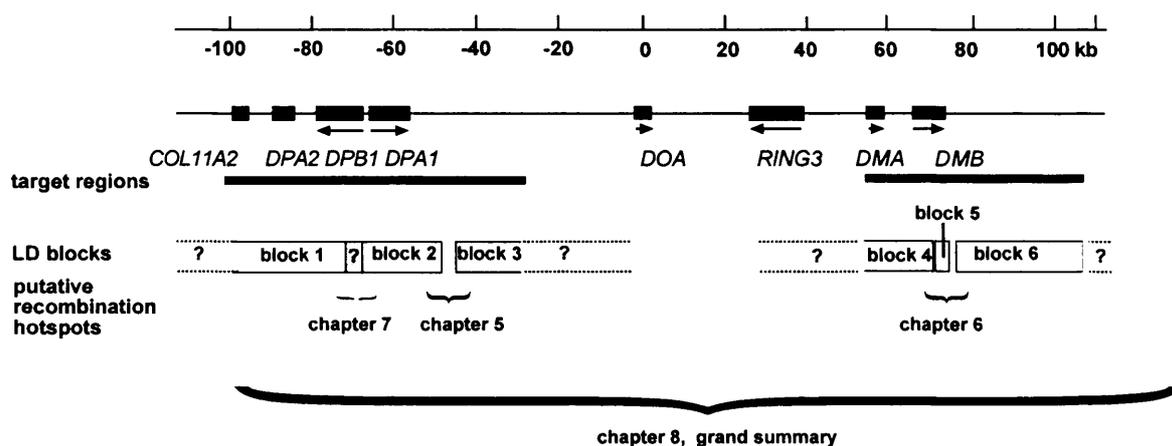


Figure 5.1. Overview of LD blocks and LD breakdown regions, and the organisation of the following four chapters. Crossover data on the two more convincing regions of LD breakdown (3' of *DPA1* and 3' end of *DMB*) are presented first (chapters 5 and 6), and in chapter 7, I describe the crossover assay across the *DPB1* region.

5.1. INTRODUCTION

LD data pointed to a region of LD breakdown roughly 8 kb downstream the *DPA1* gene (see Chapter 4 and Figure 5.1). The surrounding area was fully re-sequenced in seven semen donors to identify more SNPs in this region, allowing further refinement of the location of LD block boundaries. In the process, two large insertion/deletion polymorphisms a few kilobases 5' of the LD breakdown region were identified accidentally by PCR (see also Chapter 3). These polymorphisms were identified as SVA elements, each about 1.7 kb long, separated by ~1.7 kb; in this thesis, the insertion located further 5' is referred to as SVA1, and the one located further 3' as SVA2 (Figure 5.2). SVA elements are composite retrotransposons that consist of a short interspersed nuclear element (SINE) region, a minisatellite region and an Alu region. The MHC consensus sequence near the *DPA1* gene contains SVA2, located at position -52947, but SVA1 is missing. Hence, SVA2 occupies our basepair co-ordinates -52947 to -51235 in the MHC sequence, but the newly discovered SVA1 lacks co-ordinates, and only its insertion point (position -54646) and approximate size are known (Figure 5.2).

The LD breakdown region lies approximately 1.9 kb downstream of the 3' end of SVA2 (Figure 5.2). Thus, in SVA1/SVA2 heterozygotes, there is a long (1.7 kb) region of non-homology upstream a putative recombination hotspot. In addition to assaying crossover activity across the region of LD breakdown, the SVA polymorphisms allowed me to test whether a large heterozygosity only a few kilobases upstream has any effect on crossover rate or distribution. The hypothesis was that if sequence identity or similarity in the vicinity of a recombination hotspot is required for crossing-over, donors heterozygous at both SVA elements should have a lower recombination rate. SVA heterozygosity may also influence crossover distribution.

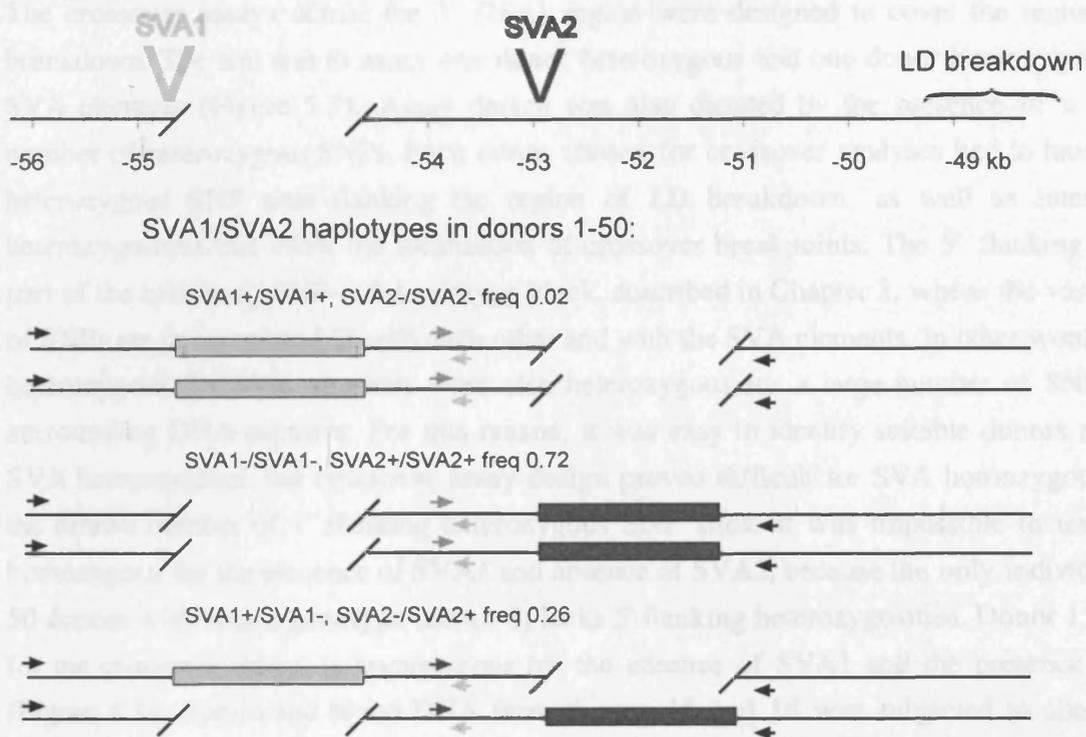


Figure 5.2. Location of the SVA elements relative to the region of LD breakdown. PCR using primers D-56.0F and D-51.0R (forward and reverse, black arrows) results in the amplification of product of the same length in all donors. However, PCR amplification of shorter fragments using D-56.0F or D-51.0R in combination with internal primers (black plus light grey or black plus dark grey arrows) reveals the presence of two large length polymorphisms, SVA1 and SVA2. In the 50 donors, only one person is homozygous for the presence of SVA1, 36 are homozygous for the presence of SVA2 and 13 are SVA1/SVA2 heterozygotes. The SVA1 or SVA2 status, *i.e.* their presence or absence, defines two extended haplotypes, which are characterised by a very high number of SNPs, each in nearly absolute association with other markers on its haplotype, as discussed in Chapter 3.

This work

Firstly, I wanted to test if the region of LD breakdown ~8 kb downstream of the *DPA1* gene corresponds to a true meiotic crossover hotspot, and secondly, if the large region of sequence non-identity (the polymorphic SVA elements) nearby suppresses recombination. Therefore, for crossover analyses, two donors, each with a different SVA status, were chosen. Donor 15 is an SVA1/SVA2 heterozygote, while donor 16 is homozygous for the SVA1 deletion and SVA2 insertion. I demonstrate that this region of LD breakdown contains an active and highly localised sperm crossover hotspot, and that SVA heterozygosity has little effect on crossover rate or distribution.

5.2. RESULTS

The crossover assays across the 3' *DPA1* region were designed to cover the region of LD breakdown. The aim was to assay one donor heterozygous and one donor homozygous for the SVA elements (Figure 5.3). Assay design was also dictated by the presence of a sufficient number of heterozygous SNPs. Each donor chosen for crossover analyses had to have suitable heterozygous SNP sites flanking the region of LD breakdown, as well as internal SNP heterozygosities that allow the localisation of crossover breakpoints. The 5' flanking region is part of the extremely SNP-rich haplotype block, described in Chapter 3, where the vast majority of SNPs are in complete LD with each other and with the SVA elements. In other words, donors heterozygous for SVA elements were also heterozygous for a large number of SNPs in the surrounding DNA segment. For this reason, it was easy to identify suitable donors among the SVA heterozygotes, but crossover assay design proved difficult for SVA homozygotes, due to the limited number of 5' flanking heterozygous SNP sites. It was impossible to test a donor homozygous for the presence of SVA1 and absence of SVA2, because the only individual in the 50 donors with such a genotype (donor 8) lacks 5' flanking heterozygosities. Donor 15, selected for the crossover assay, is homozygous for the absence of SVA1 and the presence of SVA2 (Figure 5.3). Sperm and blood DNA from donors 15 and 16 was subjected to allele-specific PCR in order to amplify recombinant molecules. Full details of allele-specific primer sequences and PCR cycling conditions can be found in Chapter 2.

5.2.1. Crossover rate

For both donors, crossover molecules were amplified from sperm DNA in PCR series containing various numbers of DNA input molecules (Table 5.1). Donor 15 was assayed for one recombinant phase only, while donor 16 was assayed in both orientations (in Figure 5.3, 5' top haplotype to 3' bottom haplotype = orientation A, 5' bottom haplotype to 3' top haplotype = orientation B) to test for reciprocity. Blood DNA PCRs, used as a control for *in vitro* artefacts and for the meiosis-specificity of crossovers, were seeded with the highest number of DNA input molecules used in sperm PCRs. The number of DNA input molecules equals the number of amplifiable DNA molecules per haplotype, which is calculated as 12 pg per amplifiable molecule. This 12 pg per amplifiable molecule estimate results from the following reasoning: 3 pg (weight of one haploid genome) x 2 (we are only detecting one of the two recombinant phases in each assay) x 2 (assumption of 50% PCR efficiency, *i.e.* 50% chance that one DNA molecule will yield PCR product, an estimate derived from extensive work with long-range PCR in our laboratory, A.J. Jeffreys, personal communication). Numbers of amplifiable molecules in both donors was further confirmed by performing PCRs at limiting dilutions at single-molecule

level, as described in Jeffreys *et al.* (1994). The number of amplifiable DNA molecules with which each PCR is seeded is referred to as the "input pool size".

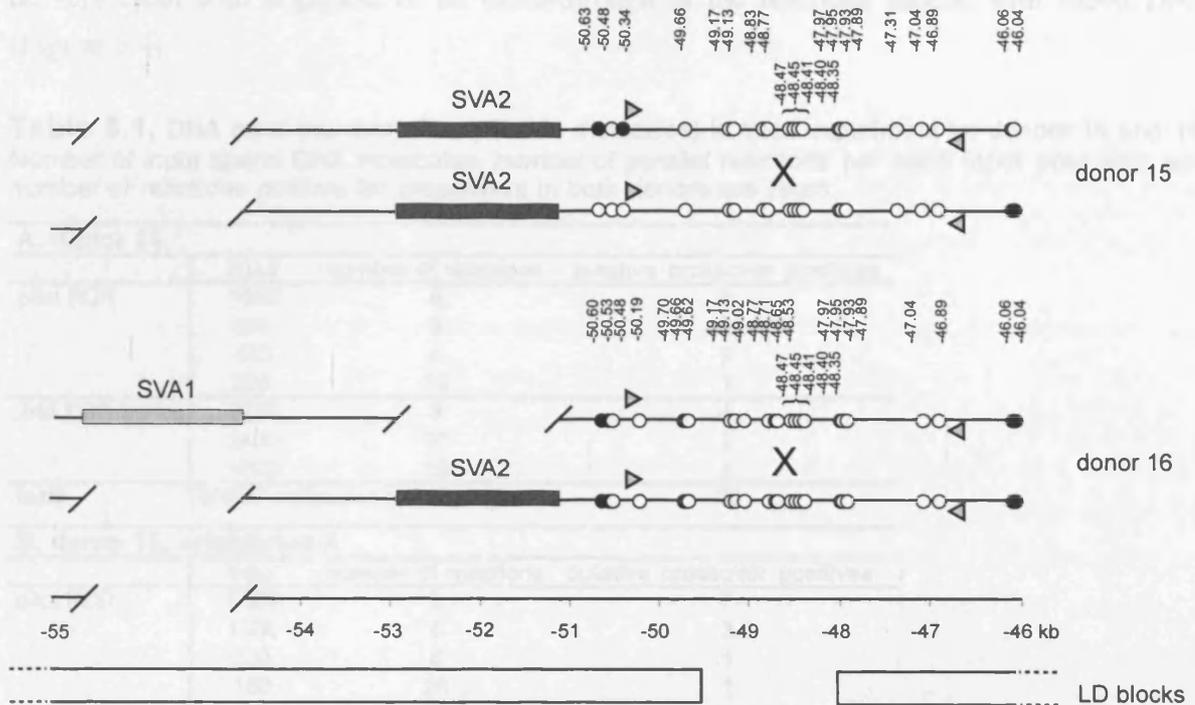


Figure 5.3. SVA1/SVA2 status and heterozygous SNP sites across the test interval in donors 15 and 16. SVA1 and SVA2 elements are shown as light and dark grey boxes. Black circles depict SNPs used as selector sites for allele-specific PCRs, white circles are other heterozygous SNPs within the assayed interval. Locations of universal primers used for tertiary PCR amplification are indicated with grey arrowheads. LD blocks (see Chapter 4) are shown as white boxes at the bottom.

To selectively amplify DNA molecules that have undergone crossover, sperm DNA digested with *BstZ171*, a restriction enzyme that cuts immediately outside the test interval to make the DNA more soluble, was subjected to one round of allele-specific PCR amplification ("primary PCR"). This was followed by S1 nuclease digestion, to rid the sample of single-stranded DNA (e.g. panhandle loop structures) that could generate PCR artefacts that could interfere with crossover detection. The S1 digested PCR products were then re-amplified ("secondary PCR") with allele-specific primers located internally to the allele-specific primers used in the primary PCR. If allele-specific primers exhibit good efficiency and specificity, crossover molecules should be detectable on an ethidium bromide stained agarose gel after the two rounds of allele-specific PCR amplification. In the 3' *DPA1* assay, signals from the PCR products were very weak on an ethidium bromide stained gel after secondary PCR amplification, due to the relatively low efficiency of primers. Therefore, all secondary PCR products were

subjected to a third round of PCR amplification with universal primers nested just inside the internal allele-specific primer sites ("tertiary PCR", see grey arrowheads in Figure 5.3). After tertiary PCR, crossover positive reactions were easily distinguishable because the assay proved to be very clean with negligible or no bleed-through in the reactions seeded with blood DNA (Figure 5.4).

Table 5.1. DNA input (number of amplifiable molecules) in each experiment for donors 15 and 16. Number of input sperm DNA molecules, number of parallel reactions per each input pool size, and number of reactions positive for crossovers in both donors are listed.

A. donor 15			
	input	number of reactions	putative crossover positives
pilot PCR	1600	8	3
	800	8	0
	400	8	2
	200	19	1
2nd PCR series	3200	8	4
	2400	10	2
	1600	10	8
total	91800 molecules (=1102 ng)		20
B. donor 16, orientation A			
	input	number of reactions	putative crossover positives
pilot PCR	5000	8	7
	1500	8	3
	500	8	1
	150	20	1
	50	20	3
2nd PCR series	2800	20	10
	2250	20	11
	1400	20	4
total	189000 molecules (=2268 ng)		40
C. donor 16, orientation B			
	input	number of reactions	putative crossover positives
pilot PCR	3350	10	8
	2800	21	8
	2250	21	9
	1400	20	3
total	167550 molecules (=2010 ng)		28

Pilot PCR series contained a range of input pool sizes to obtain an estimate of crossover rate. Once the suitable number of input molecules was established, a second PCR series was set up to harvest more crossover positive reactions if necessary (Table 5.1). After each experiment, the amplification products were examined for the number of putative crossover-positive reactions, as well as the amount of bleed-through amplification. Blood PCRs were performed only at the highest input pool sizes.

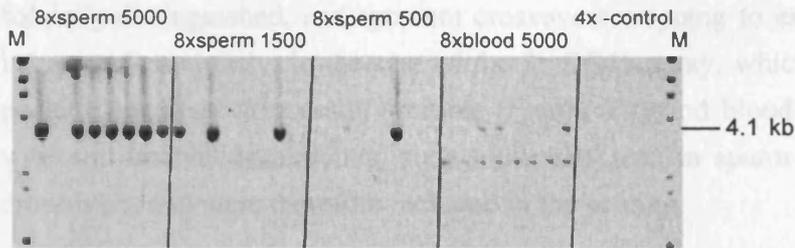


Figure 5.4. An example of tertiary crossover PCR products in donor 16 (orientation A). There are eight parallel reactions at each input pool size. Crossover positive reactions were easily distinguishable (e.g. seven positives in the eight sperm DNA PCRs at 5000 molecule input). Blood PCRs were only seeded with the highest input DNA. M, marker (*Hind*III digested λ DNA plus *Hae*III digested ϕ x DNA).

Crossover positive reactions were counted for each input pool size, and crossover rates were calculated using Poisson maximum-likelihood methods correcting for multiple crossover molecules in the positive reactions. Crossover rate in donor 15 was 0.34×10^{-3} ($0.2 - 0.53 \times 10^{-3}$ 95% confidence intervals), *i.e.* approximately one in 2940 of the screened sperm DNA molecules had undergone a crossover in the region 3' of the *DPA1* gene. For donor 16, crossover rate was 0.3×10^{-3} ($0.21 - 0.43 \times 10^{-3}$ 95% confidence intervals) in orientation A, and 0.22×10^{-3} ($0.14 - 0.33 \times 10^{-3}$ 95% confidence intervals) in orientation B. This translates to approximately one crossover in 3330 sperm DNA molecules for the orientation A assay, and one crossover in 4550 sperm DNA molecules for orientation the B assay; thus the crossover rate detected in orientation B assay is ~ 1.4 times lower than that detected in orientation A assay. When the crossover data from orientations A and B in donor 16 are pooled, then crossover rate is 0.26×10^{-3} ($0.19 - 0.34 \times 10^{-3}$ 95% confidence intervals). The number of crossovers per total number of molecules screened is not significantly different between any of the data sets in Table 5.1 (using a Fisher exact test, 2x2 contingency table, the lowest probability $P=0.224$ was found between orientations A and B in donor 16).

5.2.2. Distribution of crossover breakpoints

All crossover positive reactions were transferred onto dotblots, and crossover breakpoints were typed by ASO hybridisation. Crossover breakpoints were not randomly distributed across the test interval, but instead clustered into the same ~ 1.5 kb long DNA segment in all donors (Table 5.2 and Figure 5.5). The data therefore are consistent with the presence of a meiotic recombination hotspot ~ 8.5 kb from the end of the *DPA1* gene, henceforth referred to as the 3' *DPA1* hotspot. In donor 15, four crossovers occurred in the 5' end interval, between the selector SNP site used in secondary PCRs and the first 5' heterozygous SNP site. Such molecules could either be genuine crossovers which have their breakpoint located in the end interval, or bleed-through amplification from a progenitor haplotype. These two types cannot be

formally distinguished, and apparent crossovers mapping to end intervals should therefore be interpreted cautiously. In the case of the 3' DPA1 assay, which was "clean" *i.e.* all crossover positive reactions were easily scorable (Figure 5.4) and blood PCRs did not show molecules with end interval breakpoints, such molecules seen in sperm PCRs are presumably genuine crossovers, and were therefore included in the scoring.

Table 5.2. Distribution of crossover breakpoints in the 3' DPA1 region in donor 15, and donor 16 orientations A and B. Each row summarises the data for reactions at a particular input pool size. Per SNP interval, the number of crossover positive plus crossover negative reactions is shown. Reactions containing crossover mixtures were excluded as explained in section 2.2.7.6. SNP intervals where no crossovers were seen were pooled. Where only one crossover was observed for a given interval, no Poisson adjustment was calculated.

intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions

Crossover distribution in donor 15

input pool size	number of reactions		intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions															
	positive	negative	-50.34	-49.88	-49.13	-48.83	-48.77	-48.63	-48.47	-48.46	-48.40	-48.35	-47.97	-47.86	-47.93	-47.89	-47.51	-46.89
3200	4	4	1+7	0+8	1+7	0+7	1+7	0+8	0+8	0+8	0+8	1+7	0+8	1+7	0+8	0+8	0+8	0+8
2400	5	5	0+10	0+10	3+7	0+7	2+5	0+9	0+9	0+9	0+9	1+8	1+7	0+8	1+8	1+9	0+10	0+10
1600	11	7	3+15	0+18	2+16	0+17	3+14	0+15	1+14	0+15	1+14	6+12	0+18	0+18	0+18	0+18	0+18	0+18
800	0	8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8
400	2	6	0+8	0+8	1+7	0+8	0+8	0+8	0+8	0+8	0+8	1+7	0+8	0+8	0+8	0+8	0+8	0+8
200	1	18	0+19	0+19	0+19	1+18	0+19	0+19	0+19	0+19	0+19	0+19	0+19	0+19	0+19	0+19	0+19	0+19
Total crossovers			4	0	7	1	6	0	1	0	1	9	1	1	1	1	1	0
Poisson-adj. crossovers			4.3	0	7.3	1	6.6	0	1	0	1	9.8	1	1	1	1	1	0
Molecules screened			81600	91800	91800	79800	83000	84800	84800	84800	84800	86400	87000	87000	89400	91800	91800	91800
inter-SNP distance (bp)			891	526	332	43	233	48	19	46	47	380	19	23	36	574	428	428
cM/Mb			6.8	0	23.9	29.1	34.1	0	62.2	0	25.1	28.8	60.5	50	31.9	1.9	0	0

Crossover distribution in donor 16, orientation A

input pool size	number of reactions		intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions															
	positive	negative	-50.18	-49.82	-49.17	-48.13	-48.02	-48.77	-48.71	-48.86	-48.83	-48.35	-47.97	-47.93	-47.86	-47.89	-46.89	
5000	7	1	0+8	0+8	0+8	0+8	4+4	3+3	1+4	1+4	0+8	3+5	0+8	0+8	0+8	0+8	0+8	
2800	10	10	0+20	0+20	0+20	1+19	6+14	2+17	1+18	1+18	0+20	2+18	0+20	0+20	0+20	0+20	0+20	
2250	11	9	0+20	0+20	0+20	0+20	3+17	1+17	3+15	1+16	0+20	7+13	0+20	0+20	0+20	0+20	0+20	
1500	3	5	0+8	0+8	0+8	0+8	0+8	0+8	0+8	1+7	0+8	1+7	0+8	1+7	0+8	0+8	0+8	
1400	4	16	0+20	0+20	0+20	0+20	1+19	1+19	0+20	1+19	0+20	1+19	0+20	1+19	0+20	0+20	0+20	
500	1	7	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	1+7	0+8	0+8	0+8	0+8	0+8	
150	1	19	0+20	0+20	0+20	0+20	0+20	0+20	0+20	0+20	0+20	1+19	0+20	0+20	0+20	0+20	0+20	
50	3	17	0+20	1+19	0+20	0+20	0+20	0+20	0+20	0+20	1+19	0+20	1+19	0+20	0+20	0+20	0+20	
Total crossovers			0	1	0	1	14	7	5	6	0	17	0	1	0	1	0	
Poisson-adj. crossovers			0	1	0	1	15.1	7	5	6.6	0	18.9	0	1	0	1	0	
Molecules screened			186000	186000	186000	186000	186000	171700	186700	164450	186000	186000	186000	186000	186000	186000	186000	
inter-SNP distance (bp)			562	454	46	106	269	62	33	136	148	380	43	36	1003	1003	1003	
cM/Mb			0	1.2	0	5	29.7	64.8	90.9	29.5	0	26.3	0	14.7	0	0	0	

Crossover distribution in donor 16, orientation B

input pool size	number of reactions		intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions																
	positive	negative	-50.18	-49.82	-49.17	-48.13	-48.02	-48.77	-48.71	-48.86	-48.83	-48.47	-48.46	-47.41	-48.40	-48.35	-47.97	-47.86	-47.83
3350	8	2	0+10	2+8	0+10	1+9	4+4	0+7	0+7	1+7	0+8	0+8	0+8	0+8	0+8	3+7	0+10	0+10	0+10
2800	8	13	0+21	1+20	0+21	0+21	2+19	1+19	0+20	1+20	1+20	0+21	1+19	0+20	0+20	2+19	0+21	1+20	0+21
2250	9	12	0+21	1+20	0+21	1+20	3+16	0+17	1+16	1+16	0+17	0+17	0+17	0+17	1+17	5+16	1+20	0+21	0+21
1400	3	17	0+20	0+20	0+20	0+20	2+18	0+20	0+20	0+20	0+20	0+20	0+20	0+20	0+20	1+19	0+20	0+20	0+20
Total crossovers			0	4	0	2	11	1	1	3	1	0	1	0	1	11	1	1	0
Poisson-adj. crossovers			0	4.2	0	2	12.5	1	1	3	1	0	1	0	1	11.7	1	1	0
Molecules screened			167550	167550	167550	167550	156350	145700	145700	151850	151850	148050	148050	148050	151300	167550	167550	167550	167550
inter-SNP distance (bp)			562	454	46	106	269	62	33	136	48	19	31	13	47	380	19	23	1040
cM/Mb			0	5.5	0	9.6	29.7	11.1	20.8	14.5	13.7	0	21.6	0	14.1	18.4	31.4	25.9	0

Recombination activity in cM/Mb for each SNP interval was calculated (crossover frequency x 100 / inter-marker distance in Mb, Table 5.2), and is plotted in Figure 5.5. When the data are presented graphically in this way, short SNP intervals where single crossovers were observed get accentuated. Hence, the distribution of crossovers in the three different assays appears dissimilar at first glance (Figure 5.5). Also, the crossover distributions in donor 15 and donor 16 in orientation B look more similar to each other than donor 16 orientation A vs. B or donor 15 vs. donor 16 in orientation A. In particular, it looks as if there is a short "cold" patch

inside the hotspot, when crossovers are assayed in donor 16 in orientation A (see Figure 5.5Bi, arrowed region). Within this cold patch, no crossover breakpoints were observed. However, in donor 15 and in donor 16 orientation B, crossovers are resolved within this region. To test whether the distributions actually are different, the Kolmogorov-Smirnov test was applied. This test compares two observed distributions to see if they could have been drawn from the same underlying distribution. Each shared SNP interval, and each assay combination where distributions looked different (donor 16A vs. donor 15, donor 16B vs. donor 15, donor 16A+B pooled vs. donor 15, and donor 15 plus donor 16B pooled vs. donor 16A), was tested for differences in distribution. The Kolmogorov-Smirnov test showed no significant difference (lowest P value observed 0.32) in crossover distribution between any of the assays. Thus, regardless of SVA status of the donor, crossovers cluster into the same hotspot.

The similar crossover distribution was evident when the cumulative frequency of crossovers was plotted for each donor. From the cumulative crossover frequency of the combined data, a least-squares best-fit curve was calculated, assuming that crossover breakpoints are normally distributed across the hotspot (Figure 5.6A). The underlying normal distribution of this curve (Figure 5.6B) has a standard deviation of 415 bp; within the 3' *DPA1* hotspot 95% of crossovers therefore fall into a region ~1.6 kb wide (= 4 times standard deviation). The normal distribution curve also reveals the location of the theoretical centrepoint of the hotspot, where crossover activity is highest (position -48557). The centrepoint is located just at the end of a human endogenous retroviral (HERV) repeat element. The peak crossover activity of the 3' *DPA1* hotspot is 27 cM/Mb (Figure 5.6B).

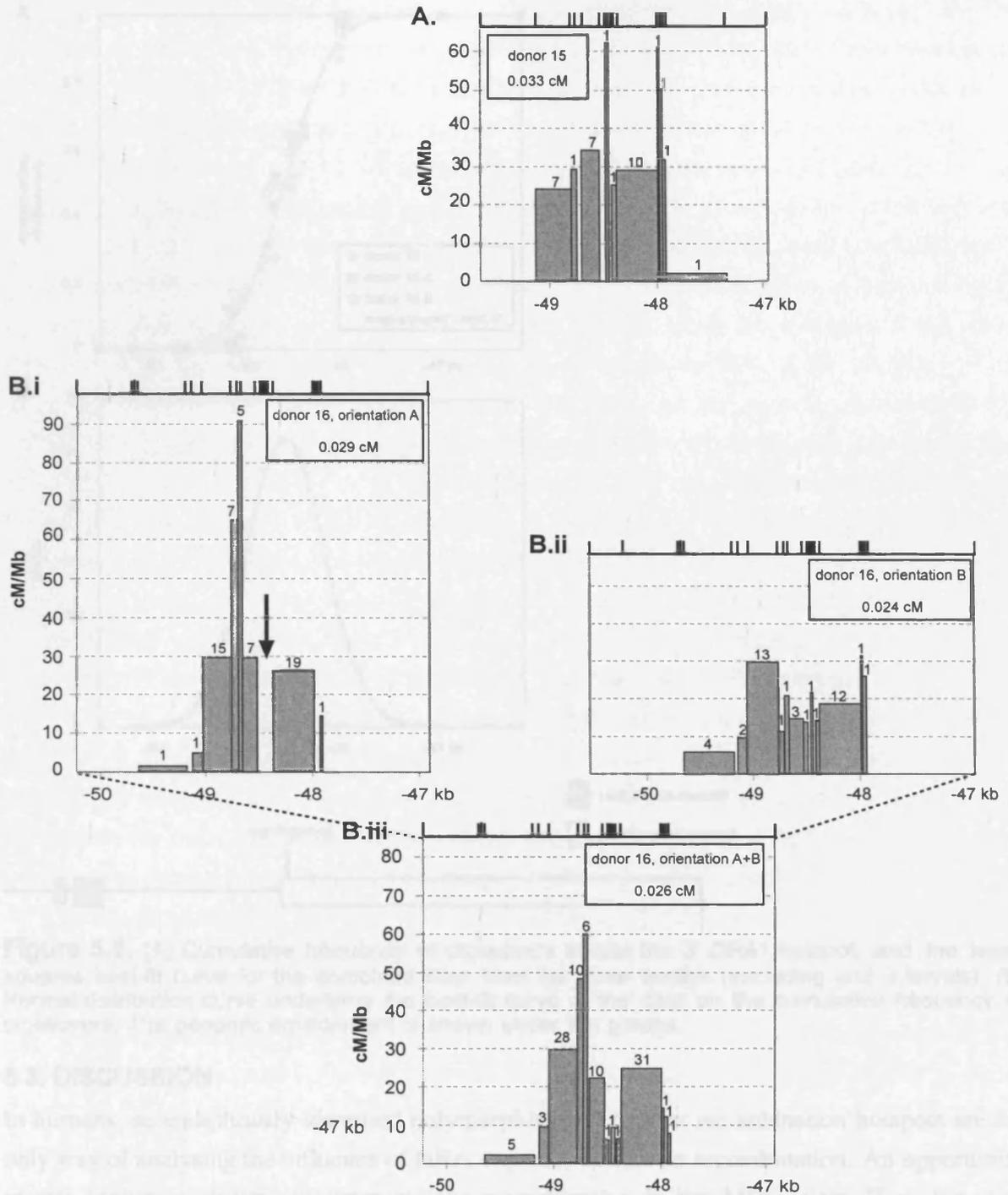


Figure 5.5. Crossover activity in each SNP interval in cM/Mb, in donor 15 (A) and donor 16 (B, *i-iii*). Heterozygous SNP sites are depicted as ticks above the graphs. All end intervals have been excluded. The number of crossovers observed in each SNP interval is shown above the bars. The 162-bp gap where no crossover breakpoints were seen in donor 16, orientation A, is indicated by an arrow. Total genetic map distance in cM is shown as an inset in each graph.

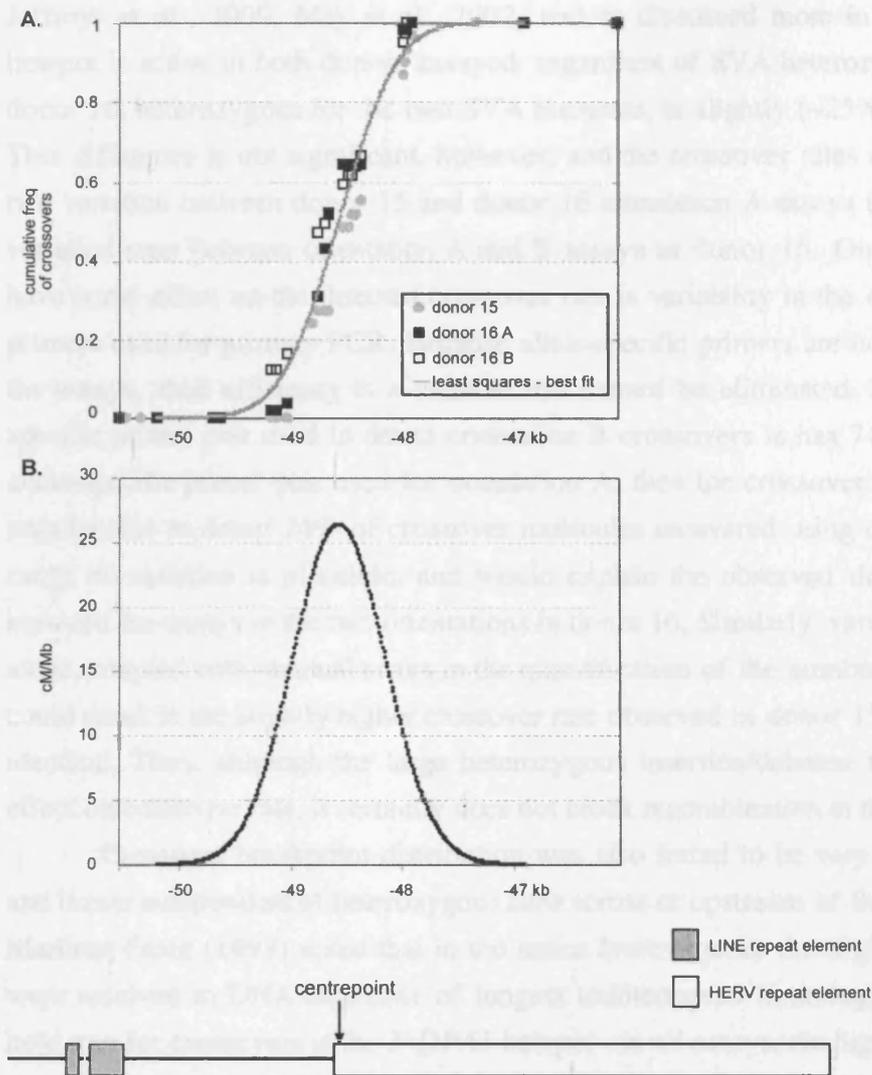


Figure 5.6. (A) Cumulative frequency of crossovers across the 3' *DPA1* hotspot, and the least-squares best-fit curve for the combined data from the three assays (excluding end intervals). (B) Normal distribution curve underlying the best-fit curve of the data on the cumulative frequency of crossovers. The genomic environment is shown under the graphs.

5.3. DISCUSSION

In humans, serendipitously identified polymorphisms at or near recombination hotspots are the only way of analysing the influence of DNA sequence factors on recombination. An opportunity to test sequence identity requirements for recombination in the MHC class II region was encountered when two polymorphic SVA insertion/deletions were discovered near a region of LD breakdown, downstream of the *DPA1* gene.

Within the region of LD breakdown, a highly localised sperm crossover hotspot named the 3' *DPA1* hotspot was identified, with a relatively high recombination activity. It has a similar morphology to other human crossover hotspots previously characterised at high resolution (e.g.

Jeffreys *et al.*, 2000, May *et al.*, 2002, and as discussed more in detail in Chapter 8). The hotspot is active in both donors assayed, regardless of SVA heterozygosity. Crossover rate in donor 16, heterozygous for the two SVA elements, is slightly (~25%) lower than in donor 15. This difference is not significant, however, and the crossover rates could be very similar. The rate variation between donor 15 and donor 16 orientation A assays is in fact similar to the rate variation seen between orientation A and B assays in donor 16. One variable which will also have some effect on the detected crossover rate is variability in the efficiency of allele-specific primers used for primary PCR. Because allele-specific primers are not identical between any of the assays, their efficiency is a variable that cannot be eliminated. For example, if the allele-specific primer pair used to detect orientation B crossovers is has 74% of the efficiency of the allele-specific primer pair used for orientation A, then the crossover assay in orientation B will only be able to detect 74% of crossover molecules recovered using orientation A primers. This range of variation is plausible, and would explain the observed difference in crossover rate between the assays in the two orientations in donor 16. Similarly, variability in primer efficiency alone, coupled with residual errors in the quantification of the number of amplifiable molecules, could result in the slightly higher crossover rate observed in donor 15, even if rates are actually identical. Thus, although the large heterozygous insertion/deletion may have a weak negative effect on crossover rate, it certainly does not block recombination at this hotspot.

Crossover breakpoint distribution was also found to be very similar in the two donors, and hence independent of heterozygous sites across or upstream of the test interval. Dooner and Martinez-Ferez (1997) noted that in the maize *bronze* gene, the highest number of crossovers were resolved in DNA segments of longest uninterrupted homology. This does generally not hold true for crossovers at the 3' *DPA1* hotspot - in all assays, the highest recombination activity (in cM/Mb) is seen within segments where SNPs are in separated by less than 50 bp (Figure 5.5). The only suggestion that multiple heterozygosities may affect recombination comes from the orientation A assay in donor 16: there is a gap in the occurrence of crossover breakpoints in the 162-bp segment between SNPs -48.53 and -48.35, containing six SNPs (see Figure 5.5, arrowed region in the cM/Mb activity), even though this region is right in the middle of the 3' *DPA1* crossover hotspot, and thus is expected to show crossovers. This suggests that crossover resolution is avoided in this region. However, statistical tests failed to detect any significant difference in crossover breakpoint distribution between any assays. Hence, there is either is no real difference, or if there is, we cannot detect it in our data set due to the limited number of crossover breakpoints landing in individual SNP intervals. Even if genuine, this phenomenon of site-specific avoidance of crossover resolution cannot be a result of simply the presence of six heterozygosities in this short DNA segment, because crossover breakpoints are observed in the same region in donor 15 and in the donor 16 orientation B assay, where the same

heterozygosities are present (see Figure 5.5). Instead it could only be explained by an effect somehow specific to the linkage phase/haplotype assayed in orientation A in donor 16. To further investigate this, more crossovers in both orientations would have to be recovered across the 3' *DPA1* hotspot, preferably from several donors.

The SVA heterozygosity, *i.e.* a long segment of non-homologous sequence less than 2 kb upstream of the crossover hotspot, has little if any effect on suppressing crossovers at the 3' *DPA1* hotspot. Hence, sequence similarity this far from the hotspot is not required for successful crossing-over. Homology search in the context of crossover is presumably only performed at an extremely localised level, likely by the single-stranded 3' overhanging DNA tails generated from the initiating double-strand break. Initial pairing and alignment of homologous chromosomes in meiosis I, prior to invasion of the single-stranded 3' overhanging DNA, will occur regardless of relatively long regions (~2 kb) of non-homology. A "homology check" mechanism that is insensitive to this type of non-homology would also be evolutionarily sensible, because the human genome contains many polymorphic insertion/deletions of variable size. If their presence inhibited proper alignment of homologues, a much larger fraction of chromosomes would likely fail to disjoin.

These results are in accordance with two studies in maize that considered the effect of large heterozygous insertion/deletions, similar in size to the SVA elements described here, on recombination rate. A 1.4 kb Mutator1 (*Mu1*) transposon insertion less than 200 bp away from the recombination hotspot within the *al* gene, reduced recombination rate by approximately half when hemizygous, but crossover breakpoint distribution remained the same (Xu *et al.*, 1995). Dooner and Martinez-Ferez (1997) reported that large hemizygous insertions at the maize *bz* gene can suppress recombination in its immediate vicinity (but only a few hundred basepairs away), and thus *create* recombination hotspots in neighbouring regions where crossovers can occur. This may be what we observe at the 3' *DPA1* hotspot, even though the heterozygous insertion in our study was further away, and therefore presumably has less of an effect on recombination. If the presence of the SVA elements (at least in the heterozygous constellation) suppresses crossovers in their immediate vicinity, say several hundred basepairs on either side, then the 3' *DPA1* hotspot may have arisen in a region where crossovers are "allowed" (Figure 5.7). Another issue to consider is the effect of multiple heterozygosities on the distribution of crossovers. The region upstream of the 3' *DPA* hotspot, in addition to containing the SVA elements, is extremely SNP rich (see Chapter 3). If high nucleotide diversity suppresses crossovers, this could be another reason why the region upstream of the 3' *DPA1* hotspot is recombinationally inactive (Figure 5.7).

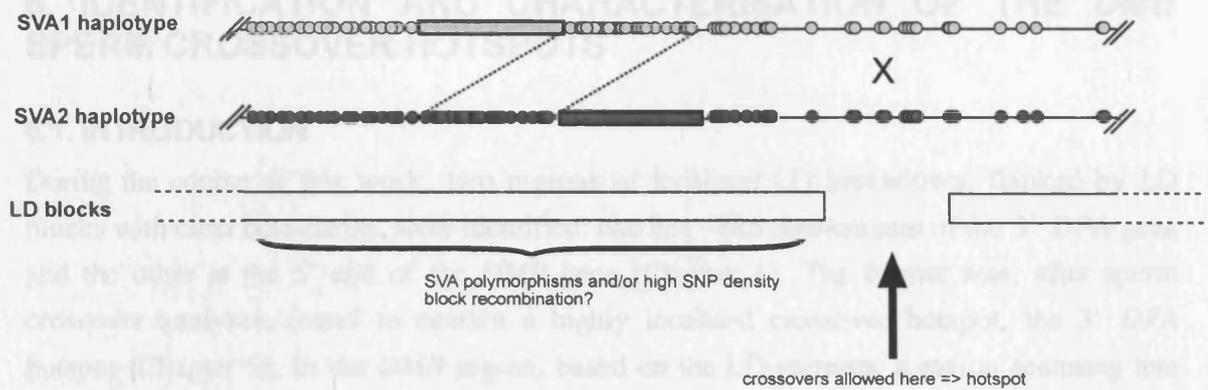


Figure 5.7. Schematic representation of how sequence diversity, observed in SVA heterozygotes upstream the 3' *DPA1* hotspot, may block crossovers. SVA1 and SVA2 are shown as light and dark gray boxes, respectively, and heterozygous SNPs on each haplotype as light or dark gray circles. The SVA heterozygosities create an out-of-register constellation of DNA segments immediately 3' of SVA1 and 5' of SVA2; heterozygous SNPs within this region, located on the two haplotypes, are connected with dashed lines.

The model shown in Figure 5.7 does not, however, explain why individuals who are homozygous for SVA elements and most SNPs in the region upstream of the 3' *DPA1* hotspot (such as donor 15, and 74% of our 50 donors), and therefore lack the crossover-suppressing heterozygosities, should also have crossover hotspot here. Population haplotype data (LD patterns) also suggests that the 3' *DPA1* hotspot has been active in the UK population for some time. It therefore seems more likely that the 3' *DPA1* hotspot is located in a region which is intrinsically recombinogenic (prone to double-strand breaks), regardless of upstream haplotypes or heterozygosities. The observed polymorphisms in upstream sequence would then simply be "passengers" on the recombinationally cold DNA segment, rather than the cause of it.

In conclusion, I have localised and characterised a novel crossover hotspot, called the 3' *DPA1* hotspot, ~8.5kb downstream the *DPA1* gene. The location of this crossover hotspot corresponds precisely to the region of LD breakdown. Approximately one sperm DNA molecule in 3000 undergoes crossover in the 3' *DPA1* hotspot. The high number of polymorphic sites, including large insertion/deletions, a few kilobases upstream has little effect on the crossover activity at the hotspot.

6. IDENTIFICATION AND CHARACTERISATION OF THE *DMB* SPERM CROSSOVER HOTSPOTS

6.1. INTRODUCTION

During the course of this work, two regions of localised LD breakdown, flanked by LD blocks with clear boundaries, were identified: one lies ~8kb downstream of the 3' *DPA* gene and the other at the 5' end of the *DMB* gene (Chapter 4). The former was, after sperm crossover analyses, found to contain a highly localised crossover hotspot, the 3' *DPA* hotspot (Chapter 5). In the *DMB* region, based on the LD patterns, a region spanning less than 7 kb (approximately from position 70000 to 77000) was identified as the most likely candidate for harbouring a recombination hotspot. Within this region, LD was found to abruptly collapse, and markers in the centre of this region formed a "mini-LD block", which did not associate with the much longer LD blocks upstream and downstream. This suggested the possibility of a double hotspot, with two separate recombination hotspots located within only a few kilobases from each other. To investigate this, a sperm crossover assay was designed with allele-specific primer sites located in the region flanking the two putative hotspots.

This work

Sperm DNA from three donors was subjected to crossover analyses across the region of LD breakdown at the 3' end of the *DMB* gene. I show below how two adjacent crossover hotspots, approximately 3.3 kb apart, were identified. They show similar morphology but a ~10-fold difference in crossover rate. *DMB1*, the weaker hotspot, lies in the third intron of the *DMB* gene, whereas *DMB2* is located downstream in non-genic sequence.

Most of the work described in this chapter has been published (Jeffreys *et al.*, 2001).

6.2. RESULTS

To identify donors suitable for the *DMB* crossover assays, genotypes from the 50 donors across the LD breakdown region were examined for individuals with a high number of heterozygosities. Initially, in donors 1-50, only one man (donor 12) was identified who had suitable heterozygous sites flanking putative hotspot regions 5' and 3', coupled with good coverage of internal heterozygous markers. Four other individuals (donors 7, 39, 41 and 45) were heterozygous for flanking SNP sites, but none of these donors had particularly good coverage of heterozygous markers across the region of LD breakdown. They were deemed unsuitable for crossover assays, as crossover data from assays in these donors would mainly yield information on crossover rate, but little information on the shape of the putative hotspots. In an attempt to identify more informative semen donors, 50 additional individuals

available as a control for donors 63 and 87; however, blood DNA from donor 7 could be used as a control for these assays, because this man has the same heterozygous selector sites in the same linkage phase as donors 63 and 87 (Figure 6.2).

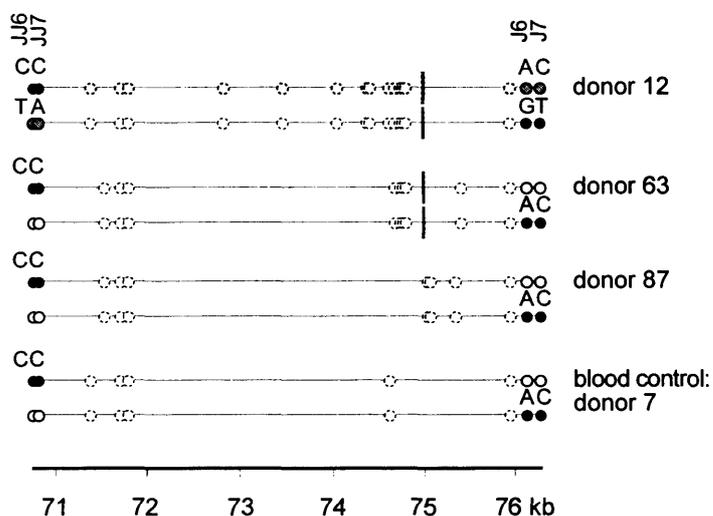


Figure 6.2. Linkage phase of *DMB* selector SNP sites in donors 12, 63 and 87 assayed for crossovers, plus donor 7, used as a blood DNA control in donor 63 and 87 crossover assays. The two leftmost and rightmost circles depict SNPs used as selector sites for allele-specific PCRs. Dashed circles are internal heterozygous SNP sites in each donor. Heterozygous GT repeats in donors 12 and 63 are shown as vertical ticks. Donor 12 was assayed for crossovers in both orientations (orientation A = top haplotype to bottom haplotype, black circles indicate selector sites, orientation B = bottom to top, gray circles indicate selector sites). Donors 63 and 87 were assayed in one orientation only, with black circles indicating selector sites.

6.2.1. Crossover rate

Sperm DNA in donor 12 was assayed for crossovers in both phases (orientations A and B) to investigate reciprocity (Figure 6.2). Crossovers in donors 63 and 87 were assayed in one recombinant phase only. Each assay consisted of at least two independent PCR experiments, with several pool sizes of input DNA molecules (Table 6.2). Blood DNA PCRs were set up either at all input pool sizes or at the highest input pool sizes only.

Following primary allele-specific PCR, S1 nuclease digestion and secondary allele-specific PCR, the crossover positive PCR products were visible on an ethidium bromide stained agarose gel (Figure 6.3). There was little or no bleed-through amplification from blood PCRs (Figure 6.3), indicating good specificity and efficiency of allele-specific primers.

Table 6.2. DNA input (number of amplifiable molecules) in each experiment for donors 12, 63 and 87. Number of input sperm DNA molecules, number of parallel reactions per each pool size, and number of reactions positive for crossovers are listed.

donor 12, orientation A			
	input	number of reactions	putative crossover positives
pilot PCR	6000	7	5
	1800	8	6
	600	8	2
	180	8	1
2nd PCR series	9600	15	14
	4800	14	10
	2400	15	8
	1200	15	1
3rd PCR series	9600	20	19
	4800	20	14
	2400	19	9
total	661440 molecules (=7937 ng)		89
donor 12, orientation B			
	input	number of reactions	putative crossover positives
1st PCR series	10000	20	20
	5000	20	17
	2500	20	10
2nd PCR series	10000	20	19
	5000	20	11
	2500	20	10
total	700000 molecules (=8400 ng)		87
donor 63			
	input	number of reactions	putative crossover positives
pilot PCR	6000	10	10
	3000	10	3
	1500	10	3
	750	14	5
2nd PCR series	2500	11	9
	1250	11	4
	625	22	7
total	170500 molecules (=2046 ng)		41
donor 87			
	input	number of reactions	putative crossover positives
pilot PCR	9600	10	10
	4800	9	9
	2400	10	5
	1200	9	3
2nd PCR series	3960	19	16
	1980	20	12
	990	18	4
total	306660 molecules (=3680 ng)		59

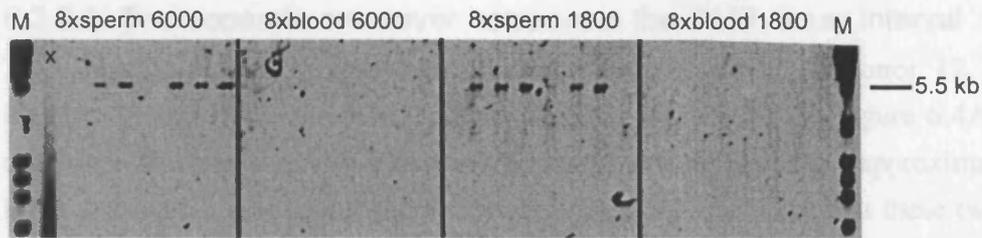


Figure 6.3. An example of secondary crossover PCR products in donor 12 (pilot PCR, orientation A), after 22 cycles of primary PCR, and 24 cycles of secondary PCR. There are eight parallel reactions at each input pool size. Crossover positive reactions were easily distinguishable (e.g. five positives in the eight sperm DNA PCRs at 6000 molecule input). Occasionally, PCR products showed a smear, with no visible band (see lane marked with an X). Such reactions were deemed as failed PCRs and were not included in any further analyses. Blood PCRs were seeded with the same quantity of input DNA as sperm PCRs. Both blood and sperm DNA was digested with *XmnI* prior to PCR amplification. M, marker (*HindIII* digested λ DNA plus *HaeIII* digested ϕ x DNA).

Crossover assays in the two orientations in donor 12 yielded indistinguishable estimates of crossover rate: 0.28×10^{-3} ($0.21 - 0.36 \times 10^{-3}$ 95% confidence intervals) in orientation A and 0.28×10^{-3} ($0.21 - 0.37 \times 10^{-3}$ 95% confidence intervals) in orientation B. This means that approximately one in ~ 3600 sperm DNA molecules had undergone crossover in the *DMB* assay interval. Crossover rate estimates for the other two donors were somewhat higher: 0.45×10^{-3} ($0.30 - 0.64 \times 10^{-3}$ 95% confidence intervals) for donor 63 and 0.42×10^{-3} ($0.30 - 0.58 \times 10^{-3}$) for donor 87. Nevertheless, the crossover rates are not statistically different between assays.

6.2.2. Crossover distribution

To examine the distribution of crossovers within the assayed interval, crossover-positive reactions and all secondary blood PCR products at the highest input pool sizes were subjected to a third round of PCR amplification, and the PCR products were transferred onto dotblots. The location of crossover breakpoints was then mapped to internal SNP intervals (Table 6.3) by ASO hybridisation. The $(GT)_n$ microsatellite located within the putative hotspot region was found to be polymorphic, with array sizes of 12, 13 and 14 repeat units seen in the re-sequenced individuals. Donors 12 and 63 were both heterozygous for 12 and 13 repeat units, and this microsatellite could therefore be used as an additional marker in mapping crossover breakpoints in these men. The microsatellite was typed using primers R74.9F and R75.0R to amplify a 96 or 98 bp fragment from the positive secondary PCRs under standard PCR conditions. The fragments were resolved on 3% Metaphor agarose gel in the presence of ethidium bromide, with internal lane standard of *HaeII* digested ϕ X marker, which includes fragments of 72 bp and 118 bp in length. The sizes of the amplified fragments containing the microsatellite were scored manually. From the number of crossovers observed in each SNP interval, crossover activity in cM/Mb was calculated (Table 6.3 and Figure 6.4).

6.2.2.1. Two separate crossover hotspots in the *DMB* assay interval

Two separate regions of crossover clustering were observed in donor 12, one between markers J8 and JJK2, the other between markers JJK13 and J5 (Figure 6.4A). These two crossover clusters, *i.e.* two crossover hotspots, are separated by approximately 1 kb of DNA where no crossover breakpoints were seen. Further support that these two clusters are genuine hotspots comes from LD analyses, where two separate regions of LD breakdown ~2 kb apart were observed (see Figure 6.1). The location of the two crossover hotspots, henceforth referred to as *DMB1* and *DMB2*, corresponds to the regions of LD breakdown. 9.6% of all crossovers detected across the assay interval occur in the *DMB1* hotspot (donor 12 crossover data pooled from orientations A and B). In the orientation A assay, some crossovers were observed in the 5' end interval, between the forward allele-specific primer site and the first internal heterozygous SNP site, as indicated in Table 6.3.

To identify all *DMB1* crossovers, even if *DMB2* crossovers were present in the same crossover-positive sperm DNA pool, crossovers falling into the *DMB1* hotspot were selectively amplified. To do this, new internal allele-specific reverse primers for secondary PCR amplification targeted towards SNP JJK18 were designed and optimised. All crossover positive reactions (as identified in the assay across the JJ6 - J7 interval) were re-amplified with allele-specific primers targeted towards SNP sites JJ7 and JJK18, encompassing the *DMB1* hotspot region only. Clustering of crossovers at *DMB1* and *DMB2* can be seen in both orientations in donor 12 (Figure 6.4A). Overall crossover distribution is also very similar, consistent with reciprocal crossover.

Table 6.3. Distribution of crossover breakpoints in the DMB region in donor 12 in orientation A and orientation B, donor 63 and donor 87. SNP intervals where no crossovers were seen were pooled. End intervals where crossovers were observed are shown as grey columns. Where only one crossover was observed for a given interval, no Poisson adjustment was calculated.

intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions

Crossover distribution in donor 12, orientation A

input pool size	number of reactions		intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions													
	positive for crossovers	negative for crossovers	J27	J8	J3	J4	J13	JJK16	JJK3	JJK5	JJK9	JJK4	JJK18	JJK14	GTn	J5
9600	33	2	1+3+4	4+30	0+30	6+26	0+30	1+29	0+30	3+27	0+27	0+27	1+26	9+16	28+7	
6000	5	2	1+6	0+6	0+6	0+6	0+6	0+6	0+6	0+6	0+6	0+6	0+6	1+5	4+3	
4800	24	10	1+33	0+33	2+32	2+30	0+31	2+29	0+30	1+29	0+30	1+29	0+29	5+24	20+14	
2400	17	17	1+33	1+32	0+33	1+32	0+33	1+32	0+33	0+33	0+33	0+33	0+33	0+33	13+21	
1800	6	2	0+8	0+8	0+8	2+6	0+6	0+6	0+6	0+6	0+6	0+6	0+6	1+5	6+2	
1200	1	14	0+15	0+15	0+15	0+15	0+15	0+15	0+15	0+15	0+15	0+15	0+15	1+14	0+15	
600	2	6	0+8	0+8	0+8	1+7	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	1+7	
180	1	7	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	0+8	1+7	
Total crossovers			4	5	2	12	0	4	0	4	0	1	1	17	73	
Poisson-adj. crossovers			4	5.1	2	12.3	0	4.3	0	4.1	0	1	1	20.2	124	
Molecules screened			691400	638600	605000	614600	587000	587000	582200	582200	553400	553400	548800	548800	661400	
inter-SNP distance (bp)			560	316	86	1053	1609	184	68	68	8	25	14	183	948	
cM/Mb			1.1	2.5	3.8	1.9	0	4	0	10.7	0	7.2	13	20.1	19.8	

Crossover distribution in donor 12, orientation B

input pool size	number of reactions		intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions													
	positive for crossovers	negative for crossovers	J8	J3	J4	J13	JJK2	JJK13	JJK18	JJK16	JJK3	JJK5	JJK9	JJK18	JJK14	J5
10000	39	1	2+38	1+39	5+35	1+39	0+40	3+37	0+40	1+39	1+39	2+38	0+40	1+31	38+2	
5000	28	12	0+40	0+40	2+38	0+40	0+40	1+39	0+40	2+38	1+39	2+38	0+40	0+38	25+15	
2500	20	20	0+40	0+40	1+39	0+40	0+40	0+40	0+40	2+38	0+40	2+38	0+40	0+40	15+25	
Total crossovers			2	1	8	1	0	4	0	5	2	6	0	1	78	
Poisson-adj. crossovers			2.1	1	8.4	1	0	4.2	0	5	2.2	3.5	0	1	156	
Molecules screened			700000	700000	700000	700000	700000	700000	700000	700000	700000	700000	700000	610000	700000	
inter-SNP distance (bp)			316	86	1053	648	582	342	34	184	68	86	34	14	1117	
cM/Mb			3.2	3.3	1.1	0.4	0	2.9	0	5.4	14.7	15.2	0	27.5	20	

Crossover distribution in donor 63

input pool size	number of reactions		intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions													
	positive for crossovers	negative for crossovers	J2	J3	J4	JJK5	JJK9	JJK18	JJK14	GTn	JJK8	J5				
6000	10	0	0+10	0+10	3+7	1+6	1+5	0+5	2+3	6+0	5+5					
3000	3	7	0+10	0+10	3+7	0+8	0+8	0+8	0+8	1+8	1+9					
2500	9	2	0+11	0+11	1+10	0+10	0+10	0+10	4+6	5+3	4+7					
1500	3	7	0+10	0+10	1+9	0+10	0+10	0+10	1+9	1+9	0+10					
1250	4	7	0+11	0+11	0+11	1+10	0+10	0+10	1+10	2+9	1+10					
750	5	9	0+14	0+14	0+14	0+14	0+14	0+14	3+11	1+13	2+12					
625	7	15	0+22	0+22	1+21	0+21	0+21	0+21	4+18	2+19	2+20					
Total crossovers			0	0	9	2	1	0	15	18	15					
Poisson-adj. crossovers			0	0	10.2	2	1	0	18.7	26.5	17					
Molecules screened			170500	170500	170500	155400	146400	143400	144000	147400	170500					
inter-SNP distance (bp)			175	86	2917	86	34	14	188	415	532					
cM/Mb			0	0	2.1	19.5	19.7	0	77.3	43.3	18.7					

Crossover distribution in donor 87

input pool size	number of reactions		intervals between heterozygous SNPs, number of crossover positive + crossover negative reactions													
	positive for crossovers	negative for crossovers	J27	J2	J3	J4	JJK12	JJK8	JJK18	J5						
9600	10	0	1+9	0+9	0+9	6+3	0+3	3+0	10+0							
4800	9	0	0+9	0+9	0+9	6+3	0+6	3+4	4+5							
3960	16	3	0+19	0+19	1+18	8+10	2+10	6+8	9+10							
2400	5	5	0+10	0+10	0+10	2+8	0+9	3+6	1+9							
1980	12	8	2+18	0+18	0+18	6+12	0+18	6+14	3+17							
1200	3	6	0+9	0+9	0+9	1+8	0+8	1+7	2+7							
990	4	14	0+18	0+18	0+18	1+17	1+17	1+17	2+16							
Total crossovers			3	0	1	30	3	23	31							
Poisson-adj. crossovers			3.1	0	1	40.8	3.8	31	46							
Molecules screened			204700	293100	293100	289140	189600	208500	306700							
inter-SNP distance (bp)			701	175	86	3268	47	265	569							
cM/Mb			1.4	0	4	4.3	42.6	56.6	26.3							

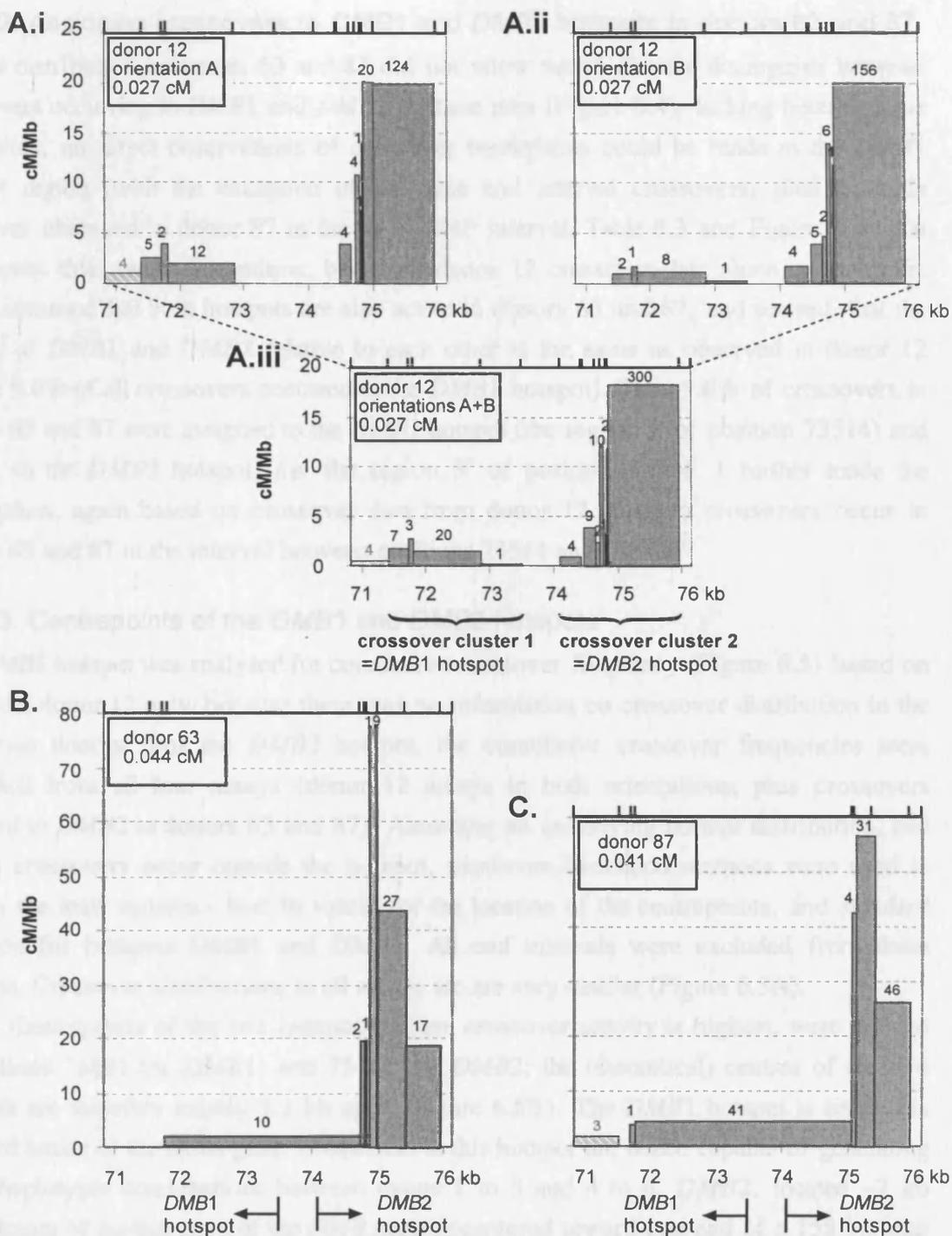


Figure 6.4. Recombination activity in each SNP interval in cM/Mb in donor 12 (A, i-iii), donor 63 (B) and donor 87 (C). Heterozygous SNP sites are depicted as ticks above the graphs. The end interval where crossovers were found are shown as grey hatched bars. The number of crossovers in each SNP interval is shown above the bars. Under the graphs in (B) and (C) the regions for which crossovers were assigned to hotspots *DMB1* and *DMB2*, based on donor 12 crossover data, are shown (see text). Total genetic map distance in cM is shown as an inset in each graph.

6.2.2.2. Assigning crossovers to *DMB1* and *DMB2* hotspots in donors 63 and 87
Marker distribution in donors 63 and 87 did not allow me to directly distinguish between crossovers occurring in *DMB1* and *DMB2* in these men (Figure 6.4); lacking heterozygous SNP sites, no direct observations of crossover breakpoints could be made in the *DMB1* hotspot region (with the exception of the three end interval crossovers, plus a single crossover observed in donor 87 in the J3-J4 SNP interval, Table 6.3 and Figure 6.4). To circumvent this, two assumptions, based on donor 12 crossover data alone, were made: first, I assumed that both hotspots are also active in donors 63 and 87, and second, that the activity of *DMB1* and *DMB2* relative to each other is the same as observed in donor 12 (where 9.6% of all crossovers occurred in the *DMB1* hotspot). Thus 9.6% of crossovers in donors 63 and 87 were assigned to the *DMB1* hotspot (the region 5' of position 73514) and 90.4% to the *DMB2* hotspot, *i.e.* the region 3' of position 73514. I further made the assumption, again based on crossover data from donor 12, that no crossovers occur in donors 63 and 87 in the interval between positions 73514 and 74096.

6.2.2.3. Centrepoinets of the *DMB1* and *DMB2* hotspots

The *DMB1* hotspot was analysed for cumulative crossover frequency (Figure 6.5) based on data from donor 12 only, because there was no information on crossover distribution in the other two donors. For the *DMB2* hotspot, the cumulative crossover frequencies were calculated from all four assays (donor 12 assays in both orientations, plus crossovers assigned to *DMB2* in donors 63 and 87). Assuming an underlying normal distribution, and that no crossovers occur outside the hotspot, maximum-likelihood methods were used to identify the least squares - best fit values for the location of the centrepoinets, and standard deviation for hotspots *DMB1* and *DMB2*. All end intervals were excluded from these analyses. Crossover distributions in all assays are are very similar (Figure 6.5A).

Centrepoinets of the two hotspots, where crossover activity is highest, were defined at positions 71881 for *DMB1*, and 75181 for *DMB2*; the (theoretical) centres of the two hotspots are therefore located 3.3 kb apart (Figure 6.5B). The *DMB1* hotspot is centred in the third intron of the *DMB* gene. Crossovers at this hotspot are hence capable of generating novel haplotypic combinations between exons 1 to 3 and 4 to 6. *DMB2*, located ~2 kb downstream of the last exon of the *DMB* gene, is centered toward one end of a 155 bp long polypurine tract (92% A/G content); the GT microsatellite is also in the immediate vicinity. Standard deviations for crossover distributions were 190 bp for *DMB1*, and 318 bp for *DMB2*. Hence, hotspot widths, within which 95% of crossovers occur, are ~750 bp and ~1250 bp for *DMB1* and *DMB2*, respectively. The width of the *DMB1* hotspot, however, is an underestimate, because the assay interval does not completely capture the 5' end of the crossover distribution (the left "shoulder" of the normal distribution curve, Figure 6.5B). This is, to a lesser extent, also true for the 3' end of the *DMB2* hotspot.

As already noted from the crossover distribution in donor 12 (Figure 6.4), crossover activity was much higher at *DMB2* than *DMB1*. This is also evident from the peak activities in cM/Mb, as revealed by the normal distribution curves in Figure 6.5B: peak activity at *DMB1* (~5cM/Mb) is nine times lower than that of *DMB2* (~45cM/Mb).

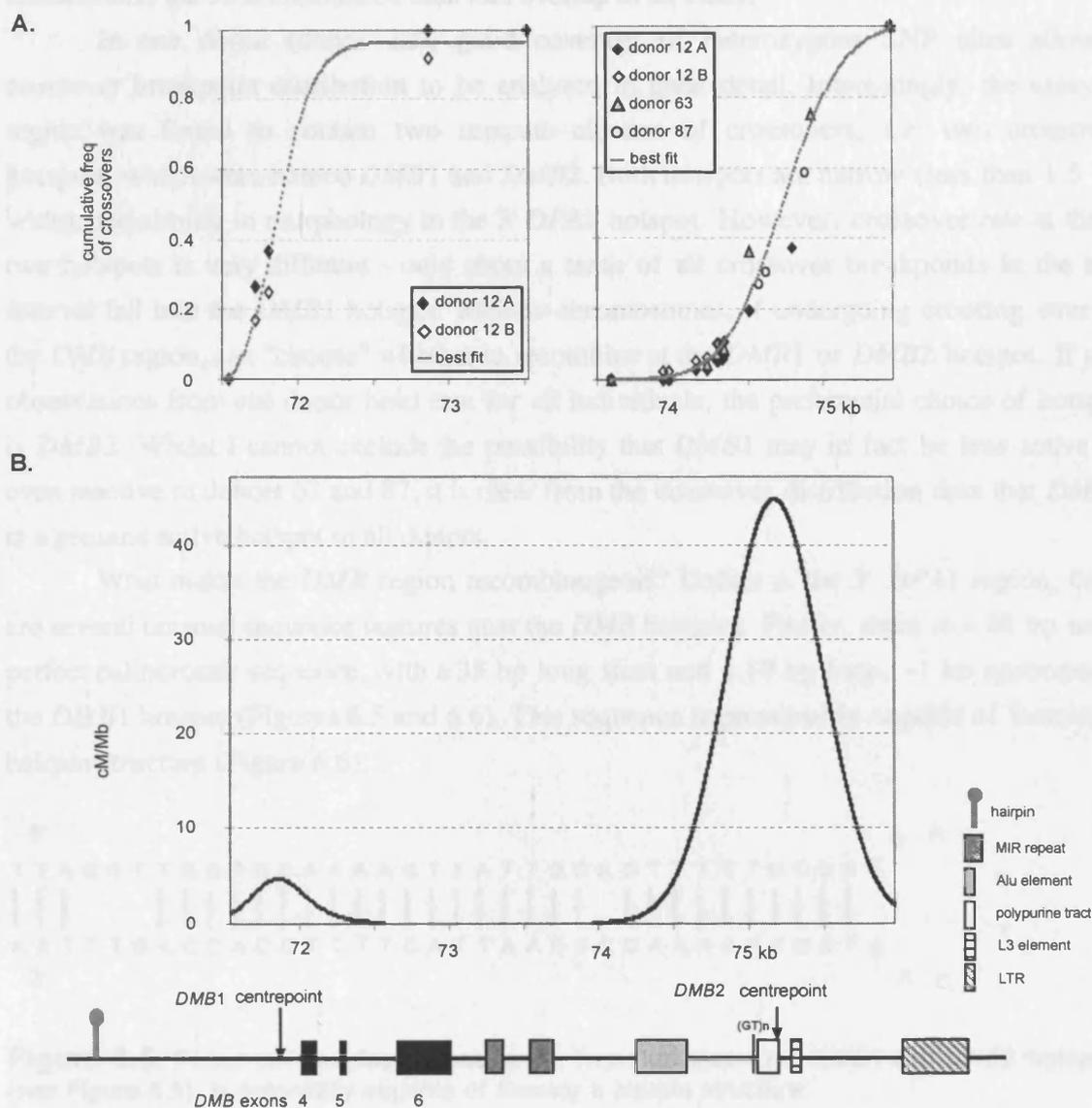


Figure 6.5. (A) Cumulative frequency of crossovers across the *DMB1* (left) and *DMB2* hotspots (right), and the least squares - best fit curve for the observed data. (B) Normal distribution curve underlying the best-fit curve of the data on cumulative frequency of crossovers. The genomic environment is shown at the bottom; black rectangles correspond to exons of the *DMB* gene. The "hairpin" (see discussion) is a ~80 bp long near-perfect palindromic sequence.

6.3. DISCUSSION

A ~5.5 kb DNA segment around the 5' end of the *DMB* gene, where LD analyses revealed two well-defined regions of LD breakdown approximately 2 kb apart, was assayed for sperm crossover activity in three different men. All three donors showed relatively high crossover rates across the test interval, with on average one recombinant per ~3000 sperm

DNA molecules. This rate is very similar to that observed at the 3' *DPA1* hotspot (Chapter 5). The moderate variation in crossover rate between donors ($0.28 - 0.45 \times 10^{-3}$) does not necessarily reflect a true biological phenomenon; it is within the limits of experimental variation (for example, PCR primer efficiency, see also discussion in Chapter 5), and furthermore, the 95% confidence intervals overlap in all cases.

In one donor (donor 12), good coverage of heterozygous SNP sites allowed crossover breakpoint distribution to be analysed in great detail. Interestingly, the assayed region was found to contain two separate clusters of crossovers, *i.e.* two crossover hotspots, which were named *DMB1* and *DMB2*. Both hotspots are narrow (less than 1.5 kb wide), and similar in morphology to the 3' *DPA1* hotspot. However, crossover rate at these two hotspots is very different - only about a tenth of all crossover breakpoints in the test interval fall into the *DMB1* hotspot. Meiotic chromosomes, if undergoing crossing over in the *DMB* region, can "choose" whether to recombine at the *DMB1* or *DMB2* hotspot. If our observations from one donor hold true for all individuals, the preferential choice of hotspot is *DMB2*. Whilst I cannot exclude the possibility that *DMB1* may in fact be less active or even inactive in donors 63 and 87, it is clear from the crossover distribution data that *DMB2* is a genuine active hotspot in all donors.

What makes the *DMB* region recombinogenic? Unlike in the 3' *DPA1* region, there are several unusual sequence features near the *DMB* hotspots. Firstly, there is a 80 bp near-perfect palindromic sequence, with a 35 bp long stem and a 10 bp loop, ~1 kb upstream of the *DMB1* hotspot (Figures 6.5 and 6.6). This sequence is presumably capable of forming a hairpin structure (Figure 6.6).

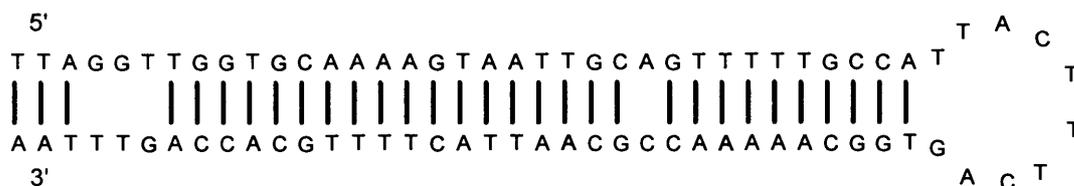


Figure 6.6. Palindromic nucleotide sequence, found upstream the *DMB1* and *DMB2* hotspots (see Figure 6.5), is potentially capable of forming a hairpin structure.

Indirect *in vitro* evidence for secondary structure was found when the region was re-sequenced - all sequencing traces stopped just before the palindromic sequence. Hairpins can extrude to form cruciform DNA structures (Dai *et al.*, 1997); in yeast, palindromic sequences (inverted repeats) have been shown to stimulate meiotic (Nag and Kurst, 1997) and mitotic (Farah *et al.*, 2002) recombination by enhancing DSB formation. The *DMB* palindrome could be a DSB site; however, because it is ~1 and ~4 kb away from the *DMB1* and *DMB2* hotspots respectively, this would mean that the crossover initiation site would be relatively distant from crossover resolution sites (where we observe crossover breakpoints). Gene conversion events, however, have been observed at and immediately near the centre of

recombination hotspots (Jeffreys and May, 2004, see also Chapter 8), indicating that recombination events are initiated and resolved in the same localised region. It is therefore more likely that, if the palindromic sequence plays any role in stimulating recombination, it would be by loosening the local chromatin structure through cruciform conformation, thus making this DNA domain more accessible. Interestingly, Cruciani *et al.* (2003) found a 144 bp palindromic sequence near the two familial crossover breakpoints they observed in the human adenosine deaminase (*ADA*) gene; this was the only long inverted repeat (>50 bp) in the 1 Mb region around *ADA*. The significance of these palindromes for recombination activity remains unknown.

The (GT)_n microsatellite is located only ~40 bp away from centre of the *DMB2* hotspot. Majewski and Ott (2000) found a strong positive correlation of GT repeats and recombination activity on a large scale on chromosome 22. A correlation between recombination rate and GT repeats ≥24 bp long has also been reported in the MHC class II region (Cullen *et al.*, 2002). From these studies, it is not obvious whether the GT tracts are the cause or effect of recombination. In mammalian cell culture cells, repair of artificially introduced double-strand breaks (*via* the non-homologous end-joining pathway) sometimes results in the insertion of (GT)_n repeats (Liang *et al.*, 1998). GT repeats in this case are by-products of recombination activity, rather than the cause of it. On the other hand, there is also evidence to suggest that GT repeats stimulate crossover, possibly because they show a high affinity to binding Rad51 (Biet *et al.*, 1999). When 39 GT repeats were inserted into the yeast *ARG4* recombination hotspot, the overall recombination frequency was not altered but a relative increase in crossovers, as opposed to conversion without crossover, was found (Gendrel *et al.*, 2000). The frequency of double crossovers within a ~5 kb region also increased in the presence of the microsatellite (Gendrel *et al.*, 2000). This is interesting in the context of *DMB*, where two crossover hotspots are located next to each other, and theoretically provide the opportunity for double crossover within the space of a few kilobases. It should be noted, however, that even if GT repeats somehow were able to completely over-ride crossover interference, the rate of double crossovers would still be very low (crossover frequencies at *DMB1* and *DMB2* multiplied, 0.00003 x 0.0003 = 0.00000009, *i.e.* approximately 1 in 10⁸ sperm would undergo double crossover at both *DMB1* and *DMB2*).

Another intriguing sequence feature is the ~150 bp long polypurine tract within which the *DMB2* hotspot centre is located. Such sequences can form unusual intramolecular structures (triplex DNA), in which the double-stranded DNA can fold back on itself, leaving one strand exposed. In the case of a 300 bp polypurine tract in the human epithelial sodium channel gamma subunit, the exposed (unpaired) strand was shown to be sensitive to S1 nuclease digestion (Auerbach *et al.*, 2000). The polypurine tract in the *DMB* region is found where crossover activity is highest. If the *DMB* polypurine tract is capable of forming triplex

DNA (Figure 6.7), this may result in a localised "openness" of DNA (for which S1 nuclease sensitivity is an indicator), making it susceptible to DSBs. Auerbach *et al.* (2000) reported that their 300 bp polypurine tract contained three different S1 sensitive sites, located within a 130 bp zone. If the same is true for the *DMB* region, the *DMB* polypurine tract could be a localised zone where DSBs are made (and hence crossovers and conversions are initiated). The fact that we observe crossover resolution sites around the same region would suggest that all steps of recombination - initiation, branch migration and resolution - are confined to a very narrow (less than 2 kb long) genomic region. If this is true, then the crossover breakpoints observed at the *DMB1* hotspot should also be initiated locally, *i.e.* the *DMB1* hotspot region should contain a separate DSB site. The other scenario is that *all* crossovers seen in the *DMB* region are initiated in the *DMB2* region, and while most are resolved within the *DMB2* hotspot, a fraction of crossovers migrate to the *DMB1* hotspot region where they are resolved. To explain the lack of crossovers between *DMB1* and *DMB2*, this model would also involve suppression of resolution in the inter-hotspot region; this suppression could be related to exon function (*DMB* exons 4 to 6 are located within the "gap" where no crossovers were seen).

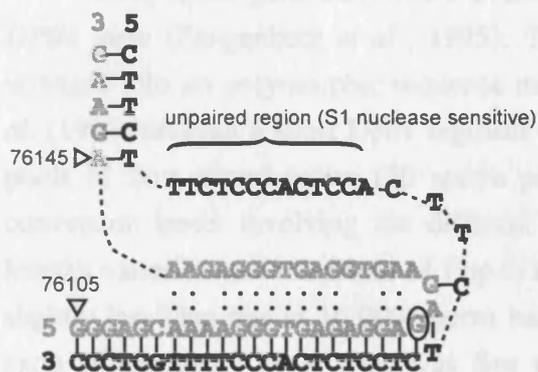


Figure 6.7. A model of how imperfect inverted repeats contained within the *DMB* polypurine tract could form intra-molecular triplex DNA. One possible inverted repeat of several contained within the polypurine tract is shown. The top strand (grey) is in the same orientation as our reference sequence; positions of bases in the reference sequence are indicated with white arrowheads. Watson and Crick base pairs are shown as solid lines, while Höögsteen base pairing is indicated with two points. The position of SNP JJK6G/A is circled; note that the presence of the A allele at this site would create an additional possibility for Höögsteen base pairing.

To summarise, I have located and analysed two novel sperm crossover hotspots (*DMB1* and *DMB2*) in the *DMB* gene region at high resolution. LD breakdown was a remarkably good predictor for the location of these hotspots, which lie approximately 3 kb away from each other. The two hotspots are very similar in morphology, but differ in crossover rate, which is roughly an order of magnitude lower for *DMB1* than *DMB2*. Three unusual sequence features (a hairpin, a GT repeat and a polypurine tract), with possible roles in promoting recombination, were identified in the vicinity of the two *DMB* hotspots.

7. CROSSOVER ANALYSIS AT THE *DPB1* GENE

7.1. INTRODUCTION

Analysis of LD patterns in the *COL11A2-DOA* interval revealed a region of incomplete and equivocal LD breakdown around the first intron of the *DPB1* gene, approximately at position -70000 (see Chapter 4). This region is defined by the partially overlapping 5' end of LD block 1, and 3' end of LD block 2, suggestive of some degree of recombination activity in this genomic interval.

The *DPB1* gene is one of the three extremely polymorphic HLA class II loci, along with *DRB1* and *DQB1*. At these loci, the second exons, which encode part of the peptide-binding groove of the class II molecule, contain a high number of coding SNPs. One-hundred and six different *DPB1* alleles had been described by July 2003 (<http://www.anthonynolan.org.uk/HIG/lists/class2list.html>). The *DPB1* gene encodes the beta glycopeptide chain of the heterodimer that forms the DP cell surface antigen, and the extremely high level of nucleotide variation at non-synonymous sites shows that the locus is under positive selection (see Li, 1997).

Inter-allelic gene conversion events have been reported in the second exon of the *DPB1* gene (Zangenberg *et al.*, 1995). This exon has a high number of coding SNPs, arranged into six polymorphic sequence motifs (regions A-F, see Figure 7.1). Zangenberg *et al.* (1995) assayed a short DNA segment (<300 bp) by performing allele-specific PCR on pools of flow-sorted sperm (50 sperm per pool) from heterozygous men. Putative gene conversion tracts involving the different motifs were in part non-overlapping and their lengths varied from a minimum of 1 bp to a maximum of 132 bp. The authors estimated that slightly less than one in 10,000 sperm had undergone gene conversion within the second exon of the *DPB1* gene. This was first experimental evidence in humans of novel HLA alleles being generated through inter-allelic gene conversion, but should be interpreted with caution, as seven out of nine events involved single site conversions. This gene conversion data across a very short region in the second exon of *DPB1* also suggested that this region is active in recombination, even though located *within* the LD block we detected in UK North Europeans (Chapter 4 and Figure 7.2). There was, however, no information on the relationship between these rare putative conversion events to crossovers (note that, as shown in Figure 7.1, some of the postulated conversions could equally well be crossovers).

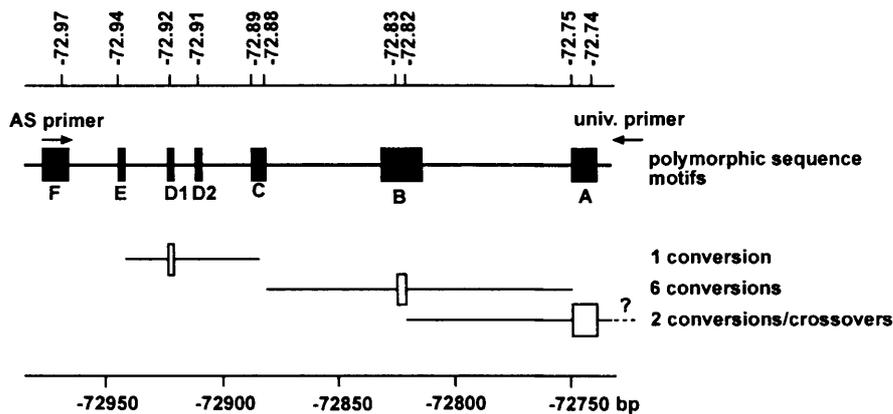


Figure 7.1. Polymorphic sequence motifs within exon 2 of *DPB1*, and Zangenberg and co-workers' assay (1995) to detect gene conversions. Polymorphic *DPB1* sequence motifs are shown as black boxes (A-F). One *DPB1* haplotype is selectively amplified from sperm DNA in a heterozygous man, using an allele-specific (AS) primer and a universal primer (black arrows). Allele-specific oligonucleotides are used to screen for the presence of alleles from the other (non-amplified) haplotype. Three different types of conversion events were identified, involving exchanges in motifs D, B and A, respectively. For exchanges in motif A (the most 3' end of the region screened) it was impossible to distinguish between conversion and crossover. Minimum conversion tract length is shown as a white box and the maximum tract length as solid line extending from the box. Modified from Zangenberg *et al.* (1995).

This work

To investigate crossover activity in a region where LD patterns point to some degree of crossover activity but which lacks discrete LD block structure, a sperm crossover assay was designed to cover a total of roughly 4 kb, including the second exon of the *DPB1* gene. I show below that crossovers are rare in this region, and that the *DPB1* gene region does not constitute a recombination hotspot.

7.2. RESULTS

The crossover assays across the *DPB1* region (Figure 7.2) were designed to cover the region of (incomplete) LD breakdown, and the short region previously assayed for gene conversions by Zangenberg *et al.* (1995). Assay design was also dictated by the presence of a sufficient number of heterozygous SNPs. Recombinant DNA molecules were selectively amplified from two semen donors (donor 9 and 13); full details of primers and PCR conditions can be found in Chapter 2.

The assay design was hampered by a paucity of high frequency SNPs, particularly at the 5' end of the region. When selecting semen donors for crossover assays, genotypes from the 50 donors (see Chapter 3) were examined. Men with multiple heterozygous SNP sites both flanking and across the assay interval are best suited. In addition, for PCR optimisation reactions of allele-specific primers, donors homozygous for both allelic states of the SNP sites required. For example, when optimising forward primer -74.16FA (specific for the A allele of SNP site -74.16A/C), DNAs from individuals homozygous for

both A and C alleles are used as template in PCR, to determine at which annealing temperature the primer shows good efficiency, in combination with specificity for the amplification of the haplotype carrying the A allele. The lack of homozygotes for both allelic states was a problem in the *DPB1* assay; for example, at SNP site -74.16A/C only A homozygotes and A/C heterozygotes were found in the 50 donors. Therefore, to carry out optimisation PCRs for allele-specific primers -74.16FA and -74.16FC, -70.18RA and -70.18RG plus -70.11RC and -70.11RT, haplotypes from a heterozygous donor had to be separated by allele-specific PCR. Diluted haplotype-specific PCR product from this heterozygous donor was then used as template for allele-specific PCR optimisations. Also, due to lack of high frequency SNPs, most allele-specific primer sites were different in the two donors assayed (Figure 7.2).

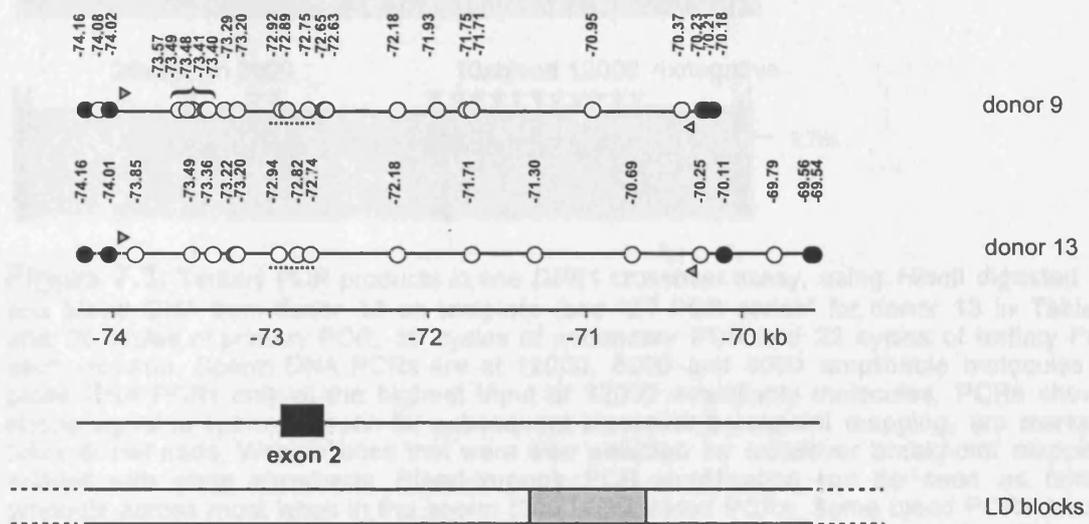


Figure 7.2. Allele-specific primer sites (black circles) and other heterozygous SNP sites (white circles) within the crossover assay interval in the two semen donors. Locations of universal primers used for tertiary PCR amplification are indicated with grey arrowheads. Dashed lines show Zangenberg *et al.*'s (1995) conversion assay interval. LD blocks are depicted as white bars at the bottom; the shaded grey area is where LD blocks 1 and 2 overlap (see Chapter 4).

7.2.1. Recovery of putative crossover positive reactions

To selectively amplify DNA molecules that had undergone crossover, sperm DNA was subjected to two rounds of allele-specific PCR amplification, after which crossover molecules should be detectable on an ethidium bromide stained agarose gel, if allele-specific primers have both good efficiency and specificity. Blood DNA was used as a control for PCR artefacts and for meiosis-specificity of crossovers. In the *DPB1* assay, after secondary PCR amplification no visible product was detectable on an ethidium bromide stained agarose gel in any reactions. Therefore, to score crossover positives, all secondary PCR products were subjected to a third round of PCR amplification with universal primers nested

immediately inside the internal allele-specific primer sites (tertiary PCR, see grey arrowheads in Figure 7.2).

For both donors, three crossover PCR experiments were conducted - a pilot PCR series to get an estimate of crossover rate and thus suitable input pool sizes, then a second and third PCR series to harvest more crossover positive reactions (Table 7.1). After each experiment, the amplification products were examined for the number of putative crossover-positive reactions, as well as the amount of bleed-through amplification. If necessary, minor adjustments were made to annealing temperature and cycle numbers in the allele-specific PCR conditions. Blood PCRs were performed only at the highest input pool sizes.

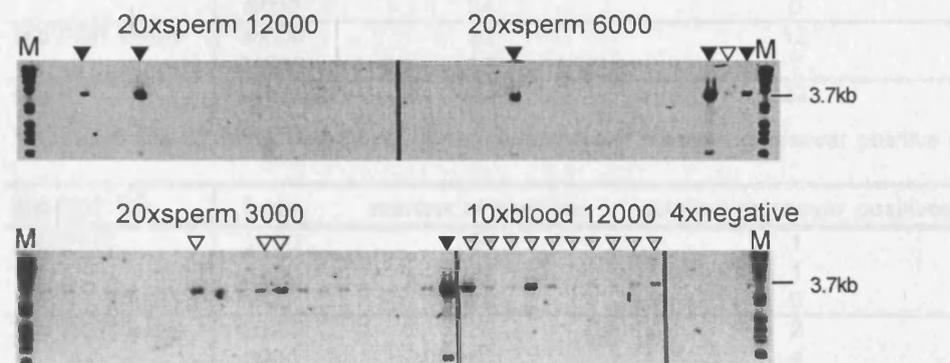


Figure 7.3. Tertiary PCR products in one *DPB1* crossover assay, using *HincII* digested sperm and blood DNA from donor 13 as template (see "2nd PCR series" for donor 13 in Table 7.1), after 26 cycles of primary PCR, 30 cycles of secondary PCR and 22 cycles of tertiary PCR on each reaction. Sperm DNA PCRs are at 12000, 6000 and 3000 amplifiable molecules each, blood DNA PCRs only at the highest input of 12000 amplifiable molecules. PCRs showing a strong signal in sperm, chosen for subsequent crossover breakpoint mapping, are marked with black arrowheads. Weaker ones that were also selected for crossover breakpoint mapping are marked with white arrowheads. Bleed-through PCR amplification can be seen as faint PCR products across most lanes in the sperm (3000) and blood PCRs. Some blood PCRs also show stronger signals (first and fourth lane of blood PCRs). To investigate the nature of these PCR products originating from blood DNA, all ten reactions (grey arrowheads) were transferred onto dotblots for crossover breakpoint mapping. M, marker (*HindIII* digested λ DNA plus *HaeIII* digested ϕ x DNA).

It became clear that the crossover rate in the assayed interval was very low and that very high numbers of amplifiable molecules were required to detect putative crossover positives in the sperm PCRs - a total of $\sim 36\mu\text{g}$ of sperm DNA was screened for donor 9, and $\sim 6\mu\text{g}$ for donor 13. Furthermore, crossover assays were complicated by bleed-through amplification of progenitor haplotypes, as shown by the PCR products generated from blood DNA (Figure 7.3).

From each experiment, all reactions showing a strong signal on an ethidium bromide stained gel, plus a number of reactions showing a weaker signal were transferred onto dotblots (Figure 7.3 and Table 7.1). All blood PCRs were also transferred to dotblots.

Table 7.1. DNA input (number of amplifiable molecules) in each experiment in donor 9 (A) and donor 13 (B). Putative crossovers recovered from each input pool (see text and Figure 7.3) were transferred onto dotblots for crossover breakpoint mapping.

donor 9	input	number of reactions	putative crossover positives
pilot PCR	20000*	15*	-
	10000*	15*	-
	5000	15	0
	2500	15	1
2nd PCR series	12000	36	2
	6000	34	0
3rd PCR series	36000	36	12
	24000	39	7
total	2980500 molecules (=35766 ng)		22

*excluded due to heavy bleed-through that would mask potential crossover positive reactions

donor 13	input	number of reactions	putative crossover positives
pilot PCR	2400	24	1
	1200	24	1
	600	24	0
2nd PCR series	12000	20	2
	6000	20	4
	3000	20	4
3rd PCR series	2500	30	11
total	595800 molecules (=7150 ng)		23

7.2.2. Rate and distribution of putative crossovers

Because of bleed-through of progenitor haplotypes, also seen in blood PCRs (see Figure 7.3), it was impossible to distinguish between genuine and artefactual crossover positive reactions amongst the tertiary PCRs. Hence, the location of crossover breakpoints in all putative crossover positives and blood PCRs was mapped to internal SNP intervals by ASO hybridisation of dotblots (Figure 7.4).

7.2.2.1. PCR artefacts map to end intervals

Blood PCRs were examined for evidence of *in vitro* artefacts and/or mitotic recombination. Only those blood PCRs that showed a strong signal after tertiary PCR on an ethidium bromide stained agarose gel (for example, two reactions in blood PCRs in Figure 7.3), gave a signal upon ASO hybridisation. ASO mapping revealed that they contained molecules whose apparent breakpoints localised to end intervals, consistent with bleed-through of one or both progenitor haplotypes. In donor 9 crossover assays, 222 000 amplifiable blood DNA molecules were screened, and five positive reactions (*i.e.* giving signal upon ASO hybridisation) were found. In donor 13 crossover assays, ~150 000 amplifiable blood DNA molecules yielded two positive reactions (Figure 7.4). In other words, on average approximately one blood DNA molecule per 53 000 gave a signal upon ASO hybridisation.

These artefacts presumably result from one or several allele-specific primers not being sufficiently specific, *i.e.* priming off both progenitor haplotypes. Another explanation would be that these "positives" are not PCR artefacts, but in fact genuine mitotic recombinants - however, during the extensive work in our laboratory on a number of meiotic crossover hotspots we have not found evidence for genuine mitotic crossovers (see e.g. Jeffreys *et al.*, 1998a, Jeffreys *et al.*, 2000, Jeffreys *et al.*, 2001, May *et al.*, 2002).

Similar molecules were also observed in sperm (Figures 7.4), at a similar rate to blood PCRs. For example, in donor 13 crossover assays, 11 pools positive for such molecules were found amongst ~600 000 sperm DNA molecules, giving a rate of 1/55000 for these putative artefacts. It is also interesting that when tertiary PCR was performed with a universal forward primer and an allele-specific reverse primer (Figure 7.4C, box at bottom), instead of universal forward and reverse primers, only six out of the eleven of the previously positive reactions amplified. This indicates there is a heterogenous population of bleedthrough molecules. For one sperm PCR (dot number 6 in Figure 7.4), the breakpoint maps between the internal allele-specific primer site and the first 5' SNP.

7.2.2.2. Crossovers mapping to internal SNP intervals

Only a small number of the sperm PCRs, which were initially deemed putative crossover positive reactions (Table 7.1), were actually found to have their crossover breakpoints in any internal SNP intervals. These are presumably genuine crossovers, since they are observed only in sperm PCRs; for example, in Figure 7.4 dot number 1 shows a clean exchange between SNPs sites -72.74 and -72.18. Altogether twelve apparently genuine crossovers were found in donor 9, and five in donor 13 (Table 7.2).

I cannot, however, exclude the possibility that these molecules, too, may be PCR artefacts, resulting from e.g. *in vitro* recombination ("jumping PCR"). No molecules of this type were observed in blood PCRs, but the number of sperm DNA molecules screened exceeds the number of blood DNA molecules. Only if blood DNA control PCRs were set up in equal numbers and equal input pool sizes to sperm PCRs, and no internal crossover breakpoints were observed, can we formally rule out the possibility that these apparent crossovers are artefacts.

7.2.2.3. Rate estimates of genuine crossovers

To estimate the rate for putative genuine crossovers, end intervals (between the inner allele-specific primer and the adjacent internal heterozygous SNP) were excluded because of evidence for bleed-through amplification, seen in both blood and sperm PCRs (Figures 7.4 and 7.5). Hence, only reactions that contained molecules where the breakpoint maps to an internal SNP interval were considered.

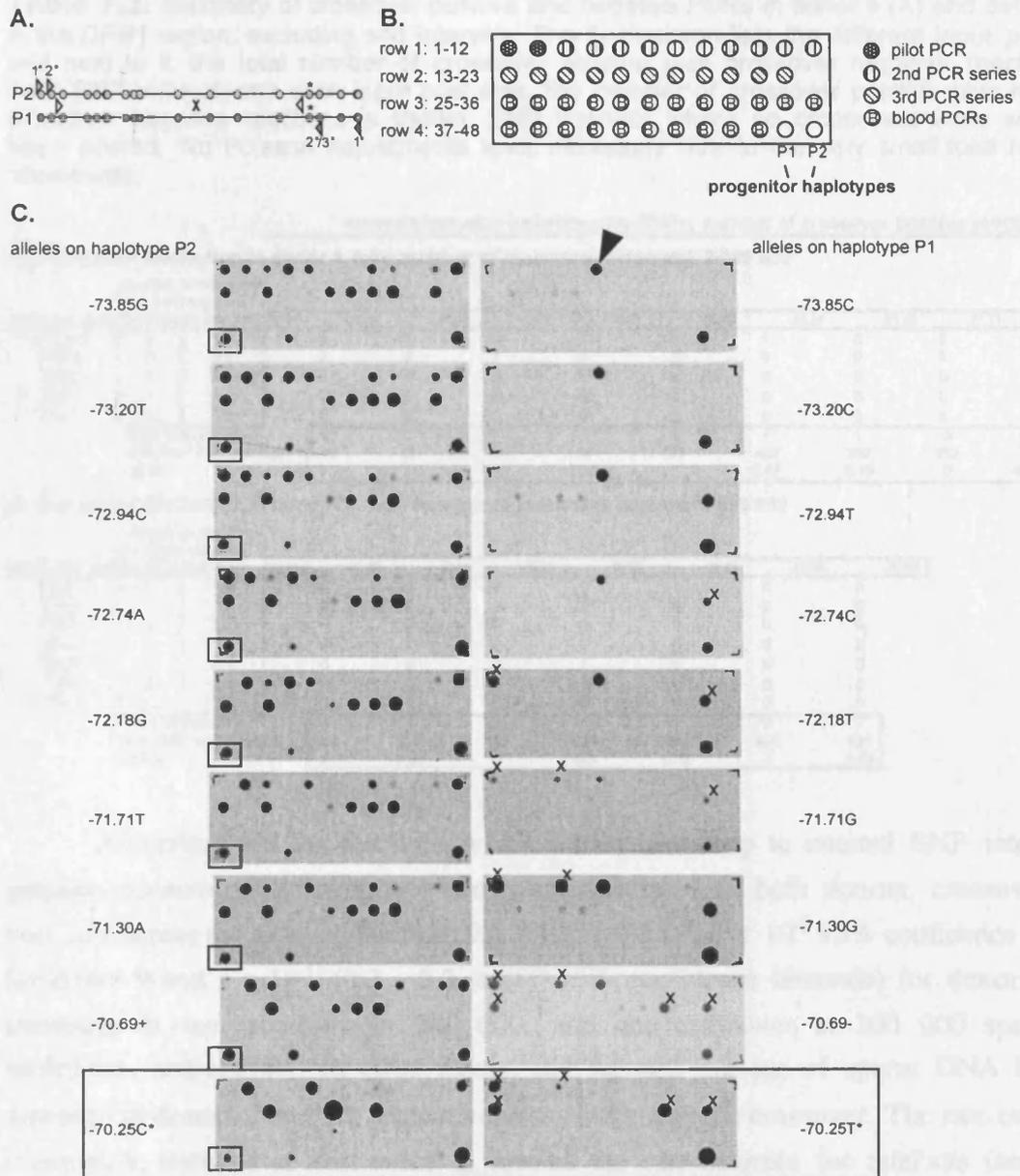


Figure 7.4. Typing of crossover breakpoints in donor 13 in the *DPB1* region. (A) PCR products were generated using allele-specific forward primers for haplotype P2 (light grey arrowheads), and allele-specific reverse primers for haplotype P1 (dark grey arrowheads) in the primary (1°) and secondary (2°) PCRs. Tertiary amplification (3°) was performed using universal primers D-74.0F and D-70.3R (white arrowheads), or in the case of one dotblot, universal forward primer D-74.0F in combination with allele-specific primer -70.11RC specific for the P1 haplotype, in order to include heterozygous SNP site -70.25, marked with an asterisk. (B) Dotblots include putative crossover-positive reactions from the three PCR experiments, as shown on the right and listed in Table 7.1. All blood PCRs were included as controls. (C) Dotblots were probed with first one ASO of a heterozygous SNP site (left), then the other (right). SNP sites are in 5' to 3' order from top to bottom. True crossovers start on the P2 progenitor haplotype and finish on the P1 progenitor haplotype. Five crossovers can be mapped to SNP intervals between -73.85 and -70.25 (marked with X on P1 haplotype blots). One reaction (marked with black arrowhead) maps between the internal allele-specific primer site and the first 5' SNP. The remaining sperm PCRs, along with the two blood PCRs that appear to be positive in the crossover assay (Figure 7.3), are bleed-through of progenitor haplotypes (for example, see blood PCR product marked with a rectangle on haplotype P2 blots).

Table 7.2. Summary of crossover positive and negative PCRs in donor 9 (A) and donor 13 (B) in the *DPB1* region, excluding end intervals. The first column lists the different input pool sizes, and next to it, the total number of crossover positive plus crossover negative reactions. For each SNP interval, and each input pool size, the number of crossover positive plus number of crossover negative reactions is shown. SNP intervals where no crossovers were seen have been pooled. No Poisson adjustments were necessary due to the very small total number of crossovers.

		intervals between heterozygous SNPs, number of crossover positive reactions											
A. Crossover distribution in donor 9, total number of molecules screened: 2 980 500													
input pool size	number of reactions												
	positive for crossovers	negative	-73.57	-72.92	-72.89	-72.75	-72.65	-72.63	-72.18	-71.93	-71.71	-70.95	-70.37
24000	5	35	0	1	0	1	0	1	0	0	0	0	2
12000	2	34	0	0	1	0	0	0	0	0	0	1	0
6000	0	34	0	0	0	0	0	0	0	0	0	0	0
5000	1	14	0	0	0	0	0	0	0	0	0	0	1
2500	0	15	0	0	0	0	0	0	0	0	0	0	0
Total crossovers			0	1	1	1	0	1	1	0	2	5	
inter-SNP distance (bp)			649	32	139	94	23	448	252	213	765	577	
cM/Mb			0	1.05	0.24	0.36	0	0.07	0.13	0	0.09	0.29	
B. Crossover distribution in donor 13, total number of molecules screened: 595 800													
input pool size	number of reactions												
	positive for crossovers	negative	-73.85	-73.20	-72.94	-72.74	-72.18	-71.71	-71.30	-70.69			
12000	1	19	0	0	0	0	1	0	0	0			
6000	1	19	0	0	0	0	0	0	0	0			
3000	0	20	0	0	0	0	0	0	0	0			
2500	3	27	0	1	0	0	0	0	0	2			
2400	1	23	0	0	0	1	0	0	0	0			
1200	0	24	0	0	0	0	0	0	0	0			
600	0	24	0	0	0	0	0	0	0	0			
Total crossovers			0	1	0	1	1	0	2				
inter-SNP distance (bp)			643	258	203	559	466	400	621				
cM/Mb			0	0.65	0	0.3	0.36	0	0.54				

Assuming that the reactions where breakpoints map to internal SNP intervals are genuine crossovers, the crossover rates were calculated. In both donors, crossover rate is very low across the assayed interval: 0.5×10^{-5} ($0.2 - 0.8 \times 10^{-5}$ 95% confidence intervals) for donor 9 and 1×10^{-5} ($0.3 - 2.3 \times 10^{-5}$ 95% confidence intervals) for donor 13. This translates to one crossover in 200 000, and one crossover in 100 000 sperm DNA molecules, respectively. In other words, 2.4 μg and 1.2 μg of sperm DNA had to be screened in donors 9 and 13, respectively, to obtain a single crossover. The rate estimate for crossovers, then, is at least twice as low as the rate estimate for artefacts (see above), showing that the crossover rates in *DPB1* region are at, if not beyond, the limit of detection for our method of crossover analysis.

For each interval, the (very approximate) crossover activity in cM/Mb was calculated (see Table 7.2 and Figure 7.5). There is no evidence for crossover clustering in the manner seen in other segments of the MHC assayed for crossover activity (for example, at the *TAP2* crossover hotspot, Jeffreys *et al.*, 2000, and the 3' *DPA1* and *DMB* hotspots described in the two previous chapters). Instead, as far as detectable within our small sample, crossovers in both donors (if genuine) seem to be distributed randomly across the test interval (Figure 7.5). This observation is supported by the cumulative frequency of crossovers, which shows an apparently linear accumulation with distance (Figure 7.6). The average crossover activity across the *DPB1* interval is 0.12 cM/Mb for donor 9 and 0.26 cM/Mb for donor 13.

No crossovers were seen in the far 5' intervals (649 bp long in donor 9 and 643 bp long in donor 13, when the end interval between the inner allele-specific forward primer site and the first 5' internal SNP site is excluded). If crossovers were fully randomly distributed, I would expect to see two crossovers in donor 9, and one crossover in donor 13 in their respective end intervals. The fact that none were observed is probably due to the small total number of crossovers. Furthermore, the DNA segments where no crossovers were seen are different in the two donors. There is therefore no evidence that these regions are genuinely suppressed for crossovers.

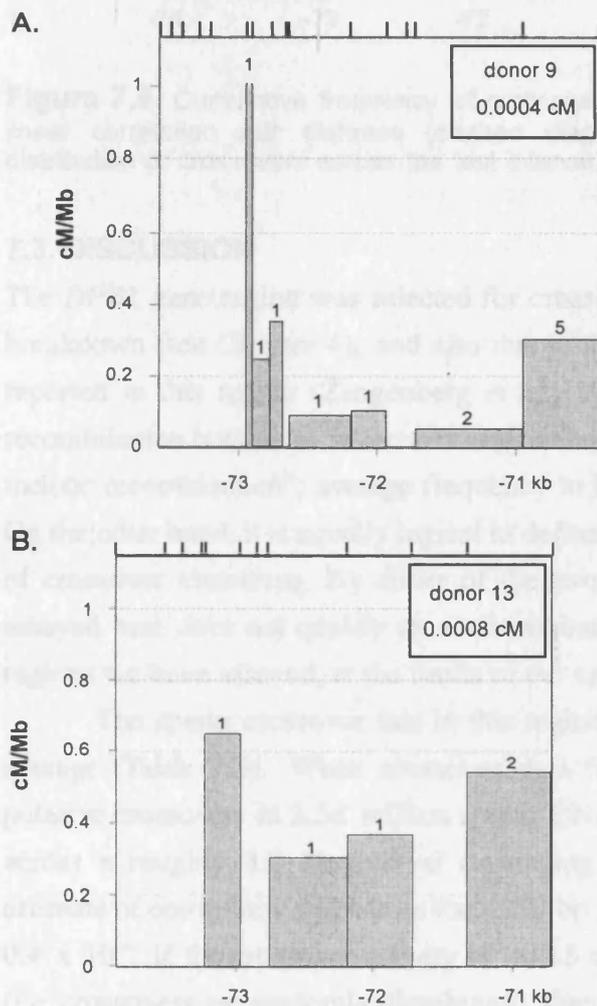


Figure 7.5. Crossover activity in cM/Mb for each SNP interval across the *DPB1* test region in donor 9 (A) and donor 13 (B). The region shown excludes end intervals (region between the inner allele-specific primer sites and first internal heterozygous SNP). Heterozygous SNP sites are depicted as ticks above the graphs. The number of crossovers in each SNP interval is shown above the bars.

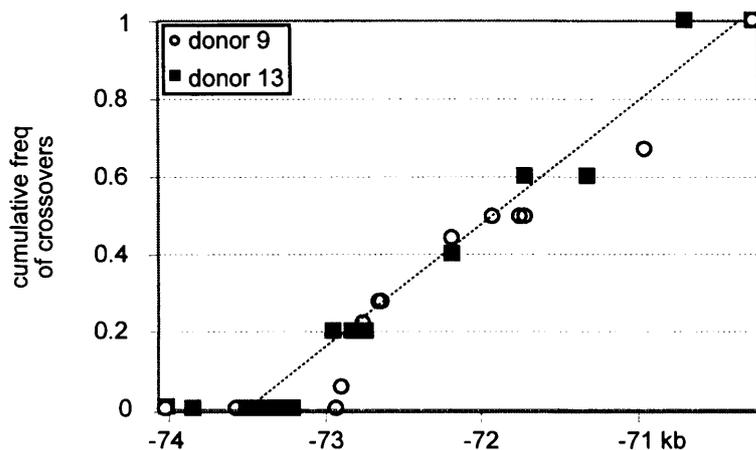


Figure 7.6. Cumulative frequency of crossovers in the *DPB1* region, showing an apparently linear correlation with distance (dashed diagonal line). This is consistent with random distribution of crossovers across the test interval.

7.3. DISCUSSION

The *DPB1* gene region was selected for crossover analysis in part due to the observed LD breakdown (see Chapter 4), and also due to the fact that gene conversion events had been reported in this region (Zangenberg *et al.*, 1995). Lichten and Goldman (1995) define a recombination hotspot as "a locus or region that displays a greater than average frequency of meiotic recombination"; average frequency in humans is generally regarded as ~ 1 cM/Mb. On the other hand, it is equally logical to define a recombination hotspot as a localised region of crossover clustering. By either of the two hotspot definitions, the *DPB1* gene region assayed here does not qualify as a recombination hotspot. It is in fact one of the coldest regions we have assayed, at the limits of our crossover analysis.

The sperm crossover rate in this region is at least 6-fold *below* the human genome average (Table 7.3). When crossover data from both donors are pooled (altogether 17 putative crossovers in 3.58 million sperm DNA molecules), the rate estimate is $\sim 0.5 \times 10^{-5}$ across a roughly 3.2 kb interval (averaging 0.15 cM/Mb). Zangenberg *et al.*'s (1995) estimate of conversion rate across the ~ 250 bp region within the second exon of *DPB1* was 0.4×10^{-4} . If the crossover activity of ~ 0.15 cM/Mb is constant across the DNA segment (*i.e.* crossovers are randomly distributed), then our estimate for Zangenberg's interval is 0.4×10^{-6} . Crossovers in the region would hence be roughly two orders of magnitude less frequent. This estimate is far greater than the 4:1 to 10:1 conversion to crossover ratio estimate by Jeffreys and May (2004), which is based on a large number of conversions mainly in the *DNA3* recombination hotspot. If the two recombinants mapping to the end interval in Zangenberg *et al.*'s assay (see Figure 7.1) are interpreted as crossovers instead of conversions, their crossover rate estimate is >10 times higher than our estimate (Table 7.3). This raises concerns about the authenticity of the *DPB1* conversions previously reported; seven out of nine are single site conversions, at least some of which could result from PCR

misincorporation, and the remaining two, where the supposed conversion involved five nucleotides, fall into the end interval *i.e.* could be crossovers.

Table 7.3. Crossover frequency estimated from this study, and conversion and crossover frequencies calculated from Zangenberg *et al.*'s (1995) data.

	sperm molecules screened	crossovers recovered	cM/Mb
donor 9	2.98 million	12	0.12
donor13	0.60 million	5	0.26
possible crossovers*	0.11 million	2	7.2
			conversion rate
conversions*	0.11 million	7	0.6x10 ⁻⁴

*Zangenberg *et al.*'s (1995) data. Both molecules that are putative crossovers (see Figure 7.1) have been interpreted as crossovers, not conversions.

Despite evidence for LD breakdown within this region, I did not observe clustering of crossovers. Instead, the data show a roughly linear accumulation of crossovers with distance (Figure 7.6). This suggests that the rare crossovers, if genuine, are randomly distributed across the test interval. The crossover rate across the *DPB1* test interval (0.15 cM/Mb) is in fact compatible with the "baseline" rate estimated from levels of LD breakdown in inter-hotspot DNA in the MHC class II region (see Table 4.1).

With such a low recombination rate, why do we still observe a localised breakdown of LD? This may simply be because chromosomes that have undergone "non-hotspot" crossovers, such as the rare events observed in the sperm DNA of donor 9 and 13, have spread in the population by random chance ("lucky" historical crossovers). Also, the region may have contained a recombination hotspot that was active in the past, so that this signature of historical recombination is still visible in the LD structure. This hotspot may have become inactivated, for example by elimination of an recombination initiation signal sequence from the population. It may also be that the crossover assay was designed too far 5', and that the bulk of crossovers cluster in the region further downstream - from the LD patterns, it is not clear where precisely LD breakdown occurs. It is also possible that this region harbours a hotspot that is much more active in individuals with a different haplotype to the two donors assayed here, and that I happened to assay two individuals with a particularly low recombination rate. This is not very likely, as I have already analysed four haplotypes (within the two heterozygotes), and not seen evidence for a crossover hotspot. Furthermore, this region could contain a recombination hotspot that is much more active in females than males, hence breaking down marker associations in the general population, a hypothesis we cannot directly test with our assay. Whatever the cause of the partial LD breakdown around the *DPB1* gene, the data presented in this chapter suggest that the assayed region does not contain a active sperm crossover hotspot. These results illustrate the fact that not all LD breakdown regions should be assumed to arise from "universal" recombination hotspot activity.

8. SUMMARY OF LD AND RECOMBINATION HOTSPOTS IN THE MHC CLASS II REGION

OVERVIEW

The two regions targeted during the course of this work, the *COL11A2-DOA* and *RING3-TAP2* intervals, for which I have presented LD and crossover data in chapters 4 to 7, were investigated as part of a larger ongoing project in our laboratory, focusing on meiotic recombination in the human MHC class II region. Taking all data from our group together, we have now examined almost 300 kb of DNA for LD patterns, and targeted eight regions of LD breakdown for crossover analysis. These assays have revealed the existence of seven highly localised sperm crossover hotspots. This chapter is an introduction to the last results chapter (Chapter 9). In Chapter 9, I describe how I extended the initial LD studies (conducted in the UK North European population) to two additional populations, surveying the effect of population history *vs.* recombination on the behaviour of one MHC LD block. To explain the reasoning behind the population study, I have included this summary chapter, where I briefly present all data available to date on MHC class II LD patterns and recombination hotspots (in part described in chapters 4 to 7, also in Jeffreys *et al.*, 2000, Jeffreys *et al.*, 2001, and Jeffreys and Neumann, 2002). LD across the approximately 292 kb long region is block-like, and narrow regions of breakdown correspond precisely to sperm recombination hotspots. All seven recombination hotspots have similar width (1-2 kb) but variable recombination rate.

8.1. LD BLOCKS IN THE MHC CLASS II REGION

Across the 292-kb long segment of the MHC class II region examined, 481 SNPs (or other bi-allelic polymorphisms) were identified and genotyped in 50 UK North Europeans. The SNP genotypes were used to calculate maximum-likelihood haplotypes, which were subjected to LD analyses, as described in Chapter 4. A block-like LD structure, observable visually from D' LD plots, was discovered (Figure 8.1); for clarity, I have used the same 5' to 3' orientation numbering of LD blocks as in Chapter 4. The overwhelming majority of DNA (271 kb of the total of 292 kb, *i.e.* 93%) is located within these LD blocks.

LD blocks can be divided into short (<5 kb) and long blocks (Table 8.1). Across this ~300 kb region, there are six long blocks and four short blocks. Block boundaries are defined by nine regions of LD breakdown. Sometimes this LD breakdown is highly localised, and it is clear where the upstream block ends and the downstream block starts. An example of such well-defined block boundaries can be found in the *TAP2* region. In other cases, such as in the *DPB1* gene region, LD only breaks down incompletely, and the location of LD breakdown is equivocal. For the two outermost LD blocks, blocks 1 and 8, only one of the two boundaries has been located, and defining them fully would require

extending SNP discovery and LD analyses further upstream and downstream, respectively. For all other blocks, we have defined both the 3' and 5' boundaries (blocks 2 to 7).

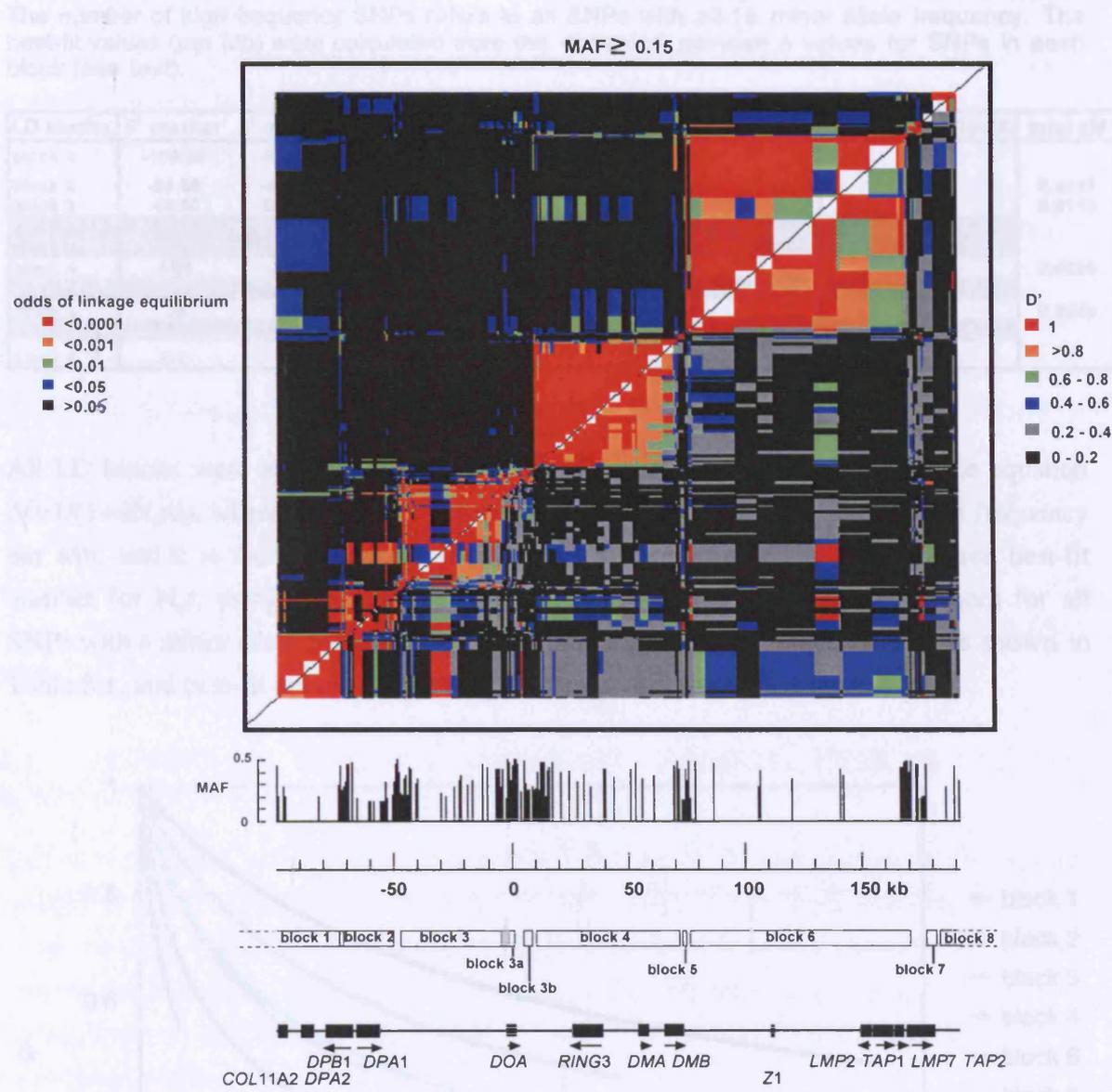


Figure 8.1. Patterns of LD in the *COL11A2-TAP2* interval. Only SNPs with a minor allele frequency ≥ 0.15 have been included. LD blocks, as identified visually, are shown as white boxes, numbered 1-8 in 5' to 3' order. Regions where LD blocks overlap are shaded in grey. Genes (names in top row) and pseudogenes (names in bottom row) are indicated as black boxes. For the first and last block only one boundary has been defined.

The long LD blocks that have been fully defined vary in length, from ~19 kb to ~94 kb (Table 8.1); block 8 was also classified as a long block because it encompasses at least 5.3 kb. The short LD blocks are 2 to 4 kb long.

Figure 8.2. Decay of the r^2 measure of LD with distance within the long LD blocks in the *COL11A2-TAP2* interval. Lines (color-coded as shown on the right) show the total squared correlation curves for the relationship $r^2 = 1/(1 + 4N_e d)$ where N_e is effective population size, c is recombination rate per Mb interval and d is the distance in Mb between markers.

Table 8.1. LD block sizes in the *COL11A2-TAP2* region. For the first and last block (blocks 1 and 8, in grey text) only one boundary has been defined. Blocks shorter than 5 kb are in grey shaded rows. The 5' and 3' SNPs that define each block, along with block size, are shown. The number of high frequency SNPs refers to all SNPs with ≥ 0.15 minor allele frequency. The best-fit values (per Mb) were calculated from the observed pairwise Δ values for SNPs in each block (see text).

LD blocks	5' marker	3' marker	size	high freq SNPs	best fit value for $N_e \times r$	cM/Mb if $N_e=10000$	total cM
block 1	-100.49	-71.30	≥ 29.2 kb	18	31	0.31	
block 2	-68.68	-49.23	19.4 kb	52	58	0.58	0.0113
block 3	-44.50	DD14	40.4 kb	37	28	0.28	0.0113
block 3a	DE10	F9	3.4 kb	18	652	6.5	
block 3b	GA6	A10	3.1 kb	13	163	1.63	
block 4	AB6	JJ7	60.1 kb	35	16	0.16	0.0096
block 5	J13	JJK14	2.0 kb	10	495	4.95	
block 6	J5	T16	94.1 kb	26	9	0.09	0.0085
block 7	TA33	T47	4.0 kb	8	528	5.28	
block 8	U2	U8	≥ 5.3 kb	3	113	1.13	

All LD blocks were analysed for the rate of LD decay with distance. From the equation $\Delta^2 = 1/(1+4N_e r d)$, where N_e is the effective population size, r is the recombination frequency per Mb, and d is the inter-marker distance in Mb, I calculated the least squares best-fit number for $N_e r$, using the observed pairwise Δ values and inter-marker distances for all SNPs with a minor allele frequency ≥ 0.15 within a block. The best-fit values are shown in Table 8.1, and best-fit curves for the long LD blocks are shown in Figure 8.2.

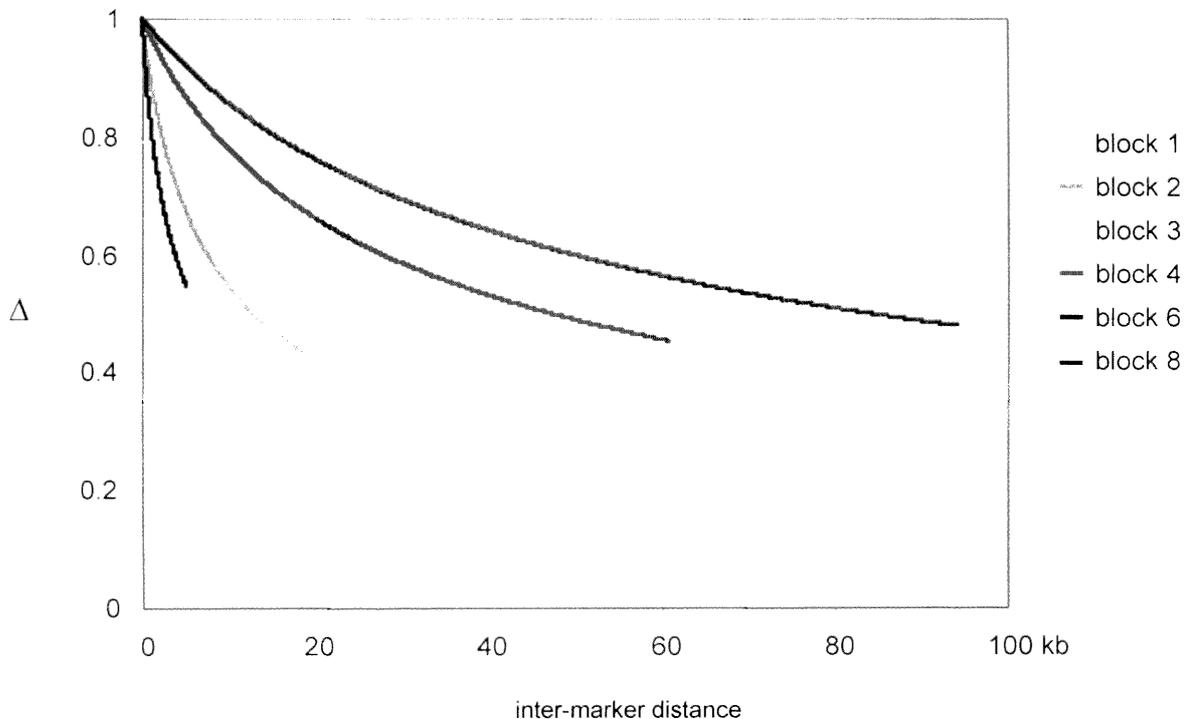


Figure 8.2. Decay of the $|\Delta|$ measure of LD with distance within the long LD blocks in the *COL11A2-TAP2* interval. Lines (colour-coded as shown on the right) show the least squares best-fit curves for the relationship $\Delta^2 = 1/(1+4N_e r d)$ where N_e = effective population size, r = recombination rate per Mb interval and d is the distance in Mb between markers.

Apart from varying in length, blocks also vary in the rate of LD decay with distance. This is in some respect self-evident, because the longer blocks are long specifically as a result of their slower decay of LD with distance. All blocks, regardless of size, show the LD measure of Δ levelling off at very similar values, *i.e.* markers at the end of each block show Δ values between 0.4 and 0.5 (Figure 8.2). This reflects the threshold of visual LD block identification, where on the D' LD plots, green rectangles correspond to inter-marker D' values of 0.6 - 0.8, and blue rectangles to D' values of 0.4 - 0.6 (Figure 8.1). However, there seems to be a genuine tendency of MHC LD blocks to end rather abruptly, at locations where most pairwise D' values are still relatively high ($D' > 0.6$). Another observation is that, regardless of block length, the total genetic map distance contained within each fully defined long LD block is almost identical. If effective population size is assumed to be $N_e = 10000$, the total genetic map distance contained within each of the long LD blocks is ~ 0.01 cM.

Because we want to characterise meiotic recombination hotspots at high resolution, the regions of LD breakdown, indicating historical recombination activity, are of particular interest to us. Altogether nine such regions were identified, in 5' to 3' order: one in the *DPB1* gene region, one 3' of the *DPA1* gene, three within a ~ 15 kb region around the *DOA* gene, two in and near the *DMB* gene, and two in and near the *TAP2* gene.

8.2. CROSSOVER HOTSPOT ANALYSES ACROSS REGIONS OF LD BREAKDOWN

Each of the regions of LD breakdown (apart from the far 3' one, where a complex re-arrangement polymorphism has prevented further analysis) was then analysed for sperm crossover activity. This involved designing allele-specific PCR assays anchored in the DNA flanking the region of LD breakdown on either side, and recovering crossover molecules directly from sperm DNA, as described in chapters 5 -7. Below, I summarise the results of these crossover analyses.

8.2.1. LD breakdown but no crossover hotspot at the *DPB1* gene

The *DPB1* crossover assay (discussed in detail in Chapter 7) was designed to cover the region of LD breakdown between blocks 1 and 2. Crossovers in the *DPB1* region were found to be extremely rare, to the extent that they are at the limit of detection for our PCR method. We concluded that the assayed region most likely does not contain an active sperm crossover hotspot. This result emphasizes the fact that not all LD breakdown regions necessarily arise from recombination hotspot activity.

8.2.2. Crossover hotspots discovered at regions of localised LD breakdown

Other crossover assays designed to cover regions of LD breakdown resulted in the identification of a total of seven sperm crossover hotspots (defined as localised clusters of

crossover breakpoints). These were named, in 5' to 3' order, the 3'*DPA1*, the *DNA1*, the *DNA2*, the *DNA3*, the *DMB1*, the *DMB2* and the *TAP2* hotspots (Figure 8.3). All seven hotspots have similar widths (~1-2 kb) but widely differing crossover rates (Table 8.2). Crossovers at the most active crossover hotspot, *DNA3*, are ~250 times more frequent than at the weakest hotspot, *DNA1*.

At all hotspots, almost all crossovers were simple and showed clean exchanges from one progenitor haplotype to the other, mapping to one SNP interval. The rare crossovers (less than 1%), which in addition to the "normal" exchange showed exchange of additional markers between haplotypes near the site of crossover, presumably result from patchy heteroduplex DNA repair.

Assays for all but one crossover hotspot recovered DNA molecules that are consistent with fully reciprocal crossover products; rate and distribution of crossovers was indistinguishable in both orientations ("A to B" and "B to A"). The exception to this was the *DNA2* hotspot, where in some men, "A to B" crossovers mapped to a different location within the hotspot than "B to A" crossovers (Jeffreys and Neumann, 2002). This phenomenon, termed reciprocal crossover asymmetry, can be explained by a haplotype-specific preference in the initiation of recombination. A single A/G SNP site, located at the centre of *DNA2*, in the heterozygous state seems to be sufficient to generate this asymmetry. The chromosome carrying the A allele is preferentially (in ~80% of crossovers) used for initiation. This allele then, as predicted by the double-strand break repair model, will in 80% of crossover events be lost due to biased gene conversion. Because the crossover rate at *DNA2* is low (Table 8.2), the effect on the population level is weak; nevertheless, over evolutionary time, this initiation-proficient allele will decrease in frequency and ultimately, the *DNA2* hotspot should go extinct (Jeffreys and Neumann, 2002).

There is some evidence for crossover hotspot clustering. Near the *DOA* gene, we found three hotspots within a space of ~12 kb, and in the *DMB* region, two hotspots within just 5 kb (Figure 8.3). Furthermore, LD patterns suggest that the *TAP2* hotspot, too, might be one hotspot in a cluster of two; there is a clear region of LD breakdown between blocks 7 and 8, *i.e.* a putative hotspot region but an unknown re-arrangement polymorphism in this region has prevented further analysis. The spacing between hotspot clusters and "lone" hotspots varies from ~55 to 95 kb (Figure 8.3).

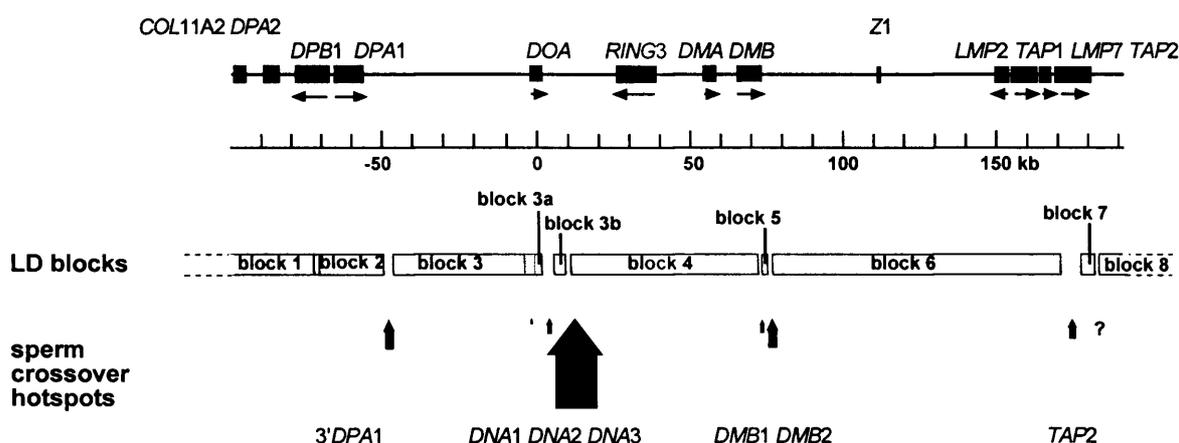


Figure 8.3. Location of sperm crossover hotspots and LD blocks across the 292-kb interval. The size of the block arrows corresponds approximately to the crossover activity at each hotspot.

Table 8.2. Summary of sperm crossover hotspots across the 292-kb interval. The approximate locations of centrepoinets are given, along with the hotspot widths, within which 95% of crossovers occur, and peak crossover activities (as determined by normal-distribution fitting). The mean male genetic map distance contributed by each hotspot was calculated from the observed crossover frequency per sperm across each test interval. There seems to be no common denominator for the genomic environment where hotspots occur.

Crossover hotspot	centrepoinet location	width (kb)	peak activity (cM/Mb)	genetic map distance (cM)	genomic environment	references
3' DPA	-48557	1.6	27	0.028	intergenic, in HERV element	this work
DNA 1	-1976	1.9	0.4	0.0005	promoter	Jeffreys <i>et al.</i> (2001)
DNA 2	2034	1.3	8	0.0037	intergenic, in Alu element	Jeffreys <i>et al.</i> (2001)
DNA 3	10008	1.2	100	0.13	intergenic, in Alu element	Jeffreys <i>et al.</i> (2001)
DMB 1	71881	0.8	5	0.0027	third intron of DMB gene	Jeffreys <i>et al.</i> (2001) + this work
DMB 2	75181	1.3	45	0.026	intergenic, in polypurine tract	Jeffreys <i>et al.</i> (2001) + this work
TAP 2	170709	1	8	0.0058	second intron of TAP 2 gene	Jeffreys <i>et al.</i> (2000)

CONCLUDING REMARKS

In the *COL11A2* to *TAP2* segment of the MHC class II, there is remarkably good correlation between LD breakdown and active sperm crossover hotspots; the only exception to this is the *DPB1* region, where LD breakdown (albeit equivocal) but no crossover clustering was observed. Within the LD breakdown regions, we have identified and characterised seven highly localised sperm crossover hotspots. In total, the seven hotspots contribute approximately 0.2 cM to the genetic map distance, of which the most active hotspot, *DNA3*, accounts for 65%. The 22 familial crossovers previously reported in our test interval suggest that the same crossover hotspots are present in females, but may have a higher recombination rate (18 of 22 crossovers were maternal in origin, Cullen *et al.*, 1997).

From LD data, we can estimate genetic distances contributed by the LD blocks (see Table 8.1). Assuming an effective population size of $N_e=10000$, each of the long LD blocks adds approximately 0.01 cM to the map distance, totalling 0.05 cM for the six blocks. The four short LD blocks (shaded in grey in Table 8.1), which we now know to separate closely

spaced hotspots, appear to be more active in recombination, as estimated from LD data: they contribute between 0.005 (block 3a) and 0.022 cM (block 3b) to genetic map distance, totalling 0.058 cM (again assuming $N_e=10000$). Hence, our estimate of the total genetic map distance, based on crossover and LD block data pooled, is 0.31 cM over a physical distance of 292 kb. The average recombination activity, then, is ~ 1.1 cM/Mb. Note that while the sperm crossover portion of our estimate is (obviously) male specific, the LD-derived estimate is not, because it is population-based and our region is autosomal. Therefore our data is in good agreement with the most recent estimate of recombination activity for chromosome 6, which is 0.72 cM/Mb for males, and 1.27 cM/Mb for females (Kong *et al.*, 2002). However, our results highlight how misleading the term "average recombination activity" can be, as regional variation is enormous, ranging from 0.13 cM across a 5.5 kb interval (the *DNA3* hotspot) to 0.009 cM across a 94 kb long DNA segment (LD block 6).

The seven hotspots are all ~ 1 -2 kb wide, but show extreme variation in crossover activity (a range of 0.4 to 100 cM/Mb in peak activity). Crossovers are simple and reciprocal, with the exception of *DNA2* which shows allele-specific preference in crossover initiation. There is no obvious similarity in the primary DNA sequence or genomic environment in which hotspots are found, suggesting a role for higher order chromatin structure; the occurrence of hotspot clusters may indicate a local "openness" of chromatin, which may make these regions accessible to the recombination machinery.

The question remains whether the LD blocks detected in UK North Europeans may harbour "hidden" recombination hotspots, which we would not have assayed for. Familial crossover data (Cullen *et al.*, 1997) suggest that we have at least not missed any hotspots with high recombination activity. It is possible, however, that population processes would have erased the signature marks of weaker crossover hotspots from LD block structure in UK North Europeans. In the following chapter, I investigate this issue further by examining the LD block between the *DNA3* and *DMB* hotspots in two additional populations with different demographic histories.

9. EFFECT OF RECOMBINATION HOTPOTS AND POPULATION HISTORY ON LINKAGE DISEQUILIBRIUM

9.1. INTRODUCTION

Recombination, population processes and selection influence the extent of linkage disequilibrium (LD) in the human genome, but their relative contributions remain unclear. The level of LD between a pair of markers depends on both molecular and population genetic factors. At the molecular level, crossover, gene conversion and recurrent mutation can reduce LD; for the MHC class II region, the effect of crossover activity on LD was shown in chapters 5, 6 and 8, whereas for some "misbehaving" SNPs within LD blocks, I inferred gene conversion and/or recurrent mutation (Chapter 4). At the population genetic level, population size, migration, bottlenecks, admixture and drift can have complex effects (Ardlie *et al.*, 2002). With so many factors contributing to LD, it is hardly surprising that estimates of the extent of LD vary greatly both in different regions of the human genome (see section 4.1) and in different populations.

Population differences in LD levels were pointed out for example in a pioneering study by Laan and Pääbo (1997). The authors showed that LD between microsatellites in Xq13 extends further in the Saami, a small and constant-sized population from the Arctic region of Europe, than in three other north European populations (Laan and Pääbo, 1997). Reich and colleagues (2001) estimated that LD in North Europeans extends on average 60 kb from a SNP with high (≥ 0.35) minor allele frequency, but less far in Nigerians. Lower levels of LD in sub-Saharan Africans were also reported by Lonjou *et al.* (1999) and Gabriel *et al.* (2002). The Evenki from Siberia and the Saami, two small and constant-sized populations, showed higher LD over short (up to 8 kb) distances compared with Finns and Swedes (Kaessmann *et al.*, 2002). In contrast, in the pseudoautosomal *SHOX* gene region, LD declines with physical distance at a similar rate in UK North Europeans, Vlax Roma (Gypsies) from Bulgaria and Saami, despite their very different histories (May *et al.*, 2002). To summarise, most studies comparing the extent of LD in different populations have observed less LD in sub-Saharan Africans, similar levels of LD in most populations of north European ancestry and, on occasion, markedly higher LD in small and isolated populations.

As shown in this thesis (chapters 5, 6 and 8), and other work (May *et al.*, 2002, Smith *et al.*, 1998, Templeton *et al.*, 2000), there is growing evidence that meiotic recombination events in human chromosomes tend to be clustered into hotspots 1-2 kb in width, and that these recombination hotspots can strongly influence patterns of LD. Hotspots tend to be separated by extended blocks of recombinationally suppressed DNA containing markers in strong LD. This structuring of diversity into LD blocks appears to be common in the human genome (Gabriel *et al.*, 2002, Daly *et al.*, 2001), and there is evidence that LD blocks and their underlying haplotypes can be shared by diverse populations (Gabriel *et al.*, 2002). It remains unclear, however, to what extent these LD

blocks are the result of chance clustering of historical crossovers, in the absence of true hotspots. Also, evidence for LD blocks created by recombination hotspots comes solely from studies of North Europeans (Jeffreys *et al.*, 2000, Jeffreys *et al.*, 2001, chapters 4-6 and 8 in this thesis), and it is unclear how population history might influence block structure in the presence of hotspots.

9.1.1. Target LD block between the *DNA3* and *DMB* hotspots

To determine how recombination hotspots and population history affect LD, I chose a region over which we had already characterised fine-scale LD patterns and meiotic crossovers in sperm in UK North Europeans (see chapters 4, 5 and 8, and Jeffreys *et al.*, 2001): the 75-kb long *HLA-DOA/HLA-DMB* interval in the MHC class II region. It contains, centromeric to telomeric, the highly active *DNA3* hotspot, followed by a 60-kb long recombinationally-suppressed LD block (seen in UK North Europeans, block 4 in Chapter 8) that terminates at the *DMB1* and *DMB2* hotspots, which show weak and intermediate male crossover activity respectively. I wanted to investigate the extent of LD in this region in two additional populations chosen for their very different demographic histories, the Saami and the Zimbabweans (see below). The geographical locations of these populations are shown in Figure 9.1.

9.1.2. Populations studied - UK North Europeans, Saami and Zimbabweans

UK North Europeans are genetically a relatively heterogeneous population, consisting of descendants of Celtic-speaking peoples, Anglo-Saxon tribes, Jutes and others. They are also a population that has undergone recent expansion. Genetic trees, based *e.g.* on classical protein polymorphisms, can be drawn to examine genetic distances between populations. In such trees, the English fall into the same branch as Dutch and the Danish (Cavalli-Sforza *et al.*, 1994).

The Saami are an old European population who speak a Finno-Ugric language. They now inhabit areas near the Polar Circle in Norway, Sweden, Finland and Russia. Traditionally they have been hunters and gatherers, and since approximately the 17th century, also reindeer herders. It is not clear where the Saami originate - in genetic trees based on classical protein markers, they are outliers among European populations (Cavalli-Sforza *et al.*, 1994). The Saami appear to have been of historically constant population size (Sajantila *et al.*, 1995) and show more extensive LD than populations that have expanded rapidly (Laan and Pääbo, 1997, Kaessmann, 2002).

In contrast, Zimbabweans from sub-Saharan Africa are expected to contain more ancient and diverse haplotypes and thus less extensive LD compared with Europeans (as a result of the "out-of-Africa" origin of modern humans, see Figure 9.1). The Zimbabwean population sample should therefore provide a test for erosion of the extended LD block by population processes. Very high diversity in this population has been verified for example

by studies of the insulin gene region and its associated minisatellite, both of which show far higher levels of diversity than seen in European and Asian populations (Stead and Jeffreys, 2002, Stead *et al.*, 2003). The work presented in this chapter aims to determine how known recombination hotspots influence patterns of LD in these very different populations.

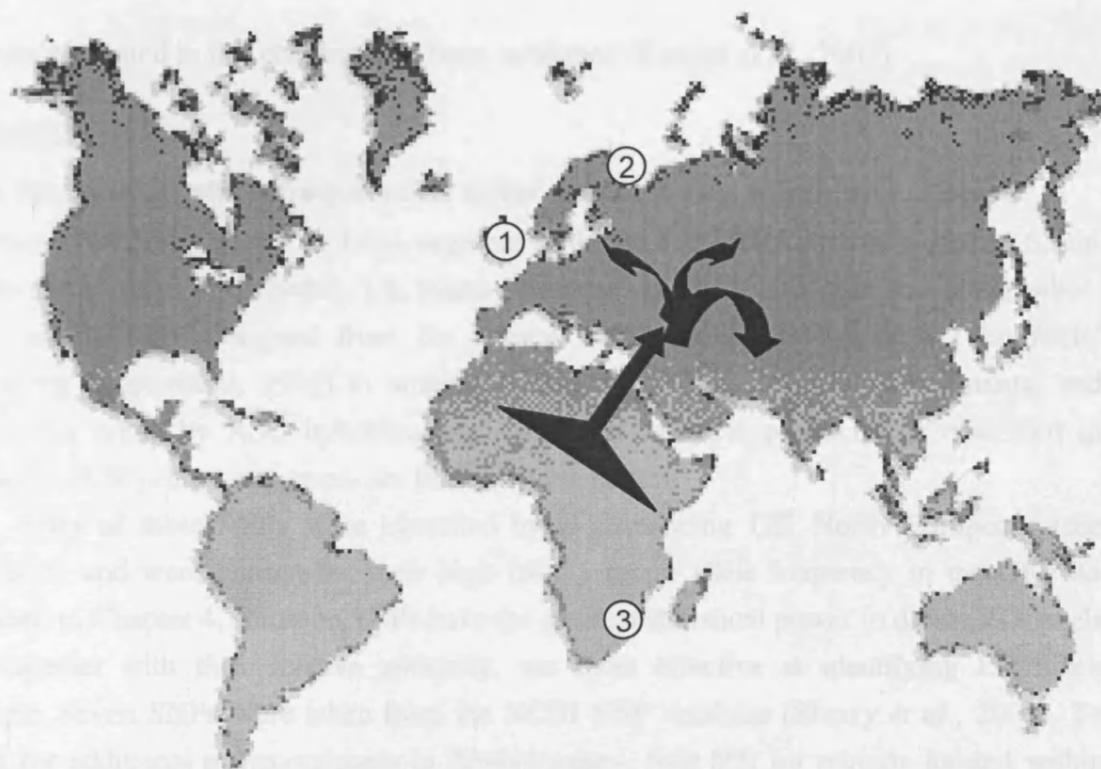


Figure 9.1. Geographical locations of the populations used in this study. 1, UK North Europeans, 2, Saami from the Kola Peninsula in Russia, 3, Zimbabweans. Block arrows indicate the "out-of-Africa" migration of modern humans ~100 000 years ago; hence, most genetic variation seen outside Africa is a subset of the much greater diversity within Africa.

This work

Characterisation of LD in the UK population (Chapter 4), followed by analysis of meiotic recombination hotspots (chapters 5, 6 and 8), showed that most regions of LD breakdown correspond precisely to the location of recombination hotspots. To investigate the effect of meiotic recombination *versus* population history on LD, three populations with different demographic histories (UK North Europeans, Saami and Zimbabweans) were genotyped for high frequency SNPs across a 75-kb DNA segment of the MHC class II region. This region spans three well-characterised crossover hotspots and a 60 kb long LD block. I demonstrate how despite a high level of underlying haplotype diversity and considerable divergence in haplotype composition between populations, all three populations show similar patterns of

LD. Surprisingly, the entire 60-kb LD block was present even in Africans, although it was relatively difficult to detect due to a systematic deficiency of high frequency SNPs (DNA within recombination hotspots did not show this low nucleotide diversity in Africans). Thus, while population history has some influence on LD, my findings suggest that recombination hotspots play a major global role in shaping LD patterns as well as helping to maintain localised SNP diversity in this region of the MHC.

The data presented in this chapter have been published (Kauppi *et al.*, 2003).

9.2. RESULTS

9.2.1. Population allele frequencies differ dramatically within an LD block

I genotyped 64 SNPs in a 75 kb DNA segment of the human MHC class II region in Saami (number of chromosomes $n=80$), UK North Europeans ($n=100$) and Zimbabweans ($n=96$). PCR primers were designed from the current MHC consensus sequence (The MHC sequencing Consortium, 1999) to amplify twelve 3.4-6.7 kb long DNA segments, and SNPs were typed by ASO hybridisation to dotblots of PCR products, as described in Chapter 2. PCR primer sequences are listed in Appendix 3.

Fifty of these SNPs were identified by re-sequencing UK North Europeans (see Chapter 3) and were chosen for their high (>0.2) minor allele frequency in the UK. As discussed in Chapter 4, common SNPs have the greatest statistical power to define LD levels and, together with their relative antiquity, are most effective at identifying LD block structure. Seven SNPs were taken from the NCBI SNP database (Sherry *et al.*, 2001). To screen for additional polymorphisms in Zimbabweans, four 600 bp regions located within the UK LD block were re-sequenced in six Zimbabweans, and seven new SNPs so discovered were typed in all populations (see Figure 9.2). Some of the SNPs (1 in the UK, 5 in Saami, 6 in Zimbabweans) could not be typed, presumably because of additional unknown SNPs that block ASO hybridisation. Amongst the scored genotypes, some were ambiguous, mainly due to poor PCR amplification resulting from low or poor quality DNA input particularly in the Zimbabwean DNA samples; these ambiguous genotypes (3 out of 3084 in UK, 7/2360 in Saami, 122/2784 in Zimbabweans) were excluded from further analyses. In the final dataset, very few markers showed deviation from Hardy-Weinberg equilibrium (2 in UK, 1 in Saami and 3 in Zimbabweans).

In Zimbabweans, I discovered a striking lack of markers with high (>0.25) minor allele frequency in the inter-hotspot region roughly between positions 17 and 70 kb (Figure 9.2, Figure 9.3A). In and near hotspots *DNA3*, *DMB1* and *DMB2*, however, the density of high frequency markers is similar to that seen in the Saami and the UK. Of the seven SNPs that were discovered through re-sequencing Zimbabweans, only one had a minor allele frequency >0.25 . This suggests that the rarity of high frequency SNPs is a genuine feature

of the inter-hotspot region in Zimbabweans, rather than arising through biased ascertainment of SNPs in Europeans.

Minor allele frequencies of SNPs differed systematically between the three populations studied, with Saami allele frequencies tending to be higher and Zimbabwean allele frequencies lower than seen in the UK (Figure 9.2). This is reflected in Nei's genetic distance (Nei, 1972), which is greatest between Saami and Zimbabweans (0.0912) and smaller for UK/Saami (0.0205) and UK/Zimbabweans (0.0340) comparisons. The GENEPOP test for genic differentiation, which compares allele frequencies between pairs of populations (Raymond and Rousset, 1995), showed that allele frequencies at 41% of markers (24 of 59 possible comparisons) were significantly different ($P < 0.05$) between UK and Saami, at 50% of markers (29/59) between UK and Zimbabweans and at 71% of markers (40/56) between Saami and Zimbabweans. These shifts in allele frequency were most marked for SNPs located within the extended LD block seen in the UK (Figure 9.2).

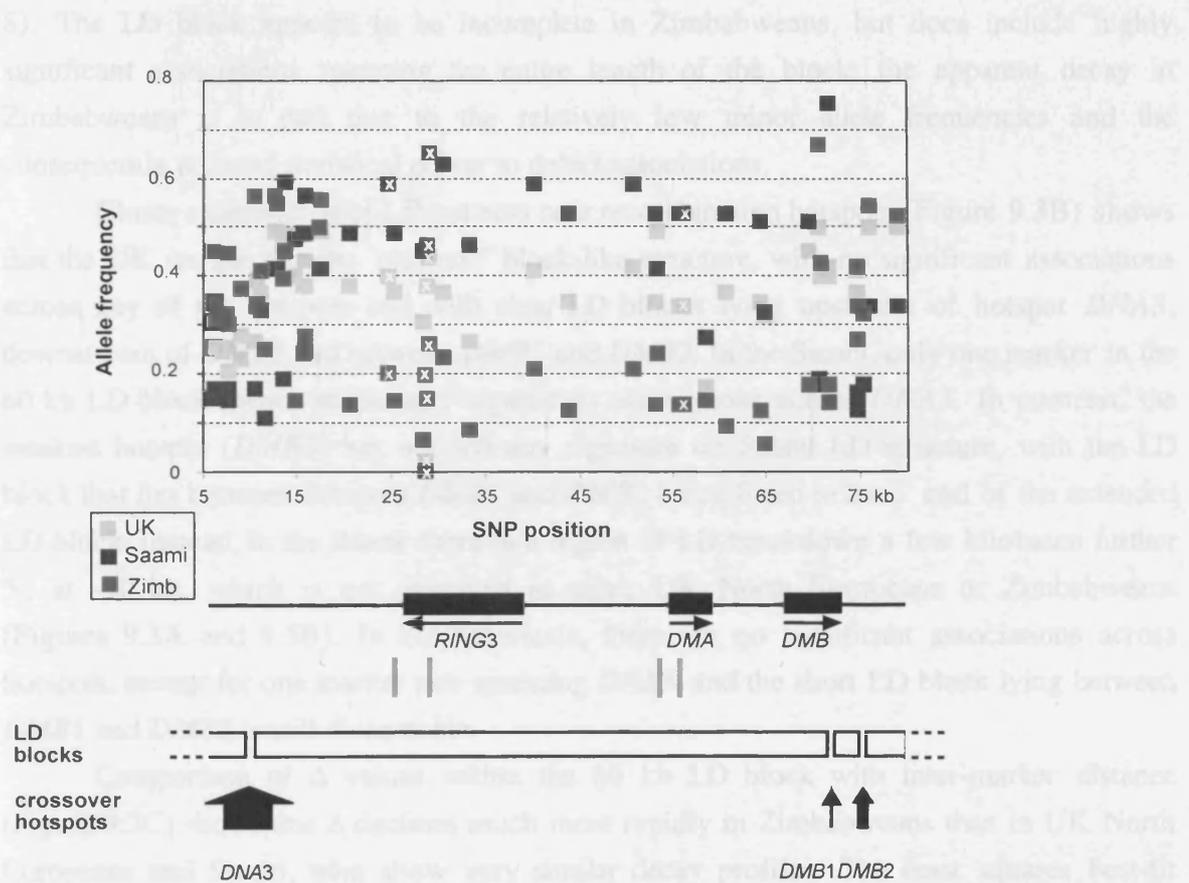


Figure 9.2. Allele-frequency comparisons between populations. Frequencies of the allele which is less common in the UK population are shown for the UK (blue), Saami (black) and Zimbabweans (red). SNPs discovered through re-sequencing Zimbabweans are marked with white crosses. Beneath are shown the location of genes (black boxes) and LD blocks seen in the UK (open boxes). Crossover hotspots are indicated with arrows sized in proportion to their activity in male meiosis (Jeffreys *et al.*, 2001). Regions re-sequenced in Zimbabweans are shown as grey bars below the genes.

9.2.2. The same LD block is observed in all populations

SNP genotypes (minor allele frequency ≥ 0.15) were used to predict maximum-likelihood haplotypes for pairs of markers which were then subjected to LD analyses, as described in Chapter 4. The D' measure of LD only has an absolute value of 1, if a marker pair shows complete association with at most only three of the four possible haplotypes (Lewontin, 1984). $|D'|$ is useful for detecting obligate historical recombination between markers. Analysis of D' values for all marker pairs (Figure 9.3A) showed LD patterns structured into LD blocks. Within these blocks most SNPs were in significant LD, separated by regions corresponding to known recombination hotspots, across which markers did not show significant associations. The block-like LD structures seen in UK North Europeans (Chapter 8 and Jeffreys *et al.*, 2001) are also present in the Saami. Somewhat surprisingly, the outline of the 60 kb LD block is still recognisable in Zimbabweans. In all three populations, this extended LD block terminates at experimentally-verified crossover hotspots (chapters 6 and 8). The LD block appears to be incomplete in Zimbabweans, but does include highly significant associations spanning the entire length of the block; the apparent decay in Zimbabweans is in part due to the relatively low minor allele frequencies and the consequently reduced statistical power to detect associations.

Closer examination of LD patterns near recombination hotspots (Figure 9.3B) shows that the UK sample has the "cleanest" block-like structure, with no significant associations across any of the hotspots and with clear LD blocks lying upstream of hotspot *DNA3*, downstream of *DMB2* and between *DMB1* and *DMB2*. In the Saami, only one marker in the 60 kb LD block shows statistically significant associations across *DNA3*. In contrast, the weakest hotspot (*DMB1*) has not left any signature on Saami LD structure, with the LD block that lies between hotspots *DMB1* and *DMB2* being fused to the 3' end of the extended LD block. Instead, in the Saami there is a region of LD breakdown a few kilobases further 5', at ~ 70 kb, which is not observed in either UK North Europeans or Zimbabweans (Figures 9.3A and 9.3B). In Zimbabweans, there are no significant associations across hotspots, except for one marker pair spanning *DNA3*, and the short LD block lying between *DMB1* and *DMB2* is still discernable.

Comparison of Δ values within the 60 kb LD block with inter-marker distance (Figure 9.3C) shows that Δ declines much more rapidly in Zimbabweans than in UK North Europeans and Saami, who show very similar decay profiles. The least squares best-fit values for the product $N_e r$ per Mb were 10 for Saami, 11 for UK North Europeans and 220 for Zimbabweans. Assuming that recombination rate is constant between populations, this suggests that the effective population size of the Zimbabwean population is approximately 20 times higher than either UK North Europeans or Saami. This high estimate for Zimbabweans provides further evidence that this population is considerably more ancient and diverse than North Europeans.

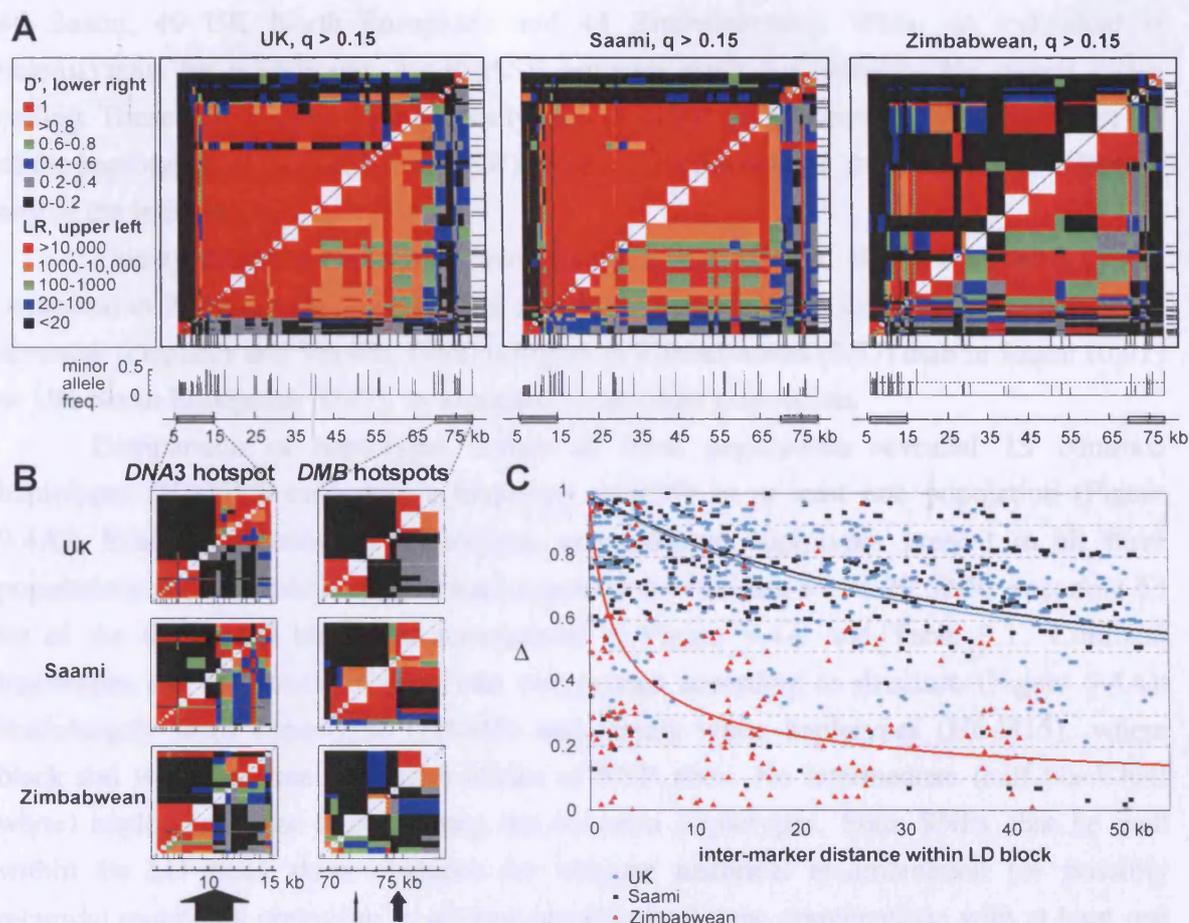


Figure 9.3. Population comparisons of LD patterns. (A) LD in UK North Europeans, Saami and Zimbabweans. $|D'|$ measures (lower right) are shown for all pairs of markers with minor allele frequency ≥ 0.15 together with the associated likelihood-ratio (LR) versus free association (upper left), and colour-coded as indicated in box. Each point is plotted as a rectangle centered on each SNP (locations shown below and to right of plot, minor allele frequencies shown under plot) and extending half way to adjacent markers. In Zimbabweans, only one marker within the LD block has a minor allele frequency ≥ 0.25 (dashed line). (B) Plots expanded to focus on regions around recombination hotspots (open boxes on distance scale in A). The location of SNPs is shown below each plot. (C) Decay of the $|\Delta|$ measure of LD with distance within the LD block. Only markers in the interval 11689-64980 bp and with minor allele frequency ≥ 0.15 were included. Lines show the least squares best-fit curves (UK, blue; Saami, black; Zimbabweans, red) for the relationship $\Delta^2 = 1/(1+4N_e r d)$ where N_e = effective population size, r = recombination rate per base pair interval and d is the distance in bp between markers.

9.2.3. Cosmopolitan haplotypes within the extended LD block are uncommon

To investigate the underlying haplotypes within the 60 kb LD block, I inferred full haplotypes using the PHASE program version 0.21 (Stephens *et al.*, 2001b) from the unambiguous genotypes of 29 SNPs that are located in the region between the *DNA3* and *DMB1* hotspots. SNPs in recombinationally active regions within hotspots were excluded. Each run was repeated 10 times with different seeds and the average haplotype frequencies generated by the runs were calculated. The unambiguous SNP genotypes used as input data for PHASE runs consisted of SNPs with minor allele frequency >0.2 in UK North Europeans, plus all SNPs identified through re-sequencing Zimbabwean DNA (genotyped in

40 Saami, 49 UK North Europeans and 44 Zimbabweans). When an individual is heterozygous for a SNP site, the PHASE program gives a probability for correct phase calling. These probabilities were generally high ($P > 0.95$) for the common haplotypes; for 11 of the haplotypes, at most only one SNP showed any evidence for uncertain phase calling in any of the individuals.

Twenty different haplotypes were found in the 49 UK North Europeans typed, compared to 21 haplotypes in 40 Saami and 40 haplotypes in 44 Zimbabweans. Haplotype diversity (Depaulis and Veuille, 1998) is higher in Zimbabweans (0.97) than in Saami (0.91) or UK North Europeans (0.87), as expected for an older population.

Comparison of haplotypes across all three populations revealed 15 common haplotypes (H1-H15) each with a frequency of $\geq 5\%$ in at least one population (Figure 9.4A). Results for common haplotypes, cosmopolitan haplotypes present in all three populations and population-specific haplotypes for the full data set of 29 SNPs spanning 53 kb of the 60 kb LD block are summarised in Figure 9.4A and Table 9.1. Common haplotypes can be loosely divided into two groups according to structure (Figure 9.4A): black/largely black haplotypes (H1-H5) and largely white haplotypes (H6-H15), where black and white indicate alternative alleles of SNP sites. No intermediate (half black-half white) haplotypes were found among the common haplotypes. Four SNPs that lie well within the LD block show evidence for obligate historical recombination (or possibly recurrent mutation), appearing in all four possible haplotype combinations with at least one other marker within the LD block.

Haplotypes tend to show a high degree of population specificity. For example, haplotype H15 is common in the UK but was not seen in Saami or Zimbabweans. When all haplotypes are considered, each population shows a large fraction of population-specific haplotypes. Cosmopolitan haplotypes shared by all three populations are uncommon, accounting for only 27% of UK chromosomes and 19% of Saami chromosomes, and for just 5% of Zimbabwean chromosomes (Figure 9.4B).

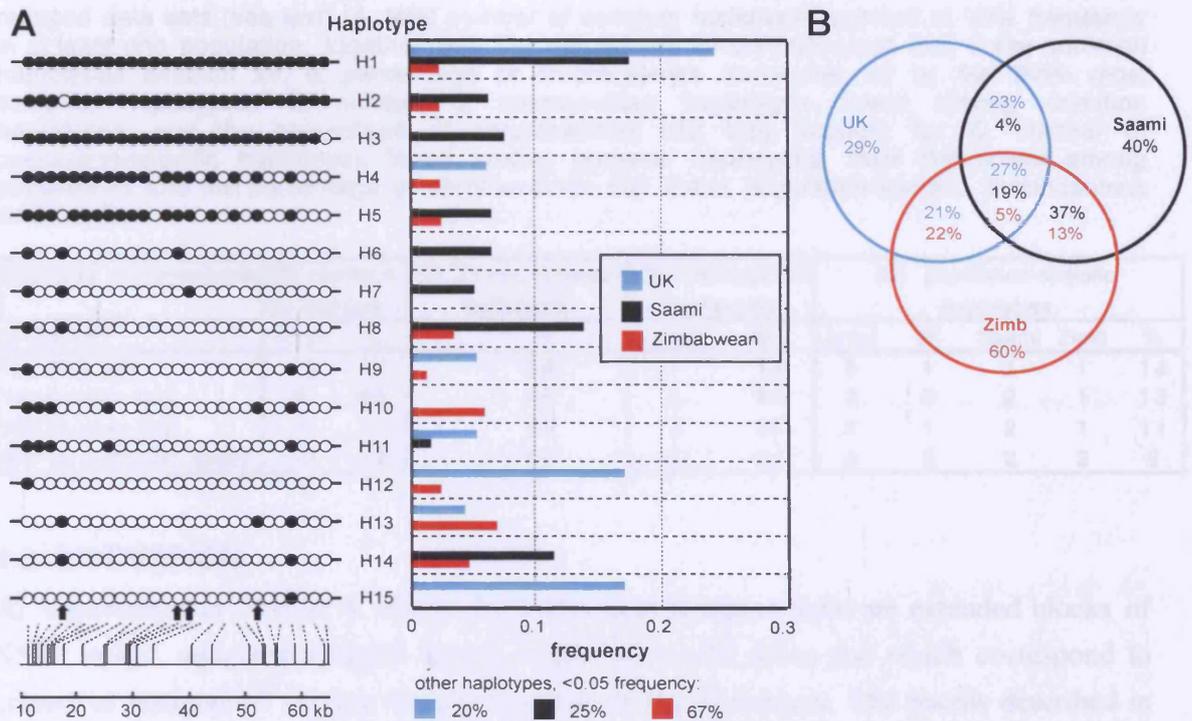


Figure 9.4. Haplotypes within the extended LD block. (A) Fifteen common haplotypes, each at $\geq 5\%$ frequency in at least one population (labelled H1-H15), were inferred from diploid genotype data. The structure of each haplotype is shown on the left, with circles representing SNP sites, shaded black or white to indicate alternative alleles. SNP positions are shown under the haplotypes. Two SNPs that were monomorphic among the 15 common haplotypes are not shown. Haplotypes can be roughly grouped according to structure, with H1-H5 largely black and H6-H15 largely white. SNPs showing evidence for gene conversion (four haplotypes with at least one other marker) are arrowed. The frequencies of each haplotype in each population are shown on the right (UK, blue; Saami, black; Zimbabweans, red). (B) Percentage of haplotypes that are shared between all three populations, between two populations and that are unique to each population (percentages in blue for UK, black for Saami and red for Zimbabweans).

Because higher marker density will tend to increase haplotype diversity, I also inferred LD block haplotypes from a reduced data set using 15 of the 29 SNPs selected at random, giving an average marker density of 1 SNP every 2.8 kb and retaining 3 of the 4 SNPs that show evidence for obligate historical recombination ("1st random half" data set in Table 9.1). I then inferred haplotypes from the remaining 14 SNPs, containing just one marker with evidence for obligate historical recombination, and from a data set where only the four SNPs showing evidence for obligate historical recombination were removed ("2nd random half" and "without recombinant SNPs" data sets in Table 9.1, respectively). The first reduction of the data set had a more dramatic effect on lowering haplotype diversity and increasing the incidence of cosmopolitan haplotypes than the second, suggesting that both recombination within the LD block and marker density contribute to haplotype diversity and the population-specificity of haplotypes.

Table 9.1. Inferred haplotypes within the extended LD block using the full data set and reduced data sets (see text). A, total number of common haplotypes present at $\geq 5\%$ frequency in at least one population, together with the percentage of chromosomes that these common haplotypes account for. B, percentage of chromosomes accounted for by the three most common haplotypes. C, number of cosmopolitan haplotypes found among common haplotypes, and the percentage of chromosomes that they account for. D, number of population-specific haplotypes found among common haplotypes, their distribution among populations and the percentage of chromosomes that these population-specific chromosomes account for.

total:266 chromosomes DATA SET	(A) common haplotypes		(B) 3 most common haplotypes	(C) cosmopolitan haplotypes		(D) population-specific haplotypes				
	total	%	%	total	%	total	UK	Saami	Zimb	%
full data set	15	62	28	1	15	5	1	3	1	14
1st random half	9	82	49	5	60	3	0	2	1	13
2nd random half	11	74	47	3	36	4	1	2	1	11
w/o recombinant SNPs	10	79	52	5	64	4	0	2	2	9

9.3. DISCUSSION

As summarised in Chapter 8, within the MHC class II region there are extended blocks of SNPs in LD, separated by short regions where LD breaks down and which correspond to crossover hotspots of varying intensities in UK North Europeans. The results described in this chapter for the *DOA* to *DMB* interval show that this LD block pattern, as defined by SNPs with high frequency alleles that are likely to be ancient, is shared by two additional populations with very different histories. This sharing is even seen in Zimbabweans, despite the evidence that this population is ancient and genetically very diverse. Surprisingly, the Saami do not exhibit either reduced haplotype diversity or increased levels of association within the extended LD block or across the strongest recombination hotspots, as might be expected of a constant-sized population.

Previous analyses of LD structure elsewhere in the human genome have similarly shown that LD blocks seem to be common and are often shared between populations (Reich *et al.*, 2001, Gabriel *et al.*, 2002). LD blocks tend to contain limited repertoires of typically 2-5 common haplotypes defined by common SNPs (Stephens *et al.*, 2001a, Gabriel *et al.*, 2002, Daly *et al.*, 2001, Patil *et al.*, 2001), some of which can be shared across major population divides, for example non-Africans and Africans (Stephens *et al.*, 2001a, Gabriel *et al.*, 2002, Patil *et al.*, 2001). The existence of these apparently cosmopolitan haplotypes suggests that much of the current haplotype structure in human populations is ancient and was established prior to the recent diversification of mankind, with more recent influences such as founder effects modifying LD patterns. For example, two or more ancient LD blocks observed in Africans may be fused into a single block in younger (non-African) populations.

This study of the MHC class II region reveals what appears to be a very different picture. First, the three populations surveyed all show very high levels of haplotype diversity within the extended LD block, with 15 common and many rare haplotypes in the overall data set. This diversity is not the result of our using an excessively high density of

SNPs; our density (averaging 1 SNP per 1.8 kb) is comparable to the density used in the LD surveys of a 500 kb region on chromosome 5 (1 SNP per 5 kb [Daly *et al.*, 2001]) and of chromosome 21 (1 SNP per 1.3 kb [Patil *et al.*, 2001]). The 60 kb LD block is therefore unusually diverse, possibly reflecting diversifying selection acting on the MHC. Alternatively, the increased diversity may reflect differences in the definition of LD blocks. Our definition is based on the very clear-cut localised breakdown of LD which correspond to recombination hotspots and serve to demarcate LD blocks (Figure 9.3A). Other studies used SNP pairs showing "strong LD" (Daly *et al.*, 2001) or "strong evidence for historical recombination" (Gabriel *et al.*, 2002) to define LD blocks. It appears that these criteria are more stringent than ours, and if applied to our LD block, could result in an artificial break-up into smaller blocks, each of lower diversity.

More significantly, there is little evidence within the MHC LD block for cosmopolitan haplotypes as seen elsewhere in the human genome (Stephens *et al.*, 2001a, Gabriel *et al.*, 2002, Patil *et al.*, 2001). Rather, many of the haplotypes appear, within the limits of our sample sizes, to be largely if not completely population-specific. This specificity even extends to the two European samples; for example, the second most common Saami haplotype (H8 at 14% frequency) was not seen in the UK, and likewise the second most common UK haplotypes (H12 and H15, both at 17% frequency) were not found in the Saami (Figure 9.4A). This implies that this region has been subjected to a high rate of haplotype turnover and that current haplotype structures were not established early in human history. Despite this turnover, LD block structure has remained remarkably constant in all three populations. It therefore follows that the crossover hotspots must have played the dominant role in shaping these patterns in all three populations, and that they have been sufficiently active to prevent LD block fusion, even in young populations. The only exception is hotspot *DMB1*, which has left no clear imprint on LD structure in the Saami. This hotspot is weak, contributing only 0.003 cM to linkage map distance in males (based on data from one sperm donor, activity in female meiosis is unknown), compared with 0.13 cM and 0.03 cM for hotspots *DNA3* and *DMB2* respectively (see Chapter 8 and Jeffreys *et al.*, 2001). It therefore appears that the more intense hotspots are sufficiently active to override effects of population history on LD patterns, a conclusion supported by a recent study of human genome-wide sequence variation (Reich *et al.*, 2002).

Haplotype turnover events within the MHC LD block include what appear to have been recombinational exchanges involving four SNPs deep within the block. However, these exchanges have not generated any obvious crossover haplotypes (half black, half white, Figure 9.4A) and may instead have involved either localised gene conversion events or alternatively recurrent mutation at these SNPs. The latter appears unlikely because two of these sites are transversions and only one is a transition at a potentially-hypermutable CpG doublet. It is also noticeable that these apparent recombination events within the LD block have not resulted in the collapse of the extended LD block into two or more clearly-defined

shorter LD blocks in Zimbabweans, who are an old population in which very weak recombination hotspots would have the potential to leave their mark on LD structure. This suggests that rare recombination events within the LD block do not cluster into weak hotspots.

While the MHC LD block structure is remarkably similar in the three populations tested, the extended LD block is relatively difficult to detect in Zimbabweans. Part of the difficulty lies in the systematic dearth of high frequency SNPs within the LD block in Zimbabweans, which limits the statistical power to detect associations. A similar systematic shift is seen in the higher frequencies of minor alleles in Saami compared to UK North Europeans (Figure 9.2). This correlated behaviour of multiple SNPs is expected for a recombinationally-suppressed region of DNA in which allele fixation or extinction at different SNPs will not occur independently - instead, loss of a haplotype from a population will result in loss of all variants restricted to that haplotype. It is noticeable that SNPs within and near the recombination hotspots appear "immune" to these correlated shifts in allele frequencies (Figure 9.2). Again, this is expected; SNPs in recombinationally active DNA can escape rapid correlated extinctions by being constantly reshuffled onto different haplotypes. The implication under this model is that hotspots should always show relatively stable levels of nucleotide diversity in the absence of additional molecular processes that could influence SNP diversity, such as meiotic drive (Jeffreys and Neumann, 2002). In contrast, recombinationally inactive LD blocks will show much more unpredictable levels of high frequency SNPs. A block will be difficult to detect by LD analysis if it happens to be in a period of low diversity, whether purely by chance or through processes such as selective sweeps.

It remains unclear whether the MHC class II region is wholly exceptional in its level of haplotype diversity and its rapid turnover, with perhaps turnover processes being aided by selection on the MHC. Additional work is needed to see if LD structuring in other regions of the human genome does not merely reflect ancient events but is instead a dynamic ongoing process involving a complex interplay between recombination hotspots, SNP turnover, drift and selection.

10. DISCUSSION

In this thesis, I have described the systematic survey of recombination activity across long segments of the human MHC class II region. Importantly, these studies were not biased by prior knowledge of familial crossover clustering; in fact, the target regions were stretches of DNA where *no* evidence for localised crossover hotspots had been found by other methods (either pedigree or data on single flow-sorted sperm). I have identified and characterised three previously unknown crossover hotspots: the 3' *DPA* plus *DMB1* and *DMB2* hotspots. In this chapter, I aim to consolidate the wealth of recent information on human LD structure and meiotic recombination hotspots, in part contributed by work presented in this thesis.

10.1. LD AND HAPLOTYPE BLOCKS IN THE HUMAN GENOME

The phenomenon of linkage disequilibrium, *i.e.* the non-random association of alleles at separate loci, between human DNA polymorphisms was first noted over 20 years ago by Antonarakis *et al.* (1982), who observed a limited number of allelic combinations (haplotypes) at the human beta globin gene cluster. LD patterns are now being studied intensively as knowledge of haplotype structure would greatly facilitate attempts to map complex diseases (Collins *et al.*, 1997, Cardon and Bell, 2002). Association mapping relies on the assumption that LD is highest between a disease-causing mutation and its closest markers. In the last three years or so, it has become clear that LD does not decay linearly with distance, as simulated by Kruglyak (1999), who assumed a random distribution of crossovers. Instead, a large part of the human genome seems to consist of LD blocks (or haplotype blocks), both terms which have been coined very recently.

Haplotypes, combinations of alleles at different loci inherited together as a "package" from one generation to the next, are usually inferred from diploid genotype data. Recently, however, Patil *et al.* (2001) analysed true chromosome 21 haplotypes that were obtained by inserting single chromosomes into hamster cells. Their study and others (our data on the MHC [Jeffreys *et al.*, 2001], Daly *et al.*, 2001, Johnson *et al.*, 2001, Gabriel *et al.*, 2002) showed that so-called haplotype blocks are common in the human genome. Within a haplotype block genetic markers, in this case SNPs, are strongly associated. Block lengths in a 500 kb region on chromosome 5q31 varied from 3 to 92 kb (Daly *et al.*, 2001). Haplotype blocks in the genome can occasionally be very long, such as the 804 kb block on chromosome 22 (Dawson *et al.*, 2002). We showed that in a 210 kb segment of the MHC, LD block structure was caused by the presence of true meiotic recombination hotspots (Jeffreys *et al.*, 2001). Our data led to speculations that most of the blocks across the human genome are caused by the clustering of recombination events (Daly *et al.*, 2001). However, the block structure may at least sometimes reflect historical random crossover events that have managed to spread in the population through drift (Zhang *et al.*, 2002, Wang *et al.*,

2002, Phillips *et al.*, 2003). To date, our study of LD and sperm crossover hotspots in the MHC is the only experimental evidence for this causal relationship.

The first evidence for the block-like structure of LD in the human genome was presented in three articles, all by separate groups but published in the same issue of Nature Genetics in October 2001 (Johnson *et al.*, Jeffreys *et al.* and Daly *et al.*). Johnson and co-workers (2001) suggested the use of "haplotype tag SNPs" (htSNPs) for attempts to identify alleles contributing to common disease. These htSNPs would capture the information of their haplotypes, and would therefore cut down the cost and labour of having to analyse a much larger number of SNPs. Gray (2000) had developed the same idea a little earlier. The prospect of using htSNPs to simplify the search for genetic determinants of common disease immediately sparked huge interest. HtSNPs would reduce the problem of disease mapping to first finding haplotype blocks, and then within the patient cohort, finding the haplotype in the block with the strongest association to disease.

10.1.1. The HapMap project

If the entire human genome consisted of haplotype blocks, the number of SNPs that need to be genotyped for disease gene mapping purposes could be reduced from 10 million (thought to be the total number of SNPs in the human genome) to about 500 000. The HapMap project, an international effort launched by the National Human Genome Research Institute in the United States in October 2002, aims to map haplotype block structure in DNA samples from Nigeria, Japan, China and the United States, and define htSNPs for each block. The project is expected to take three years to complete, with a total cost of 100 million USD (<http://genome.gov/10005336>). The HapMap project will yield no direct information on the cause of haplotype blocks, since for the human disease mapping community the main interest lies in defining block size and htSNP numbers. It will, however, begin to generate data on how long and how "universal" haplotypes are.

An MHC Haplotype Map effort is also underway, to analyse eight different MHC haplotypes (Allcock *et al.*, 2002). The aim is to map all SNPs in the MHC, and like in the HapMap project, to identify haplotype tag SNPs in order to minimise genotyping effort in disease association studies.

10.1.2. Relating LD to crossover hotspots

As discussed in Chapter 4, LD analysis can be used to infer the historical recombination activity of DNA segments. Although this approach is complicated by population genetic factors (see Chapter 9), it is a useful first step in homing in on recombinationally active regions. Work presented in this thesis demonstrates that in the MHC class II region there is a reasonably good correlation between LD and meiotic recombination hotspots. High-resolution LD mapping has certainly proved to be an extremely useful tool for mapping hotspots in the MHC – every region of discrete LD breakdown looked at thus far has been

found to contain a sperm crossover hotspot, albeit at times these were not very "hot" (Chapter 8). It is not yet clear how well-defined LD blocks are elsewhere in the human genome. For example, LD analyses of chromosome 22 (J.K. Holloway and A.J. Jeffreys, unpublished) have shown that LD can be erratic and that LD block boundaries can be extremely difficult to locate, with the lack of high-frequency SNPs posing a major problem.

Also, there is some evidence that a recombination hotspot does not necessarily break down LD. In other words, where an LD block is observed, one cannot say with certainty that this stretch of DNA does not recombine. The MS32 hotspot, identified through studies of germline instability, has a recombination activity about 50 times higher than genome average (Jeffreys *et al.*, 1998a), yet in UK North Europeans, there is little evidence for LD breakdown, as detected with SNPs with minor allele frequency ≥ 0.15 (A.J. Jeffreys, unpublished). Furthermore, in the Xp/Yp pseudoautosomal pairing region near the *PGPL* gene, only moderate LD breakdown was found, but sperm crossover analysis revealed intense (~ 90 cM/Mb) uniform recombination activity (M.T. Slingsby, unpublished).

How many hotspots may have been missed, *i.e.* how many could be located inside LD blocks and would therefore not have been tested for? All seven hotspots identified in the MHC class II to date contribute a total of 0.2 cM to the genetic map distance across 292 kb, and long LD blocks contribute ~ 0.05 cM (assuming $N_e=10000$). This would suggest that at least 80% of crossover events in this region occur within hotspots. Female recombination rate is of course not accounted for; however, familial crossovers suggest that the same hotspots are at least as active in females (Cullen *et al.*, 1997), and that at least no highly active additional female-specific hotspots are present. Neither family data nor LD block structure in the UK population (and for the *DNA3-DMB* hotspot interval, in Saami and Zimbabweans) support the existence of additional localised crossover hotspots within the LD blocks; this is not to say that weaker hotspots do not exist. However, the decay of LD with distance suggests some weak crossover activity within the LD blocks, and presumably results from the chance success of randomly distributed crossovers.

In any case, our method is extremely powerful in identifying novel hotspots of recombination. This is illustrated by the fact that within our 292-kb target region, now known to contain at least seven highly localised crossover hotspots, other methods (pedigree data and single sperm analysis combined) have only been able to identify two localised regions of crossover clustering: the *TAP2* and the *DOA* gene regions (Cullen *et al.*, 1995, 1997, 2002).

It is remarkable how data from crossover assays in sperm agree with LD, a population-based measure of recombination. The former involves direct detection of recombinants in sperm that have not been subjected to natural selection (apart from successfully completing meiosis), whereas the latter consists of a collection of different

haplotypes in the population, presumably subjected to selective forces. This suggests that in this segment of the MHC, selection has had little effect on the length of haplotype blocks and further supports the idea that recombination hotspots, rather than population genetic factors, are the key player in shaping the genome (see also Chapter 9). Of course, it is possible that the location of the hotspots is dictated by selection in the first place.

Based on our data, how feasible are the goals of the HapMap project? We found that not all regions are composed of clear LD blocks; in some regions LD structure is more complicated (for example the *DPB1* region, Chapter 7). The same seems to be true for a number of genomic regions (Wall and Pritchard, 2003). The only strategy then would be intense re-sequencing effort to properly define LD block boundaries, keeping in mind the possibility that for a given region or in a certain population high-frequency SNPs, or indeed clear block boundaries, may simply not exist. It is not yet known what number of SNPs is needed to capture all haplotype blocks, and what statistical thresholds are appropriate to define block boundaries. As discussed in Chapter 9, depending on the (somewhat arbitrary) stringency of block definition, a large LD block may be broken into shorter sub-blocks, each of lower diversity. Thus, htSNPs could be used to tag large discrete blocks but not all of the genomic sequence (Wall and Pritchard, 2003). It is also possible that aetiological variants lie *inside* recombination hotspots (for example the *SHOX* hotspot is centred in an exon [May *et al.*, 2002]); in such cases they would be impossible to identify using a haplotype tagging strategy.

10.2. WHAT ABOUT GENE CONVERSION?

In yeast, double-strand break sites (recombination hotspots) co-localise with sites of initiation and resolution of both crossovers and conversions. There is recent evidence that this may be true for mice and humans as well (see Jeffreys *et al.*, 1998a and Jeffreys and May, 2004 for conversion data at the MS32 and MHC hotspots, respectively, and Guillon and de Massy, 2002, for the mouse Eb hotspot).

Due to the nature of the crossover assay used to characterise the MHC hotspots, all gene conversions without crossover go undetected. Because conversion events can involve just one marker, a single base change detected by our PCR based method could be a conversion, but equally well a result of PCR mis-incorporation in the early cycles. Seeing that conversion tracts can be very short (often less than 300 bp), and therefore would often only capture one SNP site, this imposes a major methodological problem. An enrichment-based assay termed DEASH (DNA enrichment by allele-specific hybridisation) has recently been developed to circumvent this (Jeffreys and May, 2003). DEASH can be used to enrich for pools of molecules that contain certain allelic combinations. In the DEASH assay, crossovers are used as an internal control for conversions. Therefore, sperm crossover rate at the hotspot must be sufficiently high. So far, *DNA3* and *DMB2* crossover hotspots in the MHC, and the *SHOX* hotspot in the pseudo-autosomal region have been assayed for

conversions using the DEASH method prior to PCR amplification. It seems that in humans, all crossover hotspots are also conversion hotspots, with conversion events dominating over crossover at a 4:1 to 15:1 ratio (Jeffreys and May, 2004). Furthermore, the conversion data on MHC hotspots implies that crossovers are resolved in close proximity of the initiation site, all within the localised hotspot region.

Conversions occurring within crossover hotspots should have little if any effect on haplotypes contained within long LD blocks, as they only create novel allelic combinations at a very localised level within hotspots (assuming the majority of LD blocks do not contain additional crossover hotspots). This does not, however, rule out the existence of conversion hotspots elsewhere in the genome. My population data (Chapter 9) shows some evidence for conversion and/or recurrent mutation occurring within an LD block, but these events did not cluster into specific regions.

10.3. HOW DOES A CELL CHOOSE A RECOMBINATION HOTSPOT?

Average spacing of hotspots or hotspot clusters based on our MHC data is ~65 kb. If this is representative of the 3.2 billion bp human genome, then there should be ~50000 recombination hotspots. Therefore, chromosome 6 (190 Mb) should have ~3000 potential recombination hotspots, of which only a few will be used for crossing-over in any one meiosis. What makes a DNA segment a recombination hotspot, and how does a chromosome "decide" to cross over at one of the many potential hotspots?

So far we have characterised seven highly localised hotspots in the MHC class II region, and three hotspots elsewhere in the genome (Table 10.1). All hotspots have a very similar shape, with crossovers clustering to a region 1-2kb wide, suggesting that similar mechanisms operate at all of them. Peak recombination rates vary from below genome average at the *DNA1* hotspot to 100-fold above genome average at the *DNA3* hotspot. All hotspots are located in different genomic environments and share no obvious sequence similarities. The fact that some MHC crossover hotspots are found in clusters (three at or near the *DOA* gene, two at or near the *DMB* gene) implies that they are contained within a longer DNA segment which is prone to recombination, *i.e.* the DNA is somehow exposed to the recombination machinery. LD data and in the case of some hotspots, limited numbers of familial crossovers (Cullen *et al.*, 1997), are in good agreement with sperm crossover data. This suggests that the same hotspots exist in females as well, but the crossover activity may be different. Therefore, the same genomic locations may be potential recombination hotspots in all individuals, but there could be some mechanism acting as a "volume knob" that sets the crossover activity to a different level in the two sexes.

Comparison of these recombination hotspots gives no obvious clues as to what might be driving recombination in the human genome. Thuriaux (1977) pointed out that recombination may be associated with genes, because in a variety of organisms there is a good correlation of the overall amount of recombination with gene content (rather than

physical size of the genome). In yeast, most recombination hotspots are found in promoters. The MHC hotspots are all located at or near (within ~8 kb of) genes, but as the MHC is a gene-rich region, this association is not necessarily meaningful. Furthermore, the MS32 hotspot lies in the middle of a 77-kb non-coding DNA segment, arguing against a 1:1 correlation of genes and hotspots. Also, some MHC genes, at least within the 292-kb segment that we tested, most likely do not have adjacent hotspots. More examples of fully characterised hotspots are needed to investigate this further.

Higher-order chromatin conformation is likely to play an important role in controlling the location of crossovers. For example, non-recombining DNA could be contained in chromatin loops of variable sizes (corresponding perhaps to LD blocks of variable sizes), and the tips of the loops would be capable of recombination (Figure 10.1). Each hotspot region may contain sequence features that are capable of forming unusual secondary structures (*e.g.* palindromes or polypurine tracts) and consequently loosening the DNA locally. In this sense, the DNA sequence context would be vital for hotspot function; however, the specific sequence itself can vary between hotspots, which would explain why we are unable to identify sequence similarities.

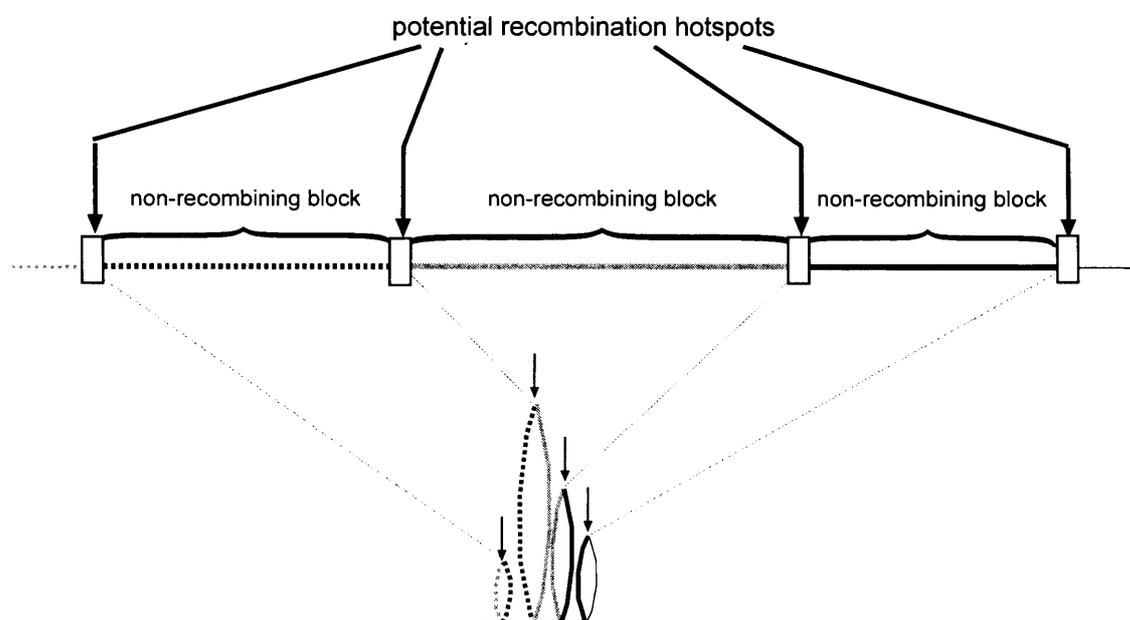


Figure 10.1. Model of how non-recombining blocks could result from the organisation of DNA into chromatin loops. In this model, only DNA located at the tips of the loops (arrowed) has the potential to recombine. If chromatin loops vary in size, then spacing of potential hotspots would vary, too (as observed in the MHC). LD blocks would most of the time correspond to non-recombining axis-associated and chromatin loop DNA. The tips of the loops may contain a larger segment of recombinogenic DNA (white boxes), *i.e.* several localised hotspots, resulting in crossover hotspot clustering, such as that observed in the *DOA* gene region.

Table 10.1. Properties of human crossover hotspots analysed at high resolution (adapted from Jeffreys *et al.*, 2003, data on *NID* hotspot, A.J. Jeffreys unpublished).

	location	width (kb)	peak activity (cM/Mb)	genomic environment	references
3' <i>DPA</i>	MHC class II region	1.6	27	intergenic, in HERV element	this work
<i>DNA 1</i>	MHC class II region	1.9	0.4	promoter	Jeffreys <i>et al.</i> (2001)
<i>DNA 2</i>	MHC class II region	1.3	8	intergenic, in Alu element	Jeffreys <i>et al.</i> (2001)
<i>DNA 3</i>	MHC class II region	1.2	100	intergenic, in Alu element	Jeffreys <i>et al.</i> (2001)
<i>DMB 1</i>	MHC class II region	0.8	5	third intron of <i>DMB</i> gene	Jeffreys <i>et al.</i> (2001) + this work
<i>DMB 2</i>	MHC class II region	1.3	45	intergenic, in polypurine tract	Jeffreys <i>et al.</i> (2001) + this work
<i>TAP 2</i>	MHC class II region	1	8	second intron of <i>TAP 2</i> gene	Jeffreys <i>et al.</i> (2000)
<i>MS32</i>	chromosome 1q42.3	1.5	50	intergenic, in LTR element	Jeffreys <i>et al.</i> (1998a)
<i>SHOX</i>	Xp/Yp <i>PAR 1</i>	2	300	<i>SHOX</i> exon	May <i>et al.</i> (2002)
<i>NID</i>	chromosome 1q42.3	1.5	20-200	intron of <i>nidogen</i> gene	A.J. Jeffreys, unpublished

It is also likely that there are additional layers of crossover control, such as *cis*-acting factors, as identified in mouse studies. In mice, haplotype specificity as well as sex-specificity has been observed for some crossover hotspots. So far, we have little evidence for this in humans. This may simply be because such hotspots have not yet been identified. However, human haplotypes tend to be a lot less divergent than those of inbred mouse strains. Therefore, it seems feasible that whatever *cis* or *trans* acting factors may modulate recombination at hotspots, they would be shared between individuals. Reciprocal crossover asymmetry (Jeffreys and Neumann, 2002), however, shows that local sequence differences between individuals can modify recombination activity.

In yeast, sequence heterology between strains at DSB sites suppresses recombination (Borts and Haber, 1987). No evidence for this has been found in humans. Our crossover assays are, however, limited to donors who are heterozygous for hotspot-flanking SNPs and a sufficient number of internal SNPs. Therefore, the relationship between within-hotspot sequence heterology and recombination rate is difficult to test in humans, and would certainly require a very large panel of semen donors.

In yeast, a greater than 40-fold variation in the frequency of chromosome breakage between DSB sites has been noted (see section 1.3.3). For the MHC class II hotspots, an even greater range of recombination activity was observed – from 0.4 cM/Mb at *DNA1* hotspot to 100 cM/Mb at *DNA3* hotspot. This rate variation results from the cell "choosing" to recombine at one hotspot more frequently than at another. It is unclear how the choice is made between the numerous sites along the DNA that constitute potential recombination hotspots.

10.4. MOUSE CROSSOVER HOTSPOTS VS. HUMAN HOTSPOTS

There is good information on mouse MHC class II meiotic recombination hotspots (see Section 1.5), enabling us to compare mouse *vs.* human hotspot distributions. It should be noted, however, that mouse hotspot data (with the exception of two recent publications, Guillon and de Massy, 2002, Yauk *et al.*, 2003) are based on pedigree analyses; hence, it is very likely that additional weaker hotspots exist in the mouse MHC.

Gene order in the mouse and human MHC is conserved, whereas crossover hotspot locations do not appear to be (Figure 10.2). Within the regions where we observe the seven highly localised human hotspots, no hotspots have been reported within equivalent homologous genomic regions in mice. This would indicate firstly, that hotspot location is not dictated by "linkage groups" of particular genes, or at least that the requirement for genes contained within a haplotype is different in the two species. Secondly, the mouse-human comparison shows that recombination hotspots are relatively transient, with a life span at least shorter than the time since mouse-human divergence. It would be interesting to see if hotspot locations are conserved between humans and evolutionarily much closer species, *e.g.* great apes.

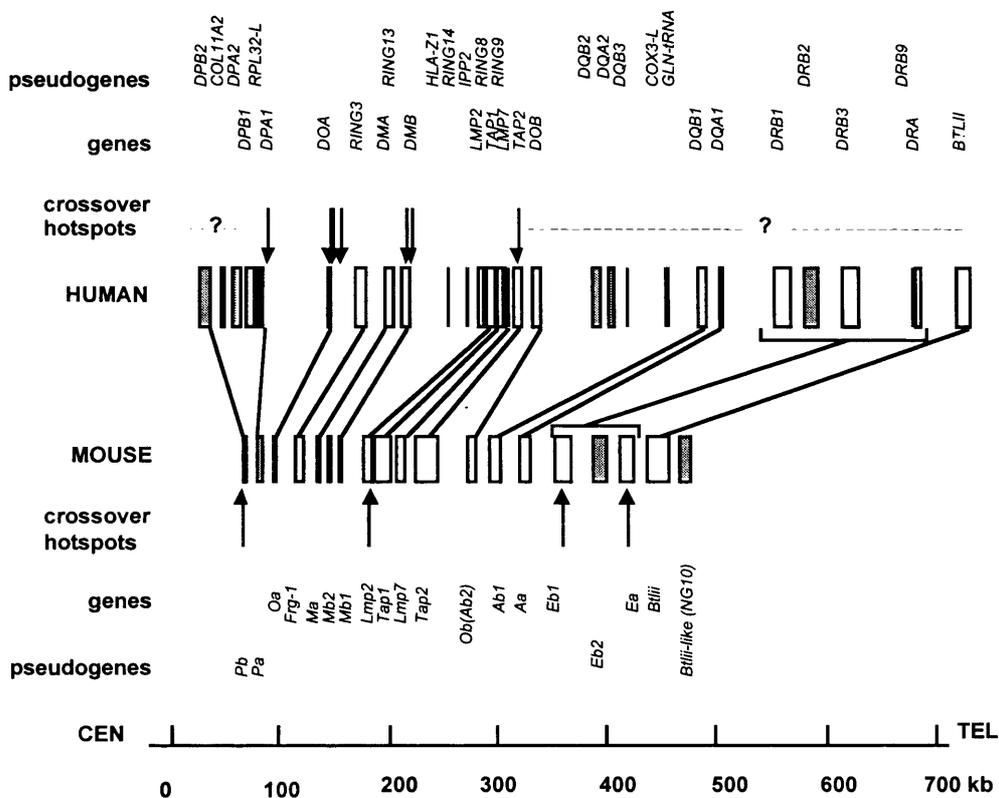


Figure 10.2. Comparison of crossover hotspot locations in the human and mouse MHC class II region. Known hotspots are indicated with arrows.

10.5. CONSEQUENCES OF RECOMBINATION ACTIVITY FOR GENOMIC DIVERSITY

Recombination hotspots seem to dominate human genome plasticity, by structuring diversity into haplotype blocks and further, by stabilising these blocks in different populations (Chapter 9). The existence of LD blocks leads to a correlated behaviour of multiple SNPs - within a recombinationally-suppressed region of DNA, allele fixation or extinction at different SNPs will not occur independently. Instead, loss of a haplotype from a population will result in loss of all variants restricted to that haplotype.

As shown in Chapter 3, I identified some extremely SNP rich regions in the human MHC, but found no clear correlation between recombination activity and SNP diversity (Chapters 5 and 6). A close inverse correlation between nucleotide diversity and LD has been noted (Ardlie *et al.*, 2002). Also, there is apparently a strong positive correlation between nucleotide variability and recombination rate on the megabase level (Nachman, 2001); it has been postulated that recombination may be mutagenic. There is no obvious biological reason why this should be true as recombination is a remarkably faithful process. One can imagine, however, as pointed out in Chapter 9, that SNPs within and near the recombination hotspots may be able to escape correlated shifts in allele frequencies within LD blocks, *i.e.* avoid rapid correlated extinctions by being constantly reshuffled onto different haplotypes. This idea is supported by the relatively stable levels of nucleotide diversity seen in and around hotspots across three populations. In contrast, recombinationally inactive LD blocks will show much more unpredictable levels of high frequency SNPs, such as the LD block in Zimbabweans that lacks high frequency SNPs (Chapter 9).

10.6. CONCLUSIONS AND FUTURE DIRECTIONS

We have for the first time been able to demonstrate the relationship between LD block structure and experimentally-verified sperm crossover hotspots in human DNA. We have shown that crossover distribution is far from random and that crossover hotspots are likely to play a major role in shaping genome diversity. The genomic region analysed for recombination hotspots is atypical of the human genome, because the MHC is extremely gene rich and presumably under strong selection; it is not clear if our results can be generalised to the rest of the human genome. On the other hand, it is difficult to choose a "typical" segment of the human genome. Data on a number of genomic regions suggest that the block-like patterns of LD are common (Gabriel *et al.*, 2002, Reich *et al.*, 2002); the HapMap project should reveal how general they are, and how long blocks are on average. Once a large number of regions of LD breakdown have been identified, they could in theory be systematically targeted for sperm crossover assays. The drawback of our method lies in the laborious fine-tuning of the assay, which makes it almost impossible to scale up. Alternative approaches are therefore necessary to analyse crossover on the whole genome scale; haplotyping individual DNA molecules across putative crossover hotspots using the PCR colony method (Mitra *et al.*, 2003) may be one way to do this.

Appendix 1. PCR and sequencing primers

primer	5'	3'	amplicon length
DPA2-DOA interval			
D-100.8F	CAGGCACTATTAGTCTTGCC		4467 bp
D-96.4R	ACTCAGGCTCATAACTGCC		
D-82.6F	CTTAGGAGTCTTGCCAGAGG		561 bp
D-82.0R	GAAGCCACTGAATCAAGCAC		
D-73.5F	ACAACGTGCCAGTTCACAGG		5021 bp
D-68.5R	AGAAGTTGGCAAGTGCAGGC		
D-67.2F	CACCTCCTTCAAAGCATCTG		1343 bp
D-65.9R	GAGAATTTTAGGTACCAGCG		
D-62.0F	CTGTCTTTATGCAGCGGACC		1981 bp
D-60.1R	TTTGTACCTGAGGTCCAGCC		
D-61.5F	GTGATTCATCCCAGAAGAGG		4343 bp
D-57.2R	CTCCCCGCTCTGAAATACTG		
D-56.0F	CATGCTCTCTACATAATGCCC		486 bp
D-55.5R	CCAATACTTCCATGGTTGGC		
D-44.3F	ATTAACAGGGATGCGCATGC		2187 bp
D-42.1R	GTCATTGGTACTTGCCTGGG		
D-42.1F	TCACTGACTGCGCAGTTCCT		2139 bp
D-39.9R	CACCATGATTCGTCCCCTA		
D-31.7F	AAACAAAACCAAAGCCCCC		3619 bp
D-28.1R	CTTTCCTGGGGAGCTTTTTTC		

Primers for complete re-sequencing around *DPB1* region of LD breakdown

D-74.2F	CCTCTCTCAACCAATTACCTCAGC		3830 bp	
D-70.4R	CTGATTCCACAATTCCCTGCATGC			
Sequencing primers				
D-73.4R	CGGGGAGGAGACTGAGATAC		3588 bp	
D-73.1F	CCGGCACTAAGGTCCCTTAG			
D-72.7F	GGCGCCACTCTCATCTAATC			
D-72.2F	GCGCTCTCTTCTCGACCAG			
D-71.2R	CTCTGCCACCAGCTGTGCAA			
D-70.7R	GCTTATGTCTACTTGGCAGG			
D-70.7F	CCTGTTTGTGGAACCCTTAGC			
D-67.2R	CAGATGCTTTGAAGGAGGTG			
Sequencing primers				
D-70.2F	TTCTGCTCTCTGCTCCCCCA			
D-69.4R	CTTAATCCAGTCTATCACTG			
D-69.0F	CAAACCTGCACGTTGTGCAC			
D-68.9R	TCTCTCAGGATACTCTCAGG			
D-68.1R	CTCCATGATGGTTCTGCAGG			
D-67.7R	TCCTGTGGGTGTGTGAGGTG			
D-67.2F	CACCTCCTTCAAAGCATCTG		1370 bp	
D-65.9R	GAGAATTTTAGGTACCAGCG			

Primers for complete re-sequencing around *DPA* region of LD breakdown

D-56.0F CATGCTCTCTACATAATGCC 5014 bp

D-51.0R CCATTAGTGTATAGTCTCTGGC

(includes two ~1800 bp long polymorphic SVA elements)

Sequencing primers

D-55.5F GCCAACCATGGAAGTATTGG

D-55.1F CAGAATACATTCTCGGACCAC

D-54.6R GTCATAATACTTTTAATGTCTG

D-54.5R TATGGAGAAAGCTTTACACC

D-53.9R CTGACCTGGAAACCATTGG

D-53.5R CCCAAAATTGAAAATTTTCAG

D-53.3F GCTGGACATTCTGTGAACAG

D-53.2R GCCAGGACCATTATTGATCA

D-52.9R GAGAGAAGATATTCTGGTGGC

D-51.2F AACTGGATGCCAGATGAGCA 2551 bp

D-48.6R TGCAAGATCATCACTCCTGG

Sequencing primers

D-50.8F AAGGAAGAGAGGGCTATGGG

D-50.5F TCTGTACCACCTGGACAGAG

D-50.1F GCTCAATTGCTCTTATTGTG

D-49.3R TATTCCCATGTGCCAACAGG

D-48.9R CTAGGGTTAACACCTTTTGTG

D-48.6F GGGTCTAACCAGGAGTGATG 4405 bp

D-44.3R TACAACCCGTAAGCATGCGC

Sequencing primers

D-48.1F GCCTATGTATCTTTTGGGTC

D-47.8F GGTTATAGTGTATTTGCCAC

D-47.5F CCTGATTTGAAGCTACCTGC

D-47.1F GGATCAATACCCTACATGAG

D-46.8F ACCACACCATACTACTGC

D-45.9R ATTAAGTCCACTCTGGAGCC

D-45.5R CAGTCAGTTGCTAGGCTATG

D-45.2R GGCAATATAGCCCTGAATTC

D-44.7R AATGAAAGGCCAGCTTGTTG

RING3-TAP2 interval

Amplicon 1

R53.4F GCTGTATTGCTACAGGAGTGG 2174 bp

R55.6R GCTCACCATGCCATACATCC

Sequencing primers

R53.9F GTTCCTCCTATTCCGCATCC

R55.0R CAGAGTAGAATCTTGAGTGG

Amplicon 2

R56.9F GCACACAGTGTACTGCCAGG 2514 bp

R59.4R AGTTATATGTTTCCTCAGGGC

Sequencing primers

R57.4F CACTTGTGCTACTATGCCCA

R58.9R ACACACACCCCAGGGATGTC

Amplicon 3

R60.7F TATCCTCTCTAGTCTGGTGC 2493 bp

R63.2R TGGCCCTAGCCAGTAACAGT

Sequencing primers

R61.3F GTCTCCATTCTTCACCACC

R62.6R CGTCAATTGGGGAGTGAGAA

Amplicon 4

R65.4F GTCGTGCTCACGTCTTTTGC 2152 bp

R67.5R CCTATATTCAGCTCAGAACC

Sequencing primers

R66.0F CAAGGAGTCTGAGAGGTGTT

R67.0R TAGGAATGTATACCAGGGAG

Amplicon 5

R68.7F CCTTCCCTGGCTCCTCTAGG 2400 bp

R71.1R CCACTGGCCCCTGAGCATGC

Sequencing primers

R69.0F AGGCCAGGGAGGATCCCACA

R70.2R CATGTTTAGGAAGGAGGGTG

R70.5R GACTGGAGTAAGTGTATGGC

Amplicon 6

R73.3F AGGCACATAGCCAGCCTGCA 2178 bp

R75.5R TGCCAAGGATACCACCTCCC

Sequencing primers

R73.8F CTGAAGACCCTCATGCTGGT

R74.4F AGATTACAGGCGTGCACCAC

Amplicon 7

R75.8F GGTGCTTCCTGCCACTACTT 2388 bp

R78.2R ACCCTCCTGTTTGACACAGC

Sequencing primers

R76.3F AGGGATTAGAAACTGACC

R77.7R CTCATTAGCATGGCTACTAT

R77.0F AGAACTCCTTTTCACCTTGC

R77.1R AGTGATAGTAGCCAGGGGCT

Amplicon 8

R81.1F CCGCTCAGCTATCCACTACT 2321 bp

R83.4R CCACATTTACAGGACACTCC

Sequencing primers

R81.7F CCAGGTTCAAGCGATTCTTC

R83.1R TTCAATCATCACTGTCCCTCC

Amplicon 9

R85.2F GTGGGACACAATCCAACCTG 2366 bp

R87.5R AAGAGACTGGGAGTGGCTGG

Sequencing primer

R87.1R TGTACAAGTAACCTGTGAG 3'

Amplicon 10

R104.1F CTGTGTAAACACATGGTGTC 2337 bp

R106.5R CGATGGTCAGGAAAGGAAGC

Sequencing primers

R104.7F GCAGATTGGCTTGTACGACT

R106.2R TCTACCCAAGTTGTGACAAC

Amplicons for complete re-sequencing around *DMB* region of LD breakdown

Amplicon A

R65.9F ACACACAGGGTTGTTTAGAG 2783 bp

R68.8R CAGCATCATCCAACAGACAG

Sequencing primers

R67.3F GCTACTCCTATCCCTGTGCT

R68.3R CATGCCTACATGATGAAGCC

Amplicon B

R70.5F GACTGGAGTAAGTGTATGGC 2997 bp

Appendix 1: PCR and Sequencing Primers

R73.5R TCTGGAGTTAAGCTCTGTCC
Sequencing primers
R71.0F AGTAAAGCCACGGCTGGCAG
R71.1R CCACTGGCCCTGAGCATGC
R71.2Rs CCAAAGCTGTGTGCAACTC
R71.5F ACCGGCCAAGGAAGTGCAG
R72.7R CTGCCTCTAGGAAATGTGCC
R73.2R TGCCCATACTCTGTGACTCC

Amplicon C

R75.4F GAGTTCCTTCAACACTTGGG 2832 bp
R78.2R ACCCTCCTGTTTGACACAGC
Sequencing primers
R75.5Fs TTCCCTGAGGGAGAGCATGT
R76.3F AGGGATTAGAAACTGACC
R77.0F2 GCAACCAATCTCCAGAACTC
R77.0F AGAACTCCTTTTCACCTTGC
R77.1R AGTGATAGTAGCCAGGGGCT
R77.7R CTCATTAGCATGGCTACTAT

Amplicon D

R78.1F GTATCATCTGCGAGCATTGT 5336 bp
R83.4R CCACATTTACAGGACACTCC
Sequencing primers
R78.6F CTGGAGATCATAACAAGGTTG
R79.2F AGCCAGGATGGTCTCAATCT
R79.5F TCTGAGATGGAGTCTTACTC
R80.6R AACTTCATCTCAAGAGCCC
R81.1F CCGCTCAGCTATCCACTACT
R81.2Rs TAAGCCAGGTACAGAAAGAC
R81.6R GTGAGTATGTGGAGAAGGCT
R81.7F CCAGGTCAAGCGATTCTTC
R83.1R TTCAATCATCACTGTCCCTCC

Appendix 2. ASO sequences

DPA2-DOA interval

ASO	5'	3'	ASO	5'	3'
-100.49G	CCAGGAAGCCTTCCCTGA		-73.49G	AGGATCCGGTCCTTTGCT	
-100.49A	CCAGGAAACCTTCCCTGA		-73.49A	AGGATCCAGTCCTTTGCT	
-99.65G	TGCATGCGGATGCCCTCA		-73.48T	CCTTTGCTCTTCCCCCTG	
-99.65A	TGCATGCAGATGCCCTCA		-73.48C	CCTTTGCCCTTCCCCCTG	
-97.62C	CCACCTACGCACACATCC		-73.47C	TGGGTCCCCTCTCCCTGG	
-97.62T	CCACCTATGCACACATCC		-73.47T	TGGGTCTCTCTCCCTGG	
-97.55T	TGCTTAATGACTCTCCCC		-73.41C	GCAGACCCTTCTGTCTAC	
-97.55C	TGCTTAACGACTCTCCCC		-73.41T	GCAGACCCTTTTGTCTAC	
-96.42G	AAGAACTGTATGTTTTGC		-73.40C	TGTCTACCTGCCCATCAG	
-96.42A	AAGAACTATATGTTTTGC		-73.40T	TGTCTACTTGCCCATCAG	
-82.49A	TTGATTTTATCTTCTCTA		-73.36C	CCAGCTACCTCCCTGYAG	
-82.49G	TTGATTTTGTCTTCTCTA		-73.36A	CCAGCTAACTCCCTGYAG	
-82.37G	TTTAACAGTCTGACATTC		-73.29T	ACCCTTGTGCGCTGTCCT	
-82.37A	TTTAACAATCTGACATTC		-73.29C	ACCCTTGCGCGCTGTCCT	
-74.16A	CTCACTCATAGCAGGTGA		-73.22G	AGCTCCCGGGGGTCTCTC	
-74.16C	CTCACTCCTAGCAGGTGA		-73.22C	AGCTCCCCGGGGTCTCTC	
-74.02A	GGGACTTGAGGGAGTAGT		-73.20T	TCTGCTCTCGTCCCTCCTC	
-74.02G	GGGACCTGAGGGAGTAGT		-73.20C	TCTGCTCCCGTCCCTCCTC	
-74.02A	GGGACTTGAGGGAGTAGT		-73.04G	CGGGCCCGGGGGCTGOC	
-74.02G	GGGACCTGAGGGAGTAGT		-73.04A	CGGGCCACGGGGCTGOC	
-74.01T	ARGTCCCTCAGGTTGGTG		-72.94C	GTCTGCACATCCTGTCCG	
-74.01A	ARGTCCCAAAGGTTGGTG		-72.94T	GTCTGCATACCCTGTCCG	
-73.85G	ATGAACTGAGACACTGCC		-72.92T	GCCCGCTTCTCCTCCAGG	
-73.85C	ATGAACTCAGACACTGCC		-72.92C	GCCCGCTCCTCCTCCAGG	
-73.83A	CCCTGTGACAGCCGTGGG		-72.91T	TCCAGGATGTCTTCTGG	
-73.83C	CCCTGTGCCAGCCGTGGG		-72.91G	TCCAGGAGGTCTTCTGG	
-73.57T	CTCACACTGCACTTGGGT		-72.89C	CTGTTCCAGTACTCCKCA	
-73.57C	CTCACACCGCACTTGGGT		-72.89G	CTGTTCCAGTAGTCCKCA	
-73.56T	GGGTTTTTTTGGTCTCTT		-72.88G	TCCGCAGCAGGCCGCCCC	
-73.56A	GGGTTTTATTGGTCTCTT		-72.88T	TCCTCATCAGGCCGCCCC	

DPA2-DOA interval (continued)

ASO	5'	3'	ASO	5'	3'
-72.83G	GAAGCGCGGWACTCCTC		-71.30A	TTTGAGTATCTTCAGGCC	
-72.83A	GAAGCGCACGWACTCCTC		-71.30G	TTTGAGTGTCTTCAGGCC	
-72.82A	GCGCRCGAACTCCTCCCG		-71.19A	CACCCCTAGCAGAAGTTG	
-72.82T	GCGCRCGTACTCCTCCCG		-71.19T	CACCCCTTGCAGAAGTTG	
-72.75C	CCAGGGACGRCAGGAATG		-71.04C	CCTTGGGCCCCAGACCCT	
-72.75A	CCAGTTACGRCAGGAATG		-71.04G	CCTTGGGCGCCAGACCCT	
-72.74A	TGGAAAAGGTAATTCTCT		-71.03A	AGACCCTAAAATTCCAGG	
-72.74C	TGGTACACGTAATTCTCT		-71.03G	AGACCCTGAAATTCCAGG	
-72.66T	TCTCTCTTAATATTAACG		-71.01A	GGCCCAGAGCCCCCAGC	
-72.66C	TCTCTCTCAATATTAACG		-71.01G	GGCCCAGGGCCCCCAGC	
-72.65C	ATATTAACGATTCTTCCC		-70.95G	CTGTTCAAGTTGCCCACT	
-72.65G	ATATTAAGGATTCTTCCC		-70.95A	CTGTTCAATTGCCCACT	
-72.63T	TCTTCCCCTCAGGGCTG		-70.69+	CATTTTTTTTTGTATATA	
-72.63C	TCTTCCCACCCAGGGCTG		-70.69-	CATTTTTTTTTGTATATA	
-72.47A	GAGCACTAAGGGGAAGGC		-70.42A	GGAATCAACCCTGCCAC	
-72.47G	GAGCACTGAGGGGAAGGC		-70.42G	GGAATCAGCCCTGCCAC	
-72.26T	TCCTTCTTTCTCTTTTCC		-70.37A	TCCAGCAAATTTTGATTG	
-72.26C	TCCTTCTCTCTCTTTTCC		-70.37G	TCCAGCAGGTTTGTATTG	
-72.18G	TTTGCATGTAGGARATGT		-70.25T	CTTTGCCTGTCCTCCCTT	
-72.18T	TTTGCATTTAGGARATGT		-70.25C	CTTTGCCGTCCTCCCTT	
-72.06A	GATGTGGAAAGTATTTTA		-70.23T	TTCACTGTCCAGSATCAC	
-72.06G	GATGTGGGAAGTATTTTA		-70.23C	TTCACTGCCCAGSATCAC	
-71.93A	CGTGGACATGAACTTTCT		-70.21T	CACATGCTCCCTTCTGCT	
-71.93G	CGTGGACGTGAACTTTCT		-70.21C	CACATGCCCCCTTCTGCT	
-71.90C	GTGCACCCTGGTACTGGG		-70.18G	CTTGGTGGCTCCCAACAC	
-71.90G	GTGCACCGTGGTACTGGG		-70.18A	CTTGGTGACTIONCCAAACAC	
-71.75A	GCTGTGGACCAGAACTT		-70.11T	CCGTCACTTCCTCCTGAC	
-71.75G	GCTGTGGGCCAGAACTT		-70.11C	CCGTCACTTCCTCCTGAC	
-71.71T	AGCATGTTTTTTTTGTTG		-69.98C	GTGACTACAAGTTCAAGT	
-71.71G	AGCATGTGTTTTTTGTTG		-69.98T	GTGACTATAAGTTCAAGT	

DPA2-DOA interval (continued)

ASO	5'	3'	ASO	5'	3'
-69.84T	GTACCTTTTAGGTGTGTC		-68.21T	ATGGCAGTTCGGCTCTTA	
-69.84C	GTACCTTCTAGGTGTGTC		-68.21C	ATGGCAGCTCGGCTCTTA	
-69.79A	CAGGAAAAGTGGAAGTCA		-67.80G	TGTTTGGGGAGATTATGC	
-69.79G	CAGGAAAGGTGGAAGTCA		-67.80C	TGTTTGGCGAGATTATGC	
-69.56G	CAAGACAGTGTGGCAATT		-67.19T	CTGATTCTATTTTGTTTA	
-69.56T	CAAGACATTGTGGCAATT		-67.19C	CTGATTCTACTTTGTTTA	
-69.54A	TCCTCAAAGATCTAGAAC		-66.77A	GTTTAATATGAGGTTTAA	
-69.54G	TCCTCAAGGATCTAGAAC		-66.77G	GTTTAATGTGAGGTTTAA	
-69.50A	TTACTGGATATATACCCA		-66.69G	TATGCAAGGTAAGGCTAC	
-69.50G	TTACTGGGTATATACCCA		-66.69C	TATGCAACGTAAGGCTAC	
-69.43A	TGTTTATAGCGGCACTAT		-66.65T	AACGTAATAATGTACTT	
-69.43T	TGTTTATTGCGGCACTAT		-66.65C	AACGTAACATATGTACTT	
-69.27C	AGCTGGAAACCATCATTC		-66.07A	AAACATCAAAGATCATCT	
-69.27T	AGCTGGAAATCATCATTC		-66.08G	AAACATCGAAGATCATCT	
-69.23A	AAGGACAAAAACCAAAG		-61.37A	ATTGACAAGTTCTTCCCA	
-69.23G	AAGGACAGAAAACCAAAG		-61.37G	ATTGACAGTTCTTCCCA	
-69.02C	ACATGTACCCTAGAACTT		-61.19T	AGAGGACTTCTATGACTG	
-69.02T	ACATGTATCCTAGAACTT		-61.19G	AGAGGACGTCTATGACTG	
-68.68A	CCAAGGGATGTGGGCTCC		-60.29T	TGGTGGGTGCCTGTA ACT	
-68.68G	CCAAGGGGTGTGGGCTCC		-60.29A	TGGTGGGAGCCTGTA ACT	
-68.54C	ATGCAGACATGAACACGC		-59.15C	AGGGTTGCTGGAGATGCA	
-68.54T	ATGCAGATGTGAACACGC		-59.15T	AGGGTTGTTGGAGATGCA	
-68.47A	GTTATCCAGGACCTCTTG		-55.89T	TTGTTTGTGGGGACATGG	
-68.47T	GTTATCCTGGACCTCTTG		-55.89C	TTGTTTGCGGGGACATGG	
-68.45+	CCATTTTCCCAAAGACA		-55.74T	ACCAGGGTCTGTCAGGGC	
-68.45-	CCCATTTTCCCAAAGACA		-55.74C	ACCAGGGCCTGTCAGGGC	
-68.44+	CCAAAGACAAGCATACTG		-55.70A	ACATACCAAATGTATGCA	
-68.44-	CCAAAGACACTGACCACT		-55.70G	ACATACCGAATGTATGCA	
-68.35G	CTATAGGGGATGAACCTC		-55.65A	TTGATAAGTGCAGCAAAC	
-68.35A	CTATAGGAGATGAACCTC		-55.65G	TTGATAGGTGCAGCAAAC	

DPA2-DOA interval (continued)

ASO	5'	3'	ASO	5'	3'
-55.60C	CATTCTGCACATGTATCC		-53.09G	AACCTTCGATATTKCCAG	
-55.60T	CATTCTGTACATGTATCC		-53.09A	AACCTTCAATATTKCCAG	
-54.64A	AGATAAGACTCAAATCCC		-53.08T	ATATTTCCAGCAGCTAYA	
-54.64G	AGATAAGGCTCAAATCCC		-53.08G	ATATTGCCAGCAGCTAYA	
-54.43C	ACTAACCCAACCTTAACAT		-53.03C	AGGACAGCAAGAAGGCAT	
-54.43T	ACTAACCTAACCTTAACAT		-53.03T	AGGACAGTGAGAAGGCAT	
-54.17C	GAGGGCTCCCCAATTCAT		-53.02A	TGACACCAGACCACACTT	
-54.17G	GAGGGCTGCCCAATTCAT		-53.02C	TGACACCCGACCACACTT	
-53.94C	GATTAATCCATAGTATTC		-51.22A	AACAAAGATTTTTTTCAT	
-53.94T	GATTAATTCATAGTATTC		-51.22C	AACAAAGCTTTTTTTCAT	
-53.81BC	TAAYTTACAAGTCCAATA		-51.13T	TGGTTTGTARAGAAGAGA	
-53.81BT	TAAYTTATAAGTCCAATA		-51.13C	TGGTTTGCARAGAAGAGA	
-53.68A	AAGGATGATATAACATGA		-51.08C	GTAATGACGGAACATCCA	
-53.68T	AAGGATGTTATAACATGA		-51.08T	GTAATGATGGAACATCCA	
-53.58+	GAATACAAAAGAAAAATT		-50.94G	GAGTGTAGAGAGATGAGC	
-53.58-	GAATACAAAGAAAAATTA		-50.94A	GAGTGTAAGAGATGAGC	
-53.55G	AATGGATGCAGAAAAAGG		-50.84G	CCTCAGCGTTTTGGGGTA	
-53.55A	AATGGATACAGAAAAAGG		-50.84A	CCTCAGCATTTTGGGGTA	
-53.42+	GGGGGGCGGGCATTTTAC		-50.76T	TCACTTATAACTATCTGG	
-53.42-	GGGGGGGAGCATTTTACTT		-50.76C	TCACTTACAACTATCTGG	
-53.38T	TATCTAATGTGAAAGTGG		-50.75+	CAACTGGATAAGATCCAA	
-53.38C	TATCTAACGTGAAAGTGG		-50.75-	ACAAGTGGATAAGACCAA	
-53.27+	ACATTTATTGAAAAAGAA		-50.74+	CCAATAGTTTGATATAC	
-53.27-	ACATTTATTTGCTTGATC		-50.74-	CCAATAGTTTGATATAC	
-53.21-	TTTTTCCTAAKGGCTGTT		-50.73C	GTTTGATATACCTGTCAA	
-53.21+	TTTTTCCTAAKGGCTGT		-50.73T	GTTTGATATACTTGTCAA	
-53.17G	AAAAATTGACCAATGCTG		-50.72AG	TGTCAAGGGAYTTAATAT	
-53.17A	AAAAATTAACCAATGCTG		-50.72AA	TGTCAAAGGAYTTAATAT	
-53.11T	AATTAGAGTTTACACAAA		-50.72BC	CAARGGACTTAATATAAC	
-53.11C	AATTAGAGTTCACACAAA		-50.72BT	CAARGGATTTAATATAAC	

DPA2-DOA interval (continued)

ASO	5'	3'	ASO	5'	3'
-50.69C	ACATTAGCCTGCATTAAT		-49.62-	ATCAGAGTGTACACGTGGA	
-50.69T	ACATTAGTCTGCATTAAT		-49.62+	ATCAGAGGTGTACACGTGG	
-50.66+	TTCTGTAATAAGGAAGGT		-49.58T	GGAAGGTTACATCYTCTG	
-50.66-	TTTCTGTATAAGGAAGGT		-49.58C	GGAAGGTCACATCYTCTG	
-50.63G	ATCCCTGCTTGAACTTT		-49.57T	ACATCTTCTGCTGTTGAG	
-50.63T	ATCCCTTCTTGAACTTT		-49.57C	ACATCCTCTGCTGTTGAG	
-50.60AG	AAGCTCTGTAYACTCAGA		-49.54A	GAAGTGTATGCCAACTAG	
-50.60AC	AAGCTCTCTAYACTCAGA		-49.54G	GAAGTGTGTGCCAACTAG	
-50.60BC	CTCTSTACACTCAGACAT		-49.49T	AGCACTCTGTGCAGCCAC	
-50.60BT	CTCTSTATACTCAGACAT		-49.49C	AGCACTCCGTGCAGCCAC	
-50.53+	AAAACAGGAGTAAAATAG		-49.43A	GATAAATAGATATGAATA	
-50.53-	GAAAACAAAGTAAAATAG		-49.43G	GATAAATAGGTATGAATA	
-50.48T	GTTTGGTTTGGTGAAGTG		-49.38A	AGGAGAAAACATAGAAAT	
-50.48C	GTTTGGTCTGGTGAAGTG		-49.38G	AGGAGAAGACATAGAAAT	
-50.34G	AGTGTACGCACGTATCTC		-49.34A	GTTAGAGAAGAATAAAAC	
-50.34T	AGTGTACTCACGTATCTC		-49.34G	GTTAGAGAAGAGTAAAC	
-50.32-	TACATGATGATTCCTTTT		-49.33A	AAAACAGAATATAAGTAC	
-50.32+	TACATGATTGATTCCTTT		-49.33G	AAAACAGGATATAAGTAC	
-50.27A	CGTGCAGATGGGGCTTAT		-49.32T	AGTCTCCTTACCTGTTGG	
-50.27G	CGTGCAGGTGGGGCTTAT		-49.32A	AGTCTCCATACCTGTTGG	
-50.24G	AGCACTCGTGTATGACAG		-49.28A	AAGGAAAACCTGCTTGGT	
-50.24A	AGCACTCATGTATGACAG		-49.28G	AAGGAAAGCCTGCTTGGT	
-50.19G	GATGAGAGAACTTCAAAC		-49.26G	CCTTGGAGCTTCAGGTTG	
-50.19C	GATGAGACAACCTTCAAAC		-49.26A	CCTTGGAACTTCAGGTTG	
-49.91C	AAAGGGACTGTTGAACAC		-49.23A	AAGGAAAACCTGCTTGGT	
-49.91T	AAAGGGATTGTTGAACAC		-49.23G	AAGGAAAGCCTGCTTGGT	
-49.70G	AGTTGCAGAGAAGGGAGA		-49.17G	TCTCAGAGGCCTGTTGAG	
-49.70C	AGTTGCACAGAAGGGAGA		-49.17T	TCTCAGATGCCTGTTGAG	
-49.66+	CAGTTCTGTTTCATAGTG		-49.13G	TTTGTTAGATCCACTGTG	
-49.66-	CAGTTCTTTTCATAGTGG		-49.13C	TTTGTTACATCCACTGTG	

DPA2-DOA interval (continued)

ASO	5'	3'	ASO	5'	3'
-49.02C	CTGGAACGGTTAGAGAA		-47.93G	AATGGAAGTATCAGTTTC	
-49.02T	CTGGAATGGTTAGAGAA		-47.93A	AATGGAAATATCAGTTTC	
-48.81C	TGCACACCCTCTTCTGTT		-47.89A	GGGTAGTATAAATTAGGC	
-48.81A	TGCACACACTCTTCTGTT		-47.89G	GGGTAGTGTAATTAGGC	
-48.83T	TGTACAGTTGTGTGCAGC		-47.31G	AGTGTGTCTGGCACCTTA	
-48.83A	TGTACAGATGTGTGCAGC		-47.31T	AGTGTTTCTGGCACCTTA	
-48.77T	TCGCTTTTATTCCGGCTT		-47.04C	ATGGTGTCTGTGTCCAGA	
-48.77C	TCGCTTTCATTCCGGCTT		-47.04T	ATGGTGTTTGTGTCCAGA	
-48.71T	TTAGCAATATTCTTTTCC		-46.89A	CGGGTTGAGTCCTAGCCA	
-48.71C	TTAGCAACATTCTTTTCC		-46.89G	CGGGTTGGGTCCTAGCCA	
-48.65C	CCAGGAGCGGGGTCTAA		-46.22G	TTTTGAAGGAGGGGACAT	
-48.65T	CCAGGAGTGGGGTCTAA		-46.22A	TTTTGAAAGAGGGGACAT	
-48.53A	TCATGAAATGYTCCCCAG		-46.06C	TGAAGGTCTCCTTGATCA	
-48.53T	TCATGAATTGYTCCCCAG		-46.06G	TGAAGGTCTGCTTGATCA	
-48.47T	TGATTGATTCTATGATAG		-46.05C	TTGATCACTGCCATAAGC	
-48.47C	TGATTGACTCTATGATAG		-46.05G	TTGATCAGTGCCATAAGC	
-48.45G	AAGTGACGCAGATATGTT		-46.04G	TAAGCCAGTGGGCCAGGC	
-48.45A	AAGTGACACAGATATGTT		-46.04A	TAAGCCAATGGGCCAGGC	
-48.41G	CAGTTTAGTTGCTGTTAG		-46.02A	CCAGGCAGAGAGCTGTGG	
-48.41A	CAGTTTAATTGCTGTTAG		-46.02T	CCAGGCAGTGAGCTGTGG	
-48.40A	TTAGAATACCAGTGCATC		-46.01A	GCTGTGGACTCGAATGTG	
-48.40G	TTAGAATGCCAGTGCATC		-46.01G	GCTGTGGGCTCGAATGTG	
-48.35G	CTCAAGGGCCTCAAGGGC		-45.98C	CATTGATCTAGCAATTGC	
-48.35A	CTCAAGGACCTCAAGGGC		-45.98T	CATTGATTTAGCAATTGC	
-48.29A	TAGACACATTTCTGGCCA		-45.89A	GATTTTGCAATAAATAAA	
-48.29G	TAGACACGTTTCTGGCCA		-45.89G	GATTTTGCGATAAATAAA	
-47.97C	GCCTGTTCTGCACCTGG		-45.71A	TCATCAGAAAACAATGTT	
-47.97T	GCCTGTTTCTGCACCTGG		-45.71T	TCATCAGTAAACAATGTT	
-47.95C	ACAAATGCGGAGTTTTGG		-45.67G	TGGAGGAGTGAACCTATTT	
-47.95A	ACAAATGAGGAGTTTTGG		-45.67A	TGGAGGAATGAACCTATTT	

DPA2-DOA interval (continued)

ASO	5'	3'
-45.66A	TTTTTGTAGGCAAGATCA	
-45.66T	TTTTTGTTGGCAAGATCA	
-45.12A	TGGGGCAAAGCCAGTTGA	
-45.12G	TGGGGCAGAGCCAGTTGA	
-44.90C	GTATTGGCTCTGTTGGGG	
-44.90G	GTATTGGGTCTGTTGGGG	
-44.50A	CAAGGACAAATAGATGGT	
-44.50T	CAAGGACTAATAGATGGT	
-44.21G	GTTAGGGGGTGCATGCTC	
-44.21A	GTTAGGGAGTGCATGCTC	
-44.16C	TGCAGACCATTTTTTATT	
-44.16A	TGCAGACAATTTTTTATT	
-43.56A	CCAGCCCATCATGAAGTT	
-43.56G	CCAGCCCGTCATGAAGTT	
-43.25T	GCCTTTCTTCTGATTGCC	
-43.25C	GCCTTTCCTCTGATTGCC	
-43.07T	AAGATCCTTGCTAAACTC	
-43.07C	AAGATCCCTGCTAAACTC	
-42.89G	TTAGCTTGTTGTTACAC	
-42.89A	TTAGCTTATTGTTACAC	
-42.05C	TTCCTCACGCGGGGCACC	
-42.05T	TTCCTCATGCGGGGCACC	
-41.27G	AAATCCAGGAGCATATAG	
-41.27A	AAATCCAAGAGCATATAG	
-40.61G	GTCTCCTGTCTGCCATCT	
-40.61A	GTCTCCTATCTGCCATCT	
-40.37G	CAAAGAGGTGAAAGATCC	
-40.37A	CAAAGAGATGAAAGATCC	
-40.32A	TACAAGGAAAACACTACAAA	
-40.32C	TACAAGGCAAACACTACAAA	
-30.23A	TTGTACAATATTAATGTT	
-30.23G	TTGTACAGTATTAATGTT	
-28.13A	CTCAGGTATTCTCTTATA	
-28.13G	CTCAGGTGTTCTCTTATA	

RING3-TAP2 interval

ASO	5'	3'	ASO	5'	3'
H6C	CCTTGCCCGAATTTGTAG		J1A	CCTTTRGAAATGGAGTGT	
H6A	CCTTGCCAGAATTTGTAG		J1T	CCTTTRGTAATGGAGTGT	
H8C	CTGATGACGTGACTATGT		J2A	GATGTAGAAGGTACACAA	
H8T	CTGATGATGTGACTATGT		J2G	GATGTAGGAGGTACACAA	
H9C	GCTTCTACGTCTAGCTGC		J3A	GGCTTTGAGGTAAAGGAA	
H9T	GCTTCTATGTCTAGCTGC		J3G	GGCTTTGGGGTAAAGGAA	
H10C	GGATCTCCTCTTAGGGTA		J4A	GAGATGGATTCCCCAGGC	
H10T	GGATCTCTTCTTAGGGTA		J4G	GAGATGGGTCCCCAGGC	
H11A	TTCCCCCAACAAGGCAAT		J13C	TTAGCAACTTGGGGGACC	
H11C	TTCCCCCACAAGGCAAT		J13T	TTAGCAATTTGGGGGACC	
H12A	ATATCTGATTTTGGTTTA		J14C	GAACACCCCCAGGCACAT	
H12G	ATATCTGGTTTTGGTTTA		J14G	GAACACCGCCAGGCACAT	
H13C	TATAATTCAGGGATCATA		JJK1C	GCAGACCCCTCATACCCC	
H13G	TATAATTGAGGGATCATA		JJK1T	GCAGACCTCTCATACCCC	
H22A	AGAGGCAATAAAAGAAGG		JJK2C	ATTTCTTCCATACCCTGA	
H22G	AGAGGCAGTAAAGAAGG		JJK2T	ATTTCTTTCATACCCTGA	
H23A	ATAATTAACAAATAAGTA		JJK13G	GTCAGTGGAAAGGAACCTTA	
H23G	ATAATTAGCAAATAAGTA		JJK13A	GTCAGTGAAAGGAACCTTA	
JJH12G	CAGACAAGGGTTTTTGTA		JJK15T	ACCACCATGCCTGGCTAA	
JJH12C	CAGACAACGGTTTTTGTA		JJK15C	ACCACCACGCCTGGCTAA	
JJ1G	GAAAGACGGAGGATTGAA		JJK16G	AGGGATGGGGTTTCACCA	
JJ1A	GAAAGACAGAGGATTGAA		JJK16C	AGGGATGCGGTTTCACCA	
JJ3T	CAATCTCTTTAGTTTGAC		JJK3G	CAAAGTGGCAGGATCCCA	
JJ3A	CAATCTCATTAGTTTGAC		JJK3A	CAAAGTGACAGGATCCCA	
JJ4T	CAGTTTATGTACTTGAAA		JJK5C	AGAGATTCCCAKTGCAGG	
JJ4C	CAGTTTACGTACTTGAAA		JJK5T	AGAGATTTCCAKTGCAGG	
JJ5C	AGCAGTGCGCACAAGACT		JJK9C	TCCCCTTCCAGGTAAGGA	
JJ5A	AGCAGTGAGCACAAGACT		JJK9T	TCCCCTTTCAGGTAAGGA	
JJ6T	GGTGTGGTTGATGTTGCT		JJK4G	AGGTAAGGAATAGGGAAA	
JJ6C	GGTGTGGCTGATGTTGCT		JJK4A	AGGTAAGAAATAGGGAAA	
JJ7A	AATGAGGATGTAAATTTG		JJK18G	TAGGTCTGCTAGGATCCA	
JJ7C	AATGAGGCTGTAAATTTG		JJK18C	TAGGTCTCCTAGGATCCA	
J8A	TCCCTTTAGWAATGGAGT		JJK14G	CCAGAAAGTCARAATACT	
J8G	TCCCTTTGGWAATGGAGT		JJK14A	CCAGAAAATCARAATACT	

RING3-TAP2 interval (continued)

ASO	5'	3'
JJK10C	AGGAGTGCGTGTGTGTGT	
JJK10T	AGGAGTGTGTGTGTGTGT	
JJK11+	GTGAGAGAAAGAGAGGAA	
JJK11-	GTGAGAGAGAGAGGAAAA	
JJK12+	AAAAGAGAAGGAGTGGAA	
JJK12-	AAAAGAGGAGTGGAAAGAG	
JJK6A	GAGAGGAAAGAAGTGGAG	
JJK6G	GAGAGGAGAGAAGTGGAG	
JJK7C	TTTAAAGCTCATCTAGTC	
JJK7G	TTTAAAGGTCATCTAGTC	
JJK19T	TCAACACTTGGGGCAGAT	
JJK19C	TCAACACCTGGGGCAGAT	
JJK8C	GCGTTGACCCACATTTG	
JJK8G	GCGTTGAGCCACATTTG	
J5C	AGGATTCCAAGAGGCCCC	
J5T	AGGATTCTAAGAGGCCCC	
J6G	AGAGGCCGGGACAGAAG	
J6A	AGAGGCCAGGGACAGAAG	
J7T	CCCTCTTCCCTCTGCTCT	
J7C	CCCTCTCCCTCTGCTCT	
JK4T	GGGACTATAGGCGCCCGC	
JK4C	GGGACTACAGGCGCCCGC	
KK3G	GATAAACGCTATTTGTTT	
KK3C	GATAAACCTATTTGTTT	
KK4A	TAGTTACACATTTACTGT	
KK4T	TAGTTACTCATTTACTGT	
KK5G	GTATTCTGTTCCCTCTTTA	
KK5C	GTATTCTTCCCTCTTTA	
K5A	TGTAGAACTGCATTGTT	
K5G	TGTAGAAGCTGCATTGTT	
K7C	GAGTTCYCGTATGTCCCC	
K7T	GAGTTCYTGTATGTCCCC	
K8G	GATTCAGGTATTTCCCTG	
K8A	GATTCAGATATTTCCCTG	

Appendix 3. PCR primers and ASOs used in population study (Chapter 9)

PCR amplicons and primers

Amplicon 1 4.4 kb

R4.9F 5' CTCAGAAGCTTATAGGATATCTGC
R9.3R 5' CCCTTCCTTCCTAAAAGTAGCC

Amplicon 2 3.9 kb

R9.3F 5' TGAGACCATGTCTCCAAAGG
R13.2R 5' CTTATTCTTGGTGTTCCTGGGATG

Amplicon 3 4.7 kb

R13.1F 5' CTCATCCATAGGCTCAAACC
R17.8R 5' GGTCCAGTTCATCCAAGTTG

Amplicon 4 5.7 kb

R19.5F 5' TGCGGGGCTCTGTAGGGTCT
R25.2R 5' CCACCTGCTACTGTTTGCCAAG

Amplicon 5 4.4 kb

R25.1F 5' CAAGAGCTTTCCTGGTGCAC 3'
R29.5R 5' TGCTCTGTCCCAGGGTCC 3'

Amplicon 6 5.0 kb

R30.7F 5' GCACCTAGGCTCCCATCAC 3'
R35.7R 5' CTGCCTGGTACTGACACC 3'

Amplicon 7 6.7 kb

R37.4F 5' CCCCCTAGCGGCCCAACC 3'
R44.1R 5' TGTCCCCACACAGCACC 3'

Amplicon 8 4.7 kb

R48.9F 5' GTGAGAGAGGCTCTCCCAGC 3'
R53.6R 5' GCTACGCCACAAGGAGGAGC 3'

Amplicon 9 3.4 kb

R55.5F 5' GTGAGAAACCCATCCATGGC 3'
R58.9R 5' ACACACACCCCAGGGATGTC 3'

Amplicon 10 4.9 kb

R60.7F 5' TATCCTCTCTAGTCTGGTGC 3'
R65.6R 5' ATCCTCCACCCTCAAGTAGG 3'

Amplicon 11 3.7 kb

R69.0F 5' AGGCCAGGGAGGATCCCACA 3'
R72.7R 5' CTGCCTCTAGGAAATGTGCC 3'

Amplicon 12 6.5 kb

R74.1F 5' CTCTTCTGGAAGGGGTTAAC 3'
R80.6R 5' AACTCTCATCTCAAGAGCCC 3'

ASO sequences

Amplicon 1

GA1C 5' CTCAAACCGTCCTCAGGG 3'
GA1T 5' CTCAAACCTGCCTCAGGG 3'

GA6G 5' GTTTTAAGTTCAATCTGT 3'
GA6A 5' GTTTTAAATTCATCTGT 3'

GA8C 5' GCTCTTTTCAGACCCACAG 3'
GA8T 5' GCTCTTTTAGACCCACAG 3'

GA9A 5' GTGACAAACCTAAATACA 3'
GA9C 5' GTGACAACCCTAAATACA 3'

GA11G 5' AGCGAATGGAGACCAGCT 3'
GA11A 5' AGCGAATAGAGACCAGCT 3'

A4C 5' CGTGTGGCACCAATTCAG 3'
A4T 5' CGTGTGGTACCAATTCAG 3'

A6+ 5' CTCCCCGGCACTTCCGTC 3'
A6- 5' ACTCCCCGCACTTCCGTC 3'

A7C 5' AAACGTTCTGCTGCCTCC 3'
A7G 5' AAACGTTGTGCTGCCTCC 3'

A11+ 5' ACTCAATAAATGTTTAAA 3'
A11- 5' AACATGGTTAATGTTTAA 3'

A12C 5' CAGCTGACTAAAACAAAA 3'
A12G 5' CAGCTGAGTAAAACAAAA 3'

A10T 5' GACTCTGTCTCTACATAA 3'
A10C 5' GACTCTGCCTCTACATAA 3'

Amplicon 2

AB9T 5' AAGGAAATAGAGCTTCTT 3'
AB9C 5' AAGGAAACAGAGCTTCTT 3'

AB6T 5' TAAAAAGTAAGCAACAAC 3'
AB6A 5' TAAAAAGAAAGCAACAAC 3'

B4- 5' TTCATCTAAATCTAGTCA 3'
B4+ 5' TTCATCTAAATCTAGTC 3'

B9G 5' TTACTIONCGGAATGAATTC 3'
B9A 5' TTACTIONCAGAATGAATTC 3'

B5C 5' ACAGAGTCTCGCTCTGTT 3'
B5T 5' ACAGAGTTTCGCTCTGTT 3'

B6G 5' ATGGTGCGATCTCAGCTC 3'
B6A 5' ATGGTGCAATCTCAGCTC 3'

B7T 5' AAATACATACATAAAAATT 3'
B7G 5' AAATACAGACATAAAAATT 3'

Amplicon 3

BC3C 5' GACTCCATCTCAWAWAAA 3'
BC3T 5' GACTCCATTTCAWAWAAA 3'

BC4T 5' CCTACCCTTACCCCCAGC 3'
BC4C 5' CCTACCCTTACCCCCAGC 3'

BC5T 5' TTA CTTATATTTATGGAA 3'
BC5C 5' TTA CTTACATTTATGGAA 3'

BC9G 5' ATAGAAGGAACAAACCTC 3'
BC9A 5' ATAGAAGAAACAAACCTC 3'

C1T 5' ACCAACCTGACAATCAAA 3'
C1A 5' ACCAACCAGACAATCAAA 3'

C2A 5' AAAATTCATATGGAACAA 3'
C2C 5' AAAATTCCTATGGAACAA 3'

C3C 5' ATGGAACAACAACAAAAA 3'
C3T 5' ATGGAACAACAATAAAAA 3'

C4A 5' ACTAAAAGTAGATCTACC 3'
C4T 5' ACTAAAAGTAGTTCTACC 3'

C6C 5' GTCATTACATCAAAAAGA 3'
C6T 5' GTCATTATATCAAAAAGA 3'

Amplicon 4

D2T 5' TATTCACTTACTGTACAA 3'
D2G 5' TATTCACGTACTGTACAA 3'

E2+ 5' GCAGTCAGACAAAATATT 3'
E2- 5' GCAGTCAAAATATTCTAT 3'

Amplicon 5

DS6C 5' CTGTAATCCTGGCTACTT 3'
DS6T 5' CTGTAATTCTGGCTACTT 3'

Amplicon 6

DS9C 5' TCACTGCCTGAGCATCCC 3'
DS9T 5' TCACTGCCTGAGCATCCC 3'

DS14C 5' GATGGGGCTTAAGCAAAA 3'
DS14A 5' GATGGGGATTAAGCAAAA 3'

Amplicon 7

DS20A 5' CGCGCGCATTTCGTGGGG 3'
DS20T 5' CGCGCGCTTTTCGTGGGG 3'

DS21T 5' AATTCTTTGTACATTCT 3'
DS21G 5' AATTCTTGGTACATTCT 3'

DS26A 5' GACAGAAAGACCATATTC 3'
DS26G 5' GACAGAAGGACCATATTC 3'

Amplicon 8

DS31T 5' GAGGTTCTTCTTTGAAGG 3'
DS31C 5' GAGGTTCCCTCTTTGAAGG 3'

HA4C 5' ATCTCCTCCGCCTCCTCT 3'
HA4T 5' ATCTCCTCTGCCTCCTCT 3'

HA5A 5' CTCCTCTAATTATGACTT 3'
HA5T 5' CTCCTCTTATTATGACTT 3'

Amplicon 9

HB2C 5' GTTTCATCATGTTTGCCA 3'
HB2T 5' GTTTCATTATGTTTGCCA 3'

HB5+ 5' GTAGAGACAGGGTCTCGC 3'
HB5- 5' GTAGAGACGGTCTCGCTG 3'

Amplicon 10

H12A/G

HD2C 5' CAATCCTCTTTTGTTTAG 3'
HD2G 5' CAATCCTGTTTGTTTAG 3'

Amplicon 11

JJ3T/A
JJ6T/C
JJ7A/C
J3A/G
J4A/G

Amplicon 12

JJK5C/T
JJK9C/T
JJK18G/C
JJK14G/A
JJK8C/G
J5C/T
J6G/A
JK4T/C

REFERENCES

- Abdullah MFF, Borts RH (2001) Meiotic recombination frequencies are affected by nutritional states in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 98: 14524-14529
- Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Harper JI (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 68, 191-197.
- Allcock RJ, Atrazhev AM, Beck S, de Jong PJ, Elliott JF, Forbes S, Halls K, Horton R, Osoegawa K, Rogers J, Sawcer S, Todd JA, Trowsdale J, Wang Y, Williams S (2002) The MHC haplotype project: a resource for HLA-linked association studies. *Tissue Antigens* 59: 520-521
- Allers T, Lichten M (2001) Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* 106: 47-57
- Anderson LK, Reeves A, Webb LM, Ashley T (1999) Distribution of crossing over on mouse synaptonemal complexes using immunofluorescent localization of MLH1 protein. *Genetics* 151:1569-1579
- Antonarakis SE, Boehm CD, Giardina PJ, Kazazian HH (1982) Nonrandom association of polymorphic restriction sites in the beta-globin gene cluster. *Proc Natl Acad Sci USA* 79: 137-141
- Aoki K, Suzuki K, Sugano T, Tasaka T, Nakahara K, Kuge O, Omori A, Kasai M (1995) A novel gene, *Translin*, encodes a recombination hotspot binding protein associated with chromosomal translocations. *Nat Genet* 10: 167-174
- Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 3: 299-309
- Auerbach SD, Loftus RW, Itani OA, Thomas CP (2000) Human amiloride-sensitive epithelial Na⁺ channel gamma subunit promoter: functional analysis and identification of a polypurine-polypyrimidine tract with the potential for triplex DNA formation. *Biochem J* 347: 105-114
- Awadalla P (2003) The evolutionary genomics of pathogen recombination. *Nat Rev Genet* 4: 50-60
- Baker SM, Plug AW, Prolla TA, Bronner CE, Harris AC, Yao X, Christie DM, Monell C, Arnheim N, Bradley A, Ashley T, Liskay RM (1996) Involvement of mouse *Mlh1* in DNA mismatch repair and meiotic crossing over. *Nat Genet* 13: 336-342
- Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. *Nat Rev Genet* 4: 99-111
- Barlow AL, Hulten MA (1998) Crossing over analysis at pachytene in man. *Eur J Hum Genet* 6: 350-358
- Baudat F, Nicolas A (1997) Clustering of meiotic double-strand breaks on yeast chromosome III. *Proc Natl Acad Sci USA* 94: 5213-5218
- Beck S, Kelly A, Radley E, Khurshid F, Alderton RP, Trowsdale J (1992) DNA sequence analysis of 66 kb of the human MHC class II region encoding a cluster of genes for antigen processing. *J Mol Biol* 228: 433-441
- Beck S, Abdulla S, Alderton RP, Glynn RJ, Gut IG, Hosking LK, Jackson A, Kelly A, Newell WR, Sanseau P, Radley E, Thorpe KL, Trowsdale J (1996) Evolutionary dynamics of non-coding sequences within the class II region of the human MHC. *J Mol Biol* 255: 1-13
- Biet E, Sun J, Dutreix M (1999) Conserved sequence preference in DNA binding among recombination proteins: an effect of ssDNA secondary structure. *Nucleic Acids Res* 27: 596-600
- Blat Y, Protacio RU, Hunter N, Kleckner N (2002) Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell* 111: 791-802
- Blumental-Perry A, Zenvirth D, Klein S, Onn I, Simchen G (2000) DNA motif associated with meiotic double-strand break regions in *Saccharomyces cerevisiae*. *EMBO Reports* 1: 232-238

- Borde V, Goldman AS, Lichten M (2000) Direct coupling between meiotic DNA replication and recombination initiation. *Science* 290: 806-809
- Borts, RH, Haber, JE (1987) Meiotic recombination in yeast: alteration by multiple heterozygosities. *Science* 237: 1459-1465
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32: 314-331
- Boulton A, Myers RS, Redfield RJ (1997) The hotspot conversion paradox and the evolution of meiotic recombination.. *Proc Natl Acad Sci USA* 94: 8058-8063
- Broman KW, Weber JL (2000) Characterization of human crossover interference. *Am J Hum Genet* 66: 1911-1926
- Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* 63: 861-869
- Bryda EC, DePari JA, Sant'Angelo DB, Murphy DB, Passmore HC (1992) Multiple sites of crossing over within the Eb recombinational hotspot in the mouse. *Mamm Genome* 2: 123-129
- Bullard SA, Kim S, Galbraith AM, Malone RE (1996) Double strand breaks at the HIS2 recombination hot spot in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 93: 13054-13059
- Cao L, Alani E, Kleckner N (1990) A pathway for generation and processing of double-strand breaks during meiotic recombination in *S. cerevisiae*. *Cell* 61: 1089-1101
- Cardon LR, Bell JI (2001) Association study designs for complex diseases. *Nat Rev Genet* 2: 91-98
- Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander, ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22: 231-238
- Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton
- Cervantes MD, Farah JA, Smith GR (2000) Meiotic DNA breaks associated with recombination in *S. pombe*. *Mol Cell* 5: 883-8
- Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm C, Kazazian HH (1984) Nonuniform recombination within the human beta-globin gene cluster. *Am J Hum Genet* 36: 1239-1258
- Cheng S, Chang SY, Gravitt P, Respass R (1994) Long PCR. *Nature* 369: 684-685
- Church GM, Gilbert W (1984) Genomic sequencing. *Proc Natl Acad Sci USA* 81: 1991-1995
- Civardi L, Xia Y, Edwards KJ, Schnable PS, Nikolau BJ (1994) The relationship between genetic and physical distances in the cloned a1-sh2 interval of the *Zea mays L.* genome. *Proc Natl Acad Sci USA* 91: 8268-72
- Collins A, Frezal J, Teague J, Morton NE (1996) A metric map of humans: 23, 500 loci in 850 bands. *Proc Natl Acad Sci USA* 93: 14771-1475
- Collins FS, Guyer MS, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. *Science* 278: 1580-1581
- Cooper DN, Krawczak M (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum Genet* 83: 181-8
- Cromie GA, Connelly JC, Leach DR (2001) Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans. *Mol Cell* 8: 1163-1174

- Cruciani F, Bernardini L, Santolamazza P, Modiano D, Torroni A, Scozzari R (2003) Linkage disequilibrium analysis of the human adenosine deaminase (*ada*) gene provides evidence for a lack of correlation between hot spots of equal and unequal homologous recombination. *Genomics* 82: 20-33
- Cullen M, Erlich H, Klitz W, Carrington M (1995) Molecular mapping of a recombination hotspot located in the second intron of the human TAP2 locus. *Am J Hum Genet* 56: 1350-1358
- Cullen M, Noble J, Erlich H, Thorpe K, Beck S, Klitz W, Trowsdale J, Carrington M (1997) Characterization of recombination in the HLA class II region. *Am J Hum Genet* 60: 397-407
- Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M (2002) High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am J Hum Genet* 71:759-776
- Dai X, Greizerstein MB, Nadas-Chinni K, Rothman-Denes LB (1997) Supercoil-induced extrusion of a regulatory DNA hairpin. *Proc Natl Acad Sci USA* 94: 2174-2179
- Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. *Nat Genet* 29: 229-232
- Davis L, Smith GR (2001) Meiotic recombination and chromosome segregation in *Schizosaccharomyces pombe*. *Proc Natl Acad Sci USA* 98: 8395-8402
- Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lohmussaer E, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418: 544-548
- de Massy B, Nicolas A (1993) The control in cis of the position and the amount of the ARG4 meiotic double-strand break of *Saccharomyces cerevisiae*. *EMBO J* 12: 1459-1466
- de Massy B (2003) Distribution of meiotic recombination sites. *Trends Genet* 19: 514-522
- Depaulis F, Veuille M (1998) Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol Biol Evol* 15: 1788-1790
- Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* 12: 387-395
- Dooner HK, Martinez-Ferez IM (1997) Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. *Plant Cell* 9: 1633-1646
- Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, Carlomagno F, Xu CF, Dawson E, Rhodes S, Ueda S, Lai E, Luben RN, Van Rensburg EJ, Mannermaa A, Kataja V, Rennart G, Dunham I, Purvis I, Easton D, Ponder BA (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* 67: 1544-1554
- Eggleston AK, West SC (1997) Recombination initiation: easy as A, B, C, D... chi? *Curr Biol* 7:745-749
- Fan Q, Xu F, Petes TD (1995) Meiosis-specific double-strand DNA breaks at the HIS4 recombination hot spot in the yeast *Saccharomyces cerevisiae*: control in cis and trans. *Mol Cell Biol* 15: 1679-1688
- Fan QQ, Xu F, White MA, Petes TD (1997) Competition between adjacent meiotic recombination hotspots in the yeast *Saccharomyces cerevisiae*. *Genetics* 145: 661-670
- Farah JA, Hartsuiker E, Mizuno K, Ohta K, Smith GR (2002) A 160-bp palindrome is a Rad50.Rad32-dependent mitotic recombination hotspot in *Schizosaccharomyces pombe*. *Genetics* 161: 461-468
- Feinberg AP, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem* 132: 6-13

- Feinberg AP, Vogelstein B (1984) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Addendum. *Anal Biochem* 137: 266-267
- Fox ME, Smith GR (1998) Control of meiotic recombination in *S. pombe*. *Prog Nucleic Acid Res Mol Biol* 61: 345-78
- Froenicke L, Anderson LK, Wienberg J, Ashley T (2002) Male mouse recombination maps for each autosome identified by chromosome painting. *Am J Hum Genet* 71: 1353-1368
- Fu H, Zheng Z, Dooner HK (2002) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc Natl Acad Sci USA* 99: 1082-1087
- 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
- Gabriel S, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. *Science* 296: 2225-2229
- Gaudieri S, Leelayuwat C, Tay GK, Townend DC, Dawkins RL (1997) The major histocompatibility complex (MHC) contains conserved polymorphic genomic sequences that are shuffled by recombination to form ethnic-specific haplotypes. *J Mol Evol* 45: 17-23
- Gaudieri S, Dawkins RL, Habara K, Kulski JK, Gojobori T (2000) SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. *Genome Res* 10: 1579-1586
- Gendrel CG, Boulet A, Dutreix M (2000) (CA/GT)_n microsatellites affect homologous recombination during yeast meiosis. *Genes Dev* 14: 1261-1268
- Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD (2000) Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 97: 11383-11390
- Gray IC, Campbell DA, Spurr NK (2000) Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* 9: 2403-2408
- Guillon H, de Massy B (2002) An initiation site for meiotic crossing-over and gene conversion in the mouse. *Nat Genet* 32: 296-299
- Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* 22: 239-247
- Harteveld KL, Losekoot M, Fodde R, Giordano PC, Bernini LF (1997) The involvement of Alu repeats in recombination events at the alpha-globin gene cluster: characterization of two alpha-zero-thalassaemia deletion breakpoints. *Hum Genet* 99: 528-534
- Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, Soeda E, Ohki M, Takagi T, Sakaki Y, Taudien S, Blechschmidt K, Polley A, Menzel U, Delabar J, Kumpf K, Lehmann R, Patterson D, Reichwald K, Rump A, Schillhabel M, Schudy A (2000) The DNA sequence of human chromosome 21. *Nature* 405: 311-319
- Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. *Genetics* 117: 331-341
- Hedrick PW (1988) Inference of recombinational hotspots using gametic disequilibrium values. *Heredity* 60: 435-438
- Heine D, Khambata S, Wydner KS, Passmore HC (1994) Analysis of recombinational hot spots associated with the p haplotype of the mouse MHC. *Genomics* 23: 168-177
- Heng HH, Chamberlain JW, Shi XM, Spyropoulos B, Tsui LC, Moens PB (1996) Regulation of meiotic chromatin loop size by chromosomal position. *Proc Natl Acad Sci USA* 93: 2795-2800

- Higgs DR, Vickers MA, Wilkie AO, Pretorius IM, Jarman AP, Weatherall DJ (1989) A review of the molecular genetics of the human alpha-globin gene cluster. *Blood* 73: 1081-1104
- Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theor Appl Genet* 38: 226-231
- Högstrand K, Böhme J (1994) A determination of the frequency of gene conversion in unmanipulated mouse sperm. *Proc Natl Acad Sci USA* 91: 9921-9925.
- Horton R, Niblett D, Milne S, Palmer S, Tubby B, Trowsdale J, Beck S (1998) Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *Mol Biol* 282: 71-97
- Hubert R, MacDonald M, Gusella J, Arnheim N (1994) High resolution localization of recombination hot spots using sperm typing. *Nat. Genet* 7: 420-424
- Hunter N, Borts RH (1997) Mlh1 is unique among mismatch repair proteins in its ability to promote crossing-over during meiosis. *Genes Dev* 11: 1573-1582
- Hunter N, Kleckner N (2001) The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination. *Cell* 106: 59-70
- Hunter N, Valentin Borner G, Lichten M, Kleckner N (2001) Gamma-H2AX illuminates meiosis. *Nat Genet* 27: 236-238
- Huttley GA, Smith MW, Carrington M, O'Brien SJ (1999) A scan for linkage disequilibrium across the human genome. *Genetics* 152: 1711-1722
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921
- Isobe T, Yoshino M, Mizuno K, Lindahl KF, Koide T, Gaudieri S, Gojobori T, Shiroishi T (2002) Molecular characterization of the Pb recombination hotspot in the mouse major histocompatibility complex class II region. *Genomics* 80: 229-235
- Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29: 217-222
- Jeffreys AJ, MacLeod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. *Nature* 354: 204-209
- Jeffreys AJ, Murray J, Neumann R (1998a) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell* 2: 267-273
- Jeffreys AJ, May CA (2003) DNA enrichment by allele-specific hybridization (DEASH): a novel method for haplotyping and for detecting low-frequency base substitutional variants and recombinant DNA molecules. *Genome Res* 13: 2316-2324
- Jeffreys AJ, May CA (2004) Intense and highly localized gene conversion activity in human meiotic crossover hotspots. *Nat Genet*, published online 4 Jan 2004
- Jeffreys AJ, Neil DL, Neumann R (1998b) Repeat instability at human minisatellites arising from meiotic recombination. *EMBO J* 17: 4147-4147
- Jeffreys AJ, Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* 31: 267-271
- Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60: 473-485
- Jeffreys AJ, Tamaki K, MacLeod A, Monckton DG, Neil DL, Armour JA (1994) Complex gene conversion events in germline mutation at human minisatellites. *Nat Genet* 6: 136-145

- Jeffreys AJ, Ritchie A, Neumann R (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum Mol Genet* 9: 725-733
- Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. *Nat Genet* 29: 233-237
- Kaessmann H, Zollner S, Gustafsson AC, Wiebe V, Laan M, Lundeberg J, Uhlen M, Pääbo S (2002) Extensive linkage disequilibrium in small human populations in Eurasia. *Am J Hum Genet* 70: 673-685
- Kauppi L, Sajantila A, Jeffreys AJ (2003) Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum Mol Genet* 12: 33-40
- Keeney S, Giroux CN, Kleckner N (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* 88: 375-384
- Khambata S, Mody J, Modzelewski A, Heine D, Passmore HC (1996) Ea recombinational hot spot in the mouse major histocompatibility complex maps to the fourth intron of the Ea gene. *Genome Res* 6: 195-201
- Kirkpatrick DT, Wang YH, Dominska M, Griffith JD, Petes TD (1999) Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes. *Mol Cell Biol* 19: 7661-7671
- Kleckner N (1996) Meiosis: how could it work? *Proc Natl Acad Sci USA* 93: 8167-8174
- Klein S, Zenvirth D, Dror V, Barton AB, Kaback DB, Simchen G (1996a) Patterns of meiotic double-strand breakage on native and artificial yeast chromosomes. *Chromosoma* 105: 276-284
- Klein S, Zenvirth D, Sherman A, Ried K, Rappold G, Simchen G (1996b) Double-strand breaks on YACs during yeast meiosis may reflect meiotic recombination in the human genome. *Nat Genet* 13: 481-484
- Kobori JA, Strauss E, Minard K, Hood L (1986) Molecular analysis of the hotspot of recombination in the murine major histocompatibility complex. *Science* 234: 173-179
- Koehler KE, Cherry JP, Lynn A, Hunt PA, Hassold TJ (2002) Genetic control of mammalian meiotic recombination. I. Variation in exchange frequencies among males from inbred mouse strains. *Genetics* 162: 297-306
- Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. *Nat Genet* 31: 241-247
- Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 22: 139-144
- Laan M, Paabo S (1997) Demographic history and linkage disequilibrium in human populations. *Nat Genet* 17: 435-438
- Lafuse WP, David CS (1986) Recombination hot spots within the I region of the mouse H-2 complex map to the E beta and E alpha genes. *Immunogenetics* 24: 352-360
- Lafuse WP, Lee ST, Castle L, David CS (1989) Restriction fragment analysis of H-2 recombinant mouse strains with crossovers between E alpha and C4 genes. *Immunogenetics* 30: 387-389
- Laurie DA, Hulten MA (1985a) Further studies on bivalent chiasma frequency in human males with normal karyotypes. *Ann Hum Genet* 49: 189-201
- Laurie DA, Hulten MA (1985b) Further studies on chiasma distribution and interference in the human male. *Ann Hum Genet* 49: 203-214
- Lawrie NM, Tease C, Hulten MA (1995) Chiasma frequency, distribution and interference maps of mouse autosomes. *Chromosoma* 140: 308-314

- Lewontin RC (1964) The interaction of selection and linkage. II. Optimum Models. *Genetics* 50: 757-782
- Lewontin RC (1984) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49: 49-67
- Li, W-H (1997) *Molecular evolution*. Sinauer Associates, Canada
- Liang F, Han M, Romanienko PJ, Jasin M (1998) Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proc Natl Acad Sci USA* 95: 5172-5177
- Lichten M, Goldman, AS (1995) Meiotic recombination hotspots. *Annu Rev Genet* 29: 423-444
- Lien S, Szyda J, Schechinger B, Rappold G, Arnheim N (2000) Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am J Hum Genet* 66: 557-566
- Lonjou C, Collins A, Morton NE (1999) Allelic association between marker loci. *Proc Natl Acad Sci USA*, 96: 1621-1626
- Lopes J, Ravise N, Vandenberghe A, Palau F, Ionasescu V, Mayer M, Levy N, Wood N, Tachi N, Bouche P, Latour P, Ruberg M, Brice A, LeGuern E (1998) Fine mapping of de novo CMT1A and HNPP rearrangements within CMT1A-REPs evidences two distinct sex-dependent mechanisms and candidate sequences involved in recombination. *Hum Mol Genet* 7: 141-148
- Lopes J, Tardieu S, Silander K, Blair I, Vandenberghe A, Palau F, Ruberg M, Brice A, LeGuern E (1999) Homologous DNA exchanges in humans can be explained by the yeast double-strand break repair model: a study of 17p11.2 rearrangements associated with CMT1A and HNPP. *Hum Mol Genet* 8: 2285-2292
- Lopez-Correa C, Dorschner M, Brems H, Lazaro C, Clementi M, Upadhyaya M, Dooijes D, Moog U, Kehrer-Sawatzki H, Rutkowski JL, Fryns JP, Marynen P, Stephens K, Legius E (2001) Recombination hotspot in NF1 microdeletion patients. *Hum Mol Genet* 10: 1387-1392
- Lupski JR, de Oca-Luna R M, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedo-Cardenas O, Barker DF, Killian JM, Garcia CA (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* 66: 219-232
- Lynn A, Kashuk C, Petersen MB, Bailey JA, Cox DR, Antonarakis SE, Chakravarti A (2000) Patterns of meiotic recombination on the long arm of human chromosome 21. *Genome Res* 10: 1319-1332
- Lynn A, Koehler KE, Judis L, Chan ER, Cherry JP, Schwartz S, Seftel A, Hunt PA, Hassold TJ (2002) Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science* 296: 2222-2225
- Mahadevaiah SK, Turner JM, Baudat F, Rogakou EP, de Boer P, Blanco-Rodriguez J, Jasin M, Keeney S, Bonner WM, Burgoyne PS (2001) Recombinational DNA double-strand breaks in mice precede synapsis. *Nat Genet* 27: 271-276
- Majewski J, Ott J (2000) GT repeats are associated with recombination on human chromosome 22. *Genome Res* 10: 1108-1114
- Malone RE, Bullard S, Lundquist S, Kim S, Tarkowski T (1992) A meiotic gene conversion gradient opposite to the direction of transcription. *Nature* 359: 154-155
- Marshall B, Leelayuwat C, Degli-Esposti MA, Pinelli M, Abraham LJ, Dawkins RL (1993) New major histocompatibility complex genes. *Hum Immunol* 38: 24-29
- May CA, Shone AC, Kalaydjieva L, Sajantila A, Jeffreys, AJ (2002) Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nat Genet* 31: 272-275
- MHC sequencing consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature*, 401, 921-923

- Miller RD, Kwok PY (2001) The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Hum Mol Genet* 10: 2195-2198
- Miller RD, Taillon-Miller P, Kwok PY (2001) Regions of low single-nucleotide polymorphism incidence in human and orangutan xq: deserts and recent coalescences. *Genomics* 71: 78-88
- Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM (2003) Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci USA* 100: 5926-5931
- Mizuno K, Koide T, Sagai T, Moriwaki K, Shiroishi T (1996) Molecular analysis of a recombinational hotspot adjacent to Lmp2 gene in the mouse MHC: fine location and chromatin structure. *Mamm Genome* 7:490-496
- Mohrenweiser HW, Tsujimoto S., Gordon L, Olsen AS (1998) Regions of sex-specific hypo- and hyper-recombination identified through integration of 180 genetic markers into the metric physical map of human chromosome 19. *Genomics* 47: 153-162
- Monckton DG, Neumann R, Guram T, Fretwell N, Tamaki K, MacLeod A, Jeffreys AJ (1994) Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nat Genet*, 8: 162-170
- Morton NE (1982) *Outline of genetic epidemiology*. Karger, Basel
- Nachman MW (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet* 17: 481-485
- Nachman MW, Bauer VL, Crowell SL, Aquadro CF (1998) DNA variability and recombination rates at X-linked loci in humans. *Genetics*, 150: 1133-1141
- Nag DK, Kurst A (1997) A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast *Saccharomyces cerevisiae*. *Genetics* 146: 835-847
- Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* 76: 5269-73
- Nei, M (1972) Genetic distance between populations. *Am Nat* 106: 283-292
- Nicolas A, Treco D, Schultes NP, Szostak JW (1989) An initiation site for meiotic gene conversion in the yeast *Saccharomyces cerevisiae*. *Nature* 338: 35-39
- Ohta K, Shibata T, Nicolas A (1994) Changes in chromatin structure at recombination initiation sites during yeast meiosis. *EMBO J* 13: 5754-5763
- Okagaki RJ, Weil CF (1997) Analysis of recombination sites within the maize waxy locus. *Genetics* 147: 815-821
- Paques F, Haber JE (1999) Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol Mol Biol Rev* 63: 349-404
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294: 1719-1723
- Peterson AC, Di Rienzo A, Lehesjoki AE, de la Chapelle A, Slatkin M, Freimer NB (1995) The distribution of linkage disequilibrium over anonymous genome regions. *Hum Mol Genet* 4: 887-894
- Petes TD (2001) Meiotic recombination hot spots and cold spots. *Nat Rev Genet* 2: 360-369
- Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, Ankener WM, Alfisi SV, Kuo FS, Camisa AL, Pazorov V, Scott KE, Carey BJ, Faith J, Katari G, Bhatti HA, Cyr JM, Derohannessian V, Elosua C, Forman AM, Grecco NM, Hock CR, Kuebler JM, Lathrop JA, Mockler MA, Nachtman EP, Restine SL, Varde SA, Hozza MJ, Gelfand CA, Broxholme J, Abecasis GR,

- Boyce-Jacino MT, Cardon LR (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet* 33: 382-387
- Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 69: 1-14
- Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenism. *J Hered* 86: 248-249
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian, SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. *Nature* 411: 199-204
- Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altschuler D (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 32: 135-142
- Reiter LT, Hastings PJ, Nelis E, De Jonghe P, Van Broeckhoven C, Lupski JR (1998) Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *Am J Hum Genet* 62: 1023-1033
- Risch NJ (2000) Searching for genetic determinants in the new millennium. *Nature* 405: 847-856
- Rocco V, de Massy B, Nicolas A (1992) The *Saccharomyces cerevisiae* ARG4 initiator of meiotic gene conversion and its associated double-strand DNA breaks can be inhibited by transcriptional interference. *Proc Natl Acad Sci USA* 89: 12068-12072
- Roeder GS (1995) Sex and the single cell: meiosis in yeast. *Proc Natl Acad Sci USA* 92: 10450-10456
- Roeder GS (1997) Meiotic chromosomes: it takes two to tango. *Genes Dev* 11: 2600-2621
- Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239: 487-491
- Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus ML, Aula P, Beckman L, Tranebjaerg L, Gedde-Dahl T, Issel-Tarver L, DiRienzo A, Pääbo S (1995) Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res* 5: 42-52
- Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning; a laboratory manual*, Second Edition. Schneider JA, Peto TE, Boone RA, Boyce AJ, Clegg JB (2002) Direct measurement of the male recombination fraction in the human beta-globin hot spot. *Hum Mol Genet* 11: 207-215
- Schultes NP, Szostak JW (1990) Decreasing gradients of gene conversion on both sides of the initiation site for meiotic recombination at the ARG4 locus in yeast. *Genetics* 126: 813-822
- Schwacha A, Kleckner N (1995) Identification of double Holliday junctions as intermediates in meiotic recombination. *Cell* 83: 783-791
- Service SK, Ophoff RA, Freimer NB (2001) The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum Mol Genet* 10: 545-551
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29: 308-311
- Shiroishi T, Hanzawa N, Sagai T, Ishiura M, Gojobori T, Steinmetz M, Moriwaki K (1990) Recombinational hotspot specific to female meiosis in the mouse major histocompatibility complex. *Immunogenetics* 31: 79-88
- Shiroishi T, Sagai T, Hanzawa N, Gotoh H, Moriwaki K (1991) Genetic control of sex-dependent meiotic recombination in the major histocompatibility complex of the mouse. *EMBO J* 10: 681-686
- Shiroishi T, Koide T, Yoshino M, Sagai T, Moriwaki K (1995) Hotspots of homologous recombination in mouse meiosis. *Adv Biophys* 31: 119-132

- Smith RA, Ho PJ, Clegg JB, Kidd JR, Thein SL (1998) Recombination breakpoints in the human beta-globin gene cluster. *Blood* 92: 4415-4421
- Snoek M, Teuscher C, van Vugt H (1998) Molecular analysis of the major MHC recombinational hot spot located within the *G7c* gene of the murine class III region that is involved in disease susceptibility. *Immunol* 160: 266-272
- Stead JDH, Jeffreys AJ (2002) Structural analysis of insulin minisatellite alleles reveals unusually large differences in diversity between Africans and non-Africans. *Am J Hum Genet* 71: 1273-1284
- Stead JD, Hurler ME, Jeffreys AJ (2003) Global haplotype diversity in the human insulin gene region. *Genome Res* 13: 2101-2111
- Steiner WW, Schreckhise RW, Smith GR (2002) Meiotic DNA breaks at the *S. pombe* recombination hot spot M26. *Mol Cell* 9: 847-855
- Steinmetz M, Stephan D, Fischer Lindahl K (1986) Gene organization and recombinational hotspots in the murine major histocompatibility complex. *Cell* 44: 895-904
- Stephens R, Horton R, Humphray S, Rowen L, Trowsdale J, Beck S (1999) Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J Mol Biol* 291: 789-799
- Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF (2001a) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293: 489-493
- Stephens M, Smith NJ, Donnelly P (2001b) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68: 978-989
- Sun H, Treco D, Schultes NP, Szostak JW (1998) Double-strand breaks at an initiation site for meiotic gene conversion. *Nature* 338: 87-90
- Sun H, Treco D, Szostak JW (1991) Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the ARG4 recombination initiation site. *Cell* 64: 1155-1161
- Sunyaev SR, Lathe WC, Ramensky VE, Bork P (2000) SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet* 16:335-337
- Sved JA (1971) Linkage disequilibrium and homozygosity of chromosomal segments in finite populations. *Theor Popul Biol* 2: 125-141
- Svetlova E, Avril-Fournout N, Ira G, Deschavanne P, Filipinski J (1998) DNase-hypersensitive sites in yeast artificial chromosomes containing human DNA. *Mol Gen Genet* 257: 292-298
- Sym M, Roeder GS (1994) Crossover interference is abolished in the absence of a synaptonemal complex protein. *Cell* 79: 283-292
- Syvanen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* 2: 930-942
- Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW (1983) The double-strand-break repair model for recombination. *Cell* 33: 25-35
- Taillon-Miller P, Bauer-Sardina I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok PY (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat Genet* 25: 324-328
- Tapper WJ, Ke X, Morton NE, Collins A (2002) Recombination, interference and sequence: comparison of chromosomes 21 and 22. *Ann Hum Genet* 66: 75-86

- Tease C, Hartshorne GM, Hulten MA (2002) Patterns of meiotic recombination in human fetal oocytes. *Am J Hum Genet* 70: 1469-1479
- Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am J Hum Genet* 66: 69-83
- Tenzen T, Yamagata T, Fukagawa T, Sugaya K, Ando A, Inoko H, Gojobori T, Fujiyama A, Okumura K, Ikemura T (1997) Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. *Mol Cell Biol* 17: 4043-4050
- Thuriaux P (1977) Is recombination confined to structural genes on the eukaryotic genome? *Nature* 268: 460-462
- Ting JP, Trowsdale J (2002) Genetic control of MHC class II expression. *Cell* 109: 21-33
- Trowsdale J (1995) "Both man & bird & beast": comparative organization of MHC genes. *Immunogenetics* 41: 1-17
- Twells RC, Mein CA, Phillips MS, Hess JF, Veijola R, Gilbey M, Bright M, Metzker M, Lie BA, Kingsnorth A, Gregory E, Nakagawa Y, Snook H, Wang WY, Masters J, Johnson G, Eaves I, Howson JM, Clayton D, Cordell HJ, Nutland S, Rance H, Carr P, Todd JA (2003) Haplotype structure, LD blocks, and uneven recombination within the LRP5 gene. *Genome Res* 13: 845-55
- Uematsu Y, Kiefer H, Schulze R, Fischer-Lindahl K, Steinmetz M (1986) Molecular characterization of a meiotic recombinational hotspot enhancing homologous equal crossing-over. *EMBO J* 5: 2123-2129
- Wahls WP, Wallace LJ, Moore PD (1990) The Z-DNA motif d(TG)₃₀ promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. *Mol Cell Biol* 10: 785-793
- Wahls WP (1998) Meiotic recombination hotspots: shaping the genome and insights into hypervariable minisatellite DNA change. *Curr Top Dev Biol* 37: 37-75
- Waldman AS, Tran H, Goldsmith EC, Resnick MA (1999) Long inverted repeats are an at-risk motif for recombination in mammalian cells. *Genetics* 153: 1873-1883
- Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 4: 587-597
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280: 1077-1082
- Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* 71: 1227-1234
- Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7: 256-276
- Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 18: 19-24
- White MA, Dominska M, Petes TD (1993) Transcription factors are required for the meiotic recombination hotspot at the HIS4 locus in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 90: 6621-6625
- Wood WI, Gitschier J, Lasky LA, Lawn RM (1985) Base composition-independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. *Proc Natl Acad Sci USA* 82: 1585-1588

- Wu TC, Lichten M (1994) Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science* 263: 515-518
- Wu TC, Lichten M (1995) Factors that affect the location and frequency of meiosis-induced double-strand breaks in *Saccharomyces cerevisiae*. *Genetics* 140: 55-66
- Xu F, Petes TD (1996) Fine-structure mapping of meiosis-specific double-strand DNA breaks at a recombination hotspot associated with an insertion of telomeric sequences upstream of the HIS4 locus in yeast. *Genetics* 143: 1115-1125
- Xu X, Hsia AP, Zhang L, Nikolau BJ, Schnable PS (1995) Meiotic recombination break points resolve at high rates at the 5' end of a maize coding sequence. *Plant Cell* 7: 2151-2161
- Yao H, Zhou Q, Li J, Smith H, Yandau M, Nikolau BJ, Schnable PS (2002) Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. *Proc Natl Acad Sci USA* 99: 6157-6162
- Yauk CL, Bois PR, Jeffreys AJ (2003) High-resolution sperm typing of meiotic recombination in the mouse MHC Ebeta gene. *EMBO J* 22: 1389-1397
- Yip SP, Lovegrove JU, Rana NA, Hopkinson DA, Whitehouse DB (1999) Mapping recombination hotspots in human phosphoglucosmutase (PGM1). *Hum Mol Genet* 8: 1699-1706
- Yoshino M, Sagai T, Lindahl KF, Toyoda Y, Shirayoshi Y, Matsumoto K, Sugaya K, Ikemura T, Moriwaki K, Shiroishi T (1994) Recombination in the class III region of the mouse major histocompatibility complex. *Immunogenetics* 40: 280-286
- Young JA, Schreckhise RW, Steiner WW, Smith GR (2002) Meiotic recombination remote from prominent DNA break sites in *S. pombe*. *Mol Cell* 9: 253-263
- Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW, Weber JL (2001) Comparison of human genetic and sequence-based physical maps. *Nature* 409: 951-953
- Yuhki N, Beck T, Stephens RM, Nishigaki Y, Newmann K, O'Brien SJ (2003) Comparative genome organization of human, murine, and feline MHC class II region. *Genome Res* 13: 1169-1179
- Zangenberg G, Huang MM, Arnheim N, Erlich H (1995) New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. *Nat Genet* 10: 407-414
- Zavattari P, Deidda E, Whalen M, Lampis R, Mulargia A, Loddo M, Eaves I, Mastio G, Todd JA, Cucca F (2000) Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Hum Mol Genet* 9: 2947-2957
- Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L (2003) Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. *Hum Genet* 113: 51-59
- Zhao Z, Boerwinkle E (2002) Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* 12: 1679-1686
- Zimmerer EJ, Passmore HC (1991) Structural and genetic properties of the Eb recombinational hotspot in the mouse. *Immunogenetics* 33: 132-140