ANALYSIS OF MEIOTIC RECOMBINATION IN THE HUMAN PSEUDOAUTOSOMAL REGIONS

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Shriparna Sarbajna BSc. (Leicester)

Department of Genetics

University of Leicester

September 2011

Analysis of meiotic recombination in the human pseudoautosomal regions

Shriparna Sarbajna

Abstract

Meiotic recombination in humans is essential for the faithful segregation of chromosomes during meiosis and is key in generating genetic diversity. Inferring past events from contemporary SNP haplotypes and studying *de novo* events in sperm DNA has shown that recombination occurs within 1-2 kb wide 'hot spots' in the human genome.

In terms of male recombination, the pseudoautosomal regions (PARs) are especially interesting. PAR1 undergoes obligatory crossover in male meiosis and therefore constitutes a male-specific recombination 'hot' domain (rate approximately 20-fold above the genome average). It is thus ideally suited to sperm DNA studies. PAR2 is not essential in male meiosis, but nonetheless recombines at a rate >6-times the genome average. Despite this, relatively little is known about the fine-scale distribution of recombination in either pseudoautosomal region. To address this, linkage disequilibrium analysis and high-resolution sperm typing was used in this work, in order to identify and characterise a collection of PAR hot spots.

This survey led to the identification of five active PAR hot spots, thereby providing relatively easy access to crossovers and noncrossovers. A second hot spot was identified in the *SHOX* region, providing information on hot spot spacing in PAR1 and facilitating important comparisons with autosomes. The first PAR1 double hot spot, a potential resource for investigating crossover interference, was also identified.

Data from the two PAR2 hot spots (*SPRY3* and PAR2A) provided direct evidence of hot spot activation by *trans*-acting PRDM9, with different protein variants activating either hot spot. The strongest meiotic drive observed at any human hot spot was also identified at *SPRY3*, providing insights into likely mechanisms of hot spot evolution. Finally, an extensive survey at *SPRY3* provided unprecedented insights into the relative frequencies of crossovers and noncrossovers at hot spots and highlighted the presence of a second pathway of noncrossover-formation in humans.

Acknowledgements

First of all, I would like to thank Alec and Celia for giving me the opportunity to work with them and for all their help and guidance throughout the last four years. Celia, thank you very much for your time, patience and advice- I would not have got through this without your help. Alec, you have and continue to be a great source of inspiration- when it comes to science you truly do lead by example. It has been a privilege to work with you both.

I would also like to thank all members, past and present, of lab G18/G19. Matt, thank you for all your help, especially during my first few months in the lab- I will always miss your savoury mince pies at Christmas time; Jon and Raheleh, for being there and for listening to me moan over many a cup of coffee; Rita, for always being ready to help; Jenny, for always cheering me up; Milly, for all your help, especially during my last few months in the lab. I miss you all very much.

Daniel, thank you for the last six years. Coming home to you and Barney is the happiest part of my day, everyday.

Finally and most importantly, thank you mum and dad, for always believing in me. This thesis would not have been possible without your support and without all the sacrifices you have made so that I might have this opportunity. I am ever so lucky to have you as parents. I love you both very much- this thesis is dedicated to you.

Abbreviations

ARGE	Autism Genetic Research Exchange
ASO/P	Allele-specific oligonucleotide/primer
bp	Base pairs
C.I.	Confidence interval
CEPH	Centre d'Etude du Polymorphisme Humaine
chr	Chromosome
cM	Centi-Morgan
CO	Crossover
Ct	C-type *PRDM9* alleles
dbSNP	Database SNP
dHJ	Double Holliday junction
DSB(s)	Double-strand break(s)
DSBR	Double-strand break repair
dsDNA	Double-stranded DNA
ECD	Early crossover decision
FHS	Framingham Heart Study
GC	Gene conversion
hDNA	Heteroduplex DNA
JM	Joint molecule
kb	Kilo base pairs
LD	Linkage disequilibrium
LDU	Linkage disequilibrium units
LINE	Long interspersed nuclear element
Mb	Mega base pairs
MDA	Maximum displacement amplification
MHC	Major histocompatibility complex
NCO	Noncrossover
PAR	Pseudoautosomal region
PCR	Polymerase chain reaction
RFLP	Restriction fragment length polymorphism
SC	Synaptonemal complex
SDSA	Synthesis dependent strand annealing
SEI	Single end invasion
SIC	Synapsis initiation complex
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
ssDNA	Single-stranded DNA
WGA	Whole genome amplification
ZMM	Zip1-Zip2-Zip3-Zip4-Spo16, Msh4-Msh5, Mer3

Contents

CHAPTER 1. INTRODUCTION 9

1.1. OVERVIEW OF MEIOSIS 10
  1.1.1. Pre-prophase 10
  1.1.2. Early to mid-prophase 11
  1.1.3. Events ensuing during and after late prophase 17
1.2. MEIOTIC RECOMBINATION IN *S. CEREVISIAE* - THE PLAYERS, PROCESSES AND PATHWAYS 18
  1.2.1. Spo11 creates recombination-initiating DSBs 19
  1.2.2. Early processing of meiotic DSBs 20
  1.2.3. Identifying an intact homologous partner and strand invasion 21
  1.2.4. Multiple pathways of recombination in the budding yeast 23
  1.2.5. Regulation of CO-formation: The obligate CO and interference 27
1.3. DISTRIBUTION OF RECOMBINATION EVENTS IN *S. CEREVISIAE* 29
  1.3.1. Features and classification of yeast hot spots 29
  1.3.2. Global distribution of COs and NCOs in S. cerevisiae 31
1.4. MEIOTIC RECOMBINATION IN MOUSE 33
  1.4.1. Conservation of yeast (S. cerevisiae) recombination processes and pathways in mice 34
  1.4.2. Properties and distributions of COs in mice 35
  1.4.3. NCOs in mice 40
1.5. MEIOTIC RECOMBINATION IN HUMANS 41
  1.5.1. Cytogenetic studies 41
  1.5.2. Insights from analysis of recombination in human pedigrees 44
  1.5.3. Insights from linkage disequilibrium studies 50
  1.5.4. Insights from sperm-typing assays 56
1.6. HUMAN X-Y PSEUDOAUTOSOMAL REGIONS 60
  1.6.1. Recombination within PAR1 is essential in male meiosis 63
  1.6.2. PAR1 constitutes a male-specific recombinationally 'hot' domain 64
  1.6.3. Double-COs in PAR1 and implications on interference mechanisms 65
  1.6.4. Distinct temporal and structural regulation of recombination in the mouse PAR 66
1.7. THIS WORK 68

CHAPTER 2. MATERIALS AND METHODS 69

2.1. MATERIALS 69
  2.1.1. Chemical reagents 69
  2.1.2. Plastic-ware 69
  2.1.3. Oligonucleotides 69
  2.1.4. Enzymes 70
  2.1.5. Molecular weight markers 70
  2.1.6. Standard solutions 70
  2.1.7. Human DNA 70
  2.1.8. Computers 71
2.2. METHODS 72
  2.2.1. PCR amplification of DNA 72
  2.2.2. Agarose gel electrophoresis 73
  2.2.3. DNA re-sequencing 74
  2.2.4. SNP typing by allele specific nucleotide (ASO) hybridisation to dotblots 75

CHAPTER 3. CHARACTERISATION OF A SECOND RECOMBINATION HOT SPOT NEAR THE XP/YP PSEUDOAUTOSOMAL *SHOX* GENE 78

3.1. INTRODUCTION 78

3.1.1. Heterogeneity in recombination activity across PAR1 ... 78
3.1.2. Identification of the first pseudoautosomal hot spot ... 79
3.1.3. An extended survey of recombination in and around the SHOX gene ... 81
3.1.4. This work ... 83
3.2. RESULTS ... 85
3.2.1. Choice of donors ... 85
3.2.2. Allele-specific primer design and optimisation ... 86
3.2.3. Linkage phasing of semen donors ... 86
3.2.4. Re-sequencing d10 and d62 over the assay interval ... 89
3.2.5. Pilot CO assay ... 89
3.2.6. Bulk recovery of CO molecules ... 91
3.2.7. Scoring COs for d10 and d62 ... 95
3.2.8. Comparison of recombination activity and distribution of COs in d10 and d62 ... 102
3.2.9. Transmission distortion in reciprocal COs ... 105
3.3. DISCUSSION ... 107
3.3.1. Variation in CO activity between d10 and d62 ... 109
3.3.2. TD accompanying COs in d10 ... 110
3.3.3. Implications of identification of this hot spot on the overall recombination landscape in PAR1 112

CHAPTER 4. SEARCH FOR THE MOST HIGHLY ACTIVE XP/YP PSEUDOAUTOSOMAL HOT SPOTS ... 114

4.1. INTRODUCTION ... 114
4.1.1. Identification of highly active autosomal sperm CO hot spots using patterns of marker association ... 114
4.1.2. This work ... 116
4.2. RESULTS ... 119
4.2.1. High-resolution LD analysis in a panel of North European semen donors ... 119
4.2.2. High-resolution sperm analysis of the proximal LDU step (assay 'B') ... 125
4.2.3. Sperm analysis of the distal LDU step (assay 'A') ... 139
4.3. DISCUSSION ... 144
4.3.1. Difficulties and limitations of using LD profiles for hot spot identification in PAR1 ... 144
4.3.2. Utilising a pair of pseudoautosomal hot spots for investigating the processes of recombination ... 145
4.3.3. NCOs at hot spot PAR1-P and comparisons with autosomal hot spots ... 147

CHAPTER 5. IDENTIFICATION AND CHARACTERISATION OF THE FIRST PAR2 RECOMBINATION HOT SPOT ... 150

5.1. INTRODUCTION ... 150
5.1.1. This work ... 152
5.2. RESULTS ... 154
5.2.1. High-resolution mapping of exchanges in PAR2-recombinant CEPH pedigrees ... 154
5.2.2. Analysis of patterns of linkage disequilibrium over the target interval ... 158
5.2.3. High-resolution sperm analysis ... 159
5.2.4. Frequency and distribution of COs across the assay interval ... 162
5.3. DISCUSSION ... 166
5.3.1. Genomic context of the hot spot identified ... 166
5.3.2. Morphology of the SPRY3 hot spot ... 167

CHAPTER 6. FACTORS INFLUENCING CO ACTIVITY AT THE PAR2 *SPRY3* HOT SPOT ... 169

6.1. INTRODUCTION ... 169
6.1.1. Cis-regulation at hot spots ... 169
6.1.2. Trans-regulation of hot spots- Evidence from mice ... 171
6.1.3. Regulation of human hot spots by PRDM9 ... 172
6.1.4. This work ... 174
6.2. RESULTS ... 175
6.2.1. Strategy for extended survey of recombination at the SPRY3 hot spot ... 175

6.2.2. Extreme variation in CO frequencies at the SPRY3 hot spot *178*
6.2.3. Heterozygosity at SNP rs700442 causes reciprocal CO asymmetry and transmission distortion accompanying COs *179*
6.2.4. 0442 influences CO frequencies in-cis at the SPRY3 hot spot *184*
6.2.5. Trans-acting factor PRDM9 profoundly influences frequency of COs at the SPRY3 hot spot *186*
6.2.6. Combined effects of PRDM9 and 0442 allelic status on CO frequencies *191*
6.3. DISCUSSION 193
6.3.1. Central hot spot SNP 0442 influences CO frequencies in cis at the SPRY3 hot spot *193*
6.3.2. Hot spot activity at SPRY3 is regulated by in trans by meiosis-specific protein PRDM9 .... *194*
6.3.3. Insights from a detailed survey investigating the impact of variation in PRDM9 on recombination activity at CO hot spots *195*
6.3.4. Evidence of additional influences on CO activity at the SPRY3 hot spot *198*

CHAPTER 7. A SURVEY OF NON-CROSSOVERS AT THE *SPRY3* HOT SPOT **199**

7.1. INTRODUCTION 199
7.1.1. Frequencies and distributions of NCOs at the mouse A3 hot spot *200*
7.1.2. Pathways of CO and NCO-formation in mouse meiosis *202*
7.1.3. Regulation of NCOs at a mouse recombination hot spot *203*
7.1.4. This work *204*
7.2. RESULTS 206
7.2.1. Analysis of NCOs at central SNP 0442 .... *208*
7.2.2. Impact of PRDM9 variation on frequency of NCOs at 0442 .... *211*
7.2.3. Co-conversions and estimated conversion tract lengths *213*
7.3. DISCUSSION 217
7.3.1. PRDM9 status greatly influences frequencies of NCOs at the SPRY3 hot spot *218*
7.3.2. Central SNP 0442 influences frequencies of NCOs in cis at the SPRY3 hot spot *219*
7.3.3. NCOs at the SPRY3 hot spot may be generated via two distinct pathways *219*
7.3.4. Comparisons of GC-tracts in two men showing very different extents of GC-bias in NCOs *223*

CHAPTER 8. IDENTIFICATION AND CHARACTERISATION OF AN AFRICAN-ENHANCED XQ/YQ PSEUDOAUTOSOMAL HOT SPOT **225**

8.1. INTRODUCTION 225
8.1.1. Shifts in LD hot spot usage in Hutterites carrying the PRDM9 'I' allele *225*
8.1.2. Differences in PRDM9 zinc-finger array lengths influence recombination hot spot usage *226*
8.1.3. This work *227*
8.1.4. Strategy for identifying hot spots activated by non-A variants *230*
8.2. RESULTS 231
8.2.1. High-resolution LD analysis in North European and African semen donor panels *231*
8.2.2. Analysis of the putative hot spot interval using sperm-based assays *232*
8.2.3. CO-breakpoints in the interval assayed cluster into a classical sperm CO hot spot *234*
8.2.4. CO activity at hot spot PAR2A is triggered by PRDM9 Ct variants *236*
8.2.5. NCOs at the PAR2A hot spot are also regulated by PRDM9 and activated by Ct variants *238*
8.3. DISCUSSION 243
8.3.1. Hot spot PAR2A shares many features with the SPRY3 hot spot *243*
8.3.2. Hot spot PAR2A is activated by at least two different PRDM9 Ct alleles *245*

CHAPTER 9. DISCUSSION **247**

9.1. FEATURES AND DISTRIBUTIONS OF PSEUDOAUTOSOMAL HOT SPOTS 248
9.2. PSEUDOAUTOSOMAL HOT SPOTS ARE SUBJECT TO THE SAME FORCES OF HOT SPOT ATTENUATION AS THEIR AUTOSOMAL COUNTERPARTS 250
9.3. PRDM9 REGULATES CO FREQUENCIES IN *TRANS* AT PAR HOT SPOTS 251
9.3.1. Activation of hot spots by the A-variant common in Europeans *252*
9.3.2. Activation of African-enhanced hot spots by Ct variants *254*
9.4. ADDITIONAL REGULATORS OF CO ACTIVITY AT PSEUDOAUTOSOMAL HOT SPOTS 255
9.4.1. Additional cis-acting influences on CO frequencies *255*
9.4.2. Other potential trans-regulators of CO-activities *256*

9.5. INSIGHTS INTO NCO ACTIVITY AT RECOMBINATION HOT SPOTS 259
9.5.1. Distribution of NCOs at SPRY3 and other pseudoautosomal hot spots 260
9.5.2. Relative frequencies of COs and NCOs at pseudoautosomal hot spots 261
9.5.3. Regulation of NCOs at pseudoautosomal hot spots 262
9.5.4. Multiple pathways of NCO-formation at the SPRY3 hot spot 263
9.6. FUTURE CONSIDERATIONS 265

CHAPTER 2 APPENDICES 266

APPENDIX 2.1 266

CHAPTER 3 APPENDICES 267

APPENDIX 3.1 267
APPENDIX 3.2. 270
APPENDIX 3.3 271

CHAPTER 4 APPENDICES 273

APPENDIX 4.1 273
APPENDIX 4.2 285
APPENDIX 4.3. 289
APPENDIX 4.4. 294

CHAPTER 5 APPENDICES 297

APPENDIX 5.1 297
APPENDIX 5.2 298
APPENDIX 5.3. 308
APPENDIX 5.4. 311
APPENDIX 5.5. 315
APPENDIX 5.6. 319

CHAPTER 6 APPENDICES 323

APPENDIX 6.1 323
APPENDIX 6.2. 325
APPENDIX 6.3. 327
APPENDIX 6.4. 332

CHAPTER 7 APPENDICES 334

APPENDIX 7.1. 334

CHAPTER 8 APPENDICES 341

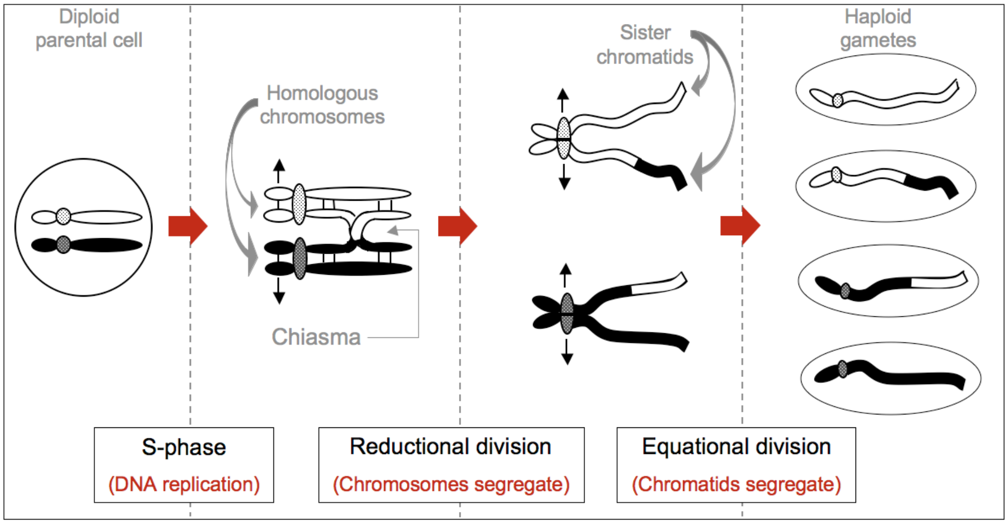
APPENDIX 8.1 341
APPENDIX 8.2 349
APPENDIX 8.3. 352
APPENDIX 8.4. 353

REFERENCES 356

Chapter 1. Introduction

Meiosis is a specialised cell division, wherein haploid gametes are produced from diploid parental cells. It is essential for sexual reproduction. Halving of the cellular chromosome complement during meiosis is achieved by a single round of DNA replication, followed by two successive rounds of chromosome segregation (Figure 1.1). Diploid chromosome complements are ultimately restored during sexual reproduction, upon fusion of haploid gametes. Meiotic recombination occurs during prophase of the first meiotic division. It is, with the exception of dipteran males and lepidopteran females, essential for the correct segregation of chromosomes during this process (McKim *et al.*, 1998; Rasmussen, 1977). Meiotic recombination is also responsible for creating genetic diversity through generation of novel allelic combinations.

Figure 1.1 Reductional and equational segregations in meiosis


Diploid parental cell
Homologous chromosomes
Chiasma
Sister chromatids
Haploid gametes
S-phase
(DNA replication)
Reductional division
(Chromosomes segregate)
Equational division
(Chromatids segregate)

Meiosis has several unique aspects compared with mitosis, the process of division utilised by somatic cells. Following DNA replication, homologous chromosomes pair and eventually synapse during the first meiotic prophase. Genetic exchange or crossing-over occurs during this stage. In the ensuing first meiotic division homologous chromosomes, each comprising of a pair of sister chromatids, move to the opposite poles of the dividing cell in a 'reductional' segregation pattern. Sister chromatids then segregate 'equationally' during meiosis II, which is in many ways analogous to mitosis.

In this chapter, concepts of meiosis and meiotic recombination will be introduced first. Over the years, the budding yeast (*Saccharomyces cerevisiae*) has emerged as the model organism of choice for investigating meiosis and meiotic recombination and has perhaps been most extensively studied. Hence, data from this organism will be used to provide an overview of these processes. More recently, the mouse has been increasingly used as a model for meiotic recombination in mammals. Data emerging from these studies will also be discussed briefly. Finally, the current understanding of meiotic recombination in humans, particularly pertaining to its frequency and distribution, will be described in detail. Since meiotic recombination in the human pseudoautosomal regions are the main emphases of this work, these genomic regions and their features and roles in male meiosis will also be introduced.

1.1. Overview of meiosis

Different stages of meiosis were mostly elucidated and defined at the chromosome level using light microscopy. These stages of meiosis represent a series of distinct changes in chromosome morphology and are conserved in virtually all diploid organisms. In the following paragraphs, various stages of meiosis will be discussed with particular emphasis on prophase I, the stage during which many of the important and unique events of meiosis are accomplished.

1.1.1. Pre-prophase

Meiosis is preceded by a G1/G0 pre-meiotic stage, characterised by expansion of the nucleus (Zickler & Kleckner, 1998). In yeast and some other organisms, homologous chromosomes pair during this stage (Weiner & Kleckner, 1994). However, this is loosened or lost during the ensuing meiotic S-phase, only to be restored in the first meiotic prophase (Weiner & Kleckner, 1994). The pre-meiotic S-phase is defined by diffused chromatin, which often contains strongly staining foci (Zickler & Kleckner, 1998). The chromosome content of the dividing cell replicates during this stage and connections between sister-chromatids start to develop (Zickler & Kleckner, 1998). It is worth noting that the meiotic S-phase is very similar to its mitotic counterpart, but usually takes much longer (Zickler & Kleckner, 1998).

1.1.2. Early to mid-prophase

1.1.2.1. Overview

The first meiotic prophase comprises several distinct stages, namely, leptotene, zygotene, pachytene and diplotene. Prior to leptotene, chromosomes undergo a cycle of compaction in several organisms, the degree of which varies considerably (Zickler & Kleckner, 1998). During leptotene, chromosomes appear as long, thin, thread-like structures, such that the overall chromosome content of individual cells appears as a tangled mass of such threads. The zygotene stage of prophase follows, during which chromosomes become shorter and fatter. Homologous chromosomes move close to each other and pairing commences. By pachytene, homologous chromosomes, which appear even shorter and thicker, have synapsed and are tightly associated with one another. Many important changes in chromosome configurations occur during these early stages of prophase and crucial interactions develop between sister-chromatids as well as between homologous chromosomes, as outlined below.

1.1.2.2. Events along each pair of sister-chromatids

During leptotene, sister-chromatids of an individual chromosome are organised into a single linear array of loops, which are tethered at their bases to a bulky proteinaceous structural axis (Rattner, Goldsmith & Hamkalo, 1981). The component or even ‘sliver’, of this axis that anchors these chromatin loops constitutes the ‘axial element’. Axial elements of homologous chromosomes eventually come together and become intimately connected along their lengths by proteins that make up the central region of the synaptonemal complex (SC) (Roeder, 1997). Axial elements in a mature SC are referred to as lateral elements. Structure and functions of the SC are described in detail in section 1.1.2.4.

The presence of a single, ‘conjoined’ sister/sister axis is unique to meiosis, as is the asymmetric positioning of chromatin loops with respect to this axis. These features of meiotic cells are likely to facilitate the inter-homologue interactions that commence during or after the organisation of chromatin into this loop-configuration (Kleckner, 1996). An axis-associated kinetochore, which constitutes the point of attachment to the spindle fibres that in turn pull chromosomes apart during anaphase, also starts developing in early prophase and is shared by the two sister chromatids (Goldstein, 1982).

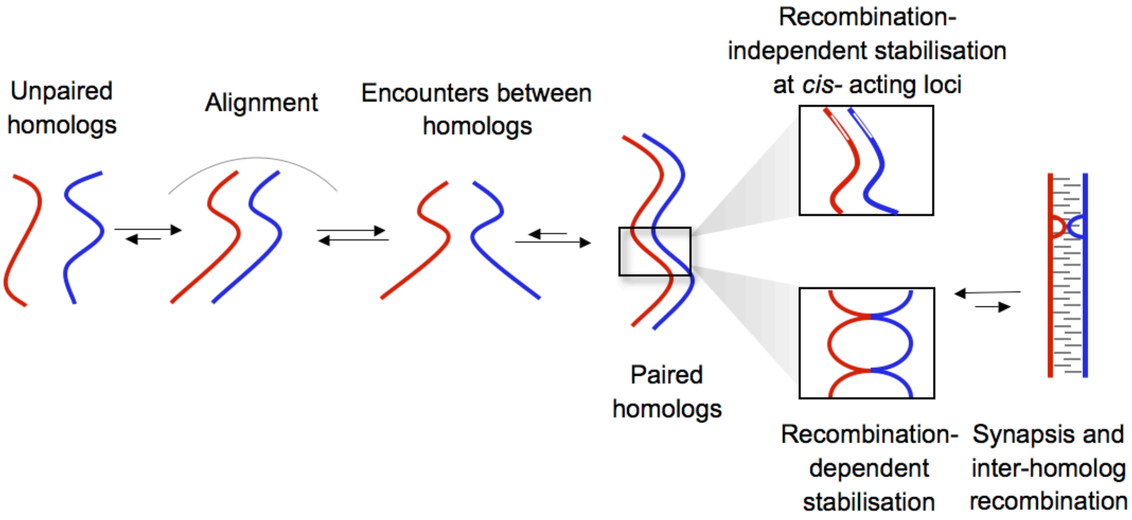
1.1.2.3. Interactions between homologous chromosomes

How homologous chromosomes find, recognise and synapse with one another is an area of much interest and discussion. Whilst many details pertaining to this remain to be elucidated, a very general overview of the likely progression from unpaired to fully conjoined homologous chromosomes is shown in Figure 1.2. The precise usage of terms described in this illustration varies in the literature and in this instance, definitions suggested by Bhalla & Dernburg (2008) are utilised.

The first process described in Figure 1.2 (see following page), that is ‘alignment’ of homologues, commences at the onset of prophase and refers to the polarisation of chromosomes within the nucleus. This polarisation is mediated through interactions with the nuclear envelope (depicted as a grey arc in Figure 1.2). Chromosome arms lie approximately parallel to each other during this stage. Ensuing ‘encounters’ mark transient and reversible interactions between various loci (either homologous or non-homologous) and represent early attempts at establishing stable homologue connections. The stabilisation of these homologue encounters is referred to as ‘pairing’ and sometimes also as ‘close, stable homologue juxtapositions’ (Peoples *et al.*, 2002). Pairing between homologues is usually stabilised through recombination-dependent mechanisms, that is, through associations between homologue axes. Nonetheless some organisms (notably *C. elegans*) utilise recombination-independent mechanisms for this purpose, such as interactions at *cis*-acting loci referred to as ‘pairing centres’. Both these mechanisms are briefly discussed in section 1.1.2.4. The pathway concludes with 'synapsis' between homologues. 'Synapsis' marks the assembly of the SC, a proteinaceous structure that develops between homologues and reinforces chromosomal interactions (see section 1.1.2.4). This is accomplished by pachytene.

Figure 1.2 Intermediates in a 'pairing pathway' of meiosis

(illustration adapted from Bhalla & Dernburg, 2008)


Recombination-
independent stabilisation
at *cis*- acting loci
Unpaired
homologs
Alignment
Encounters between
homologs
Paired
homologs
Recombination-
dependent
stabilisation
Synapsis and
inter-homolog
recombination

Homologous chromosomes are depicted in red and blue. Arrows highlight the reversibility of steps in this pathway. Reversibility allows homology to be assessed at several points, such that inappropriate connections are rejected.

One extensively studied and prominent mechanism that has been suggested to promote homologue alignment and pairing is formation of the 'meiotic bouquet' (Zickler & Kleckner, 1999). In the bouquet configuration, chromosome ends associate with the nuclear envelope in a cluster, whilst the chromosome arms lie nearly parallel to one another (Zickler & Kleckner, 1998). The 'bouquet' stage appears to consist of three distinct, but correlated processes; namely, attachment of chromosome ends to the nuclear envelope, clustering of chromosome ends and finally, cytoskeleton-mediated chromosome movements (Bhalla & Dernburg, 2008). All three processes apparently contribute towards homologue alignment and ultimately pairing. Interestingly, the importance of the meiotic bouquet varies considerably. For example, it plays a minor role in *S. cerevisiae*, but appears to be more important in the fission yeast (*S. pombe*) (Goldman & Lichten, 2000). The bouquet configuration dissolves by the end of pachytene and chromosome ends become evenly distributed around the inner surface of the nuclear envelope (Zickler & Kleckner, 1998).

1.1.2.4. Recombination, assembly of the SC and appearance of recombination nodules

Meiotic recombination or more specifically, crossing-over between homologues is employed by most sexually reproducing organisms as a means of connecting homologous chromosomes during prophase I of meiosis (Page & Hawley, 2003). These connections, referred to as 'chiasmata', are essential for faithful homologue segregation during the ensuing first meiotic division in most organisms studied. Recombination, including its initiation, the processing and resolution of intermediates and various pathways involved are discussed in detail in section 1.2 of this chapter. Nonetheless, two unique and significant features of prophase I that are concomitant with various events in the recombination pathway are briefly considered here. These are formation of the SC and the appearance of early and late nodules.

SC formation in the budding yeast and in most other organisms studied is subsequent to and dependent on recombination-initiating double-strand breaks (DSBs). Conversely, mutations affecting structural components of the SC result in recombination defects (Zickler & Kleckner, 1998). This inter-dependence highlights the existence of an intimate and complex relationship between the two (Börner, Kleckner & Hunter, 2004) and is further discussed in subsequent parts of this Chapter. Presently however, the structure and likely functions of the SC are briefly described.

The SC is a tripartite proteinaceous structure, comprising of the axial (or lateral) elements of homologous chromosomes and a central element, which bridges the two axial elements together (as shown in Figure 1.3). In *S. cerevisiae,* formation of the central element at sites of recombination is primarily marked by the loading of transverse filament protein Zip1 and is dependent on proteins Zip2 and Zip3 (Sym, Engebrecht & Roeder, 1993; Fung *et al.*, 2004). In fact, Zip1, Zip2 and Zip3 form part of a synapsis initiation complex (SIC), which is not only required for SC assembly, but also for crossover (CO) formation (discussed in section 1.2.5). Transmission electron microscopy data suggest that the lateral elements in a mature SC are separated by ~100nm (Zickler & Kleckner, 1999). A nucleus with fully formed SC is at pachytene.

Figure 1.3 The structure of the synaptonemal complex

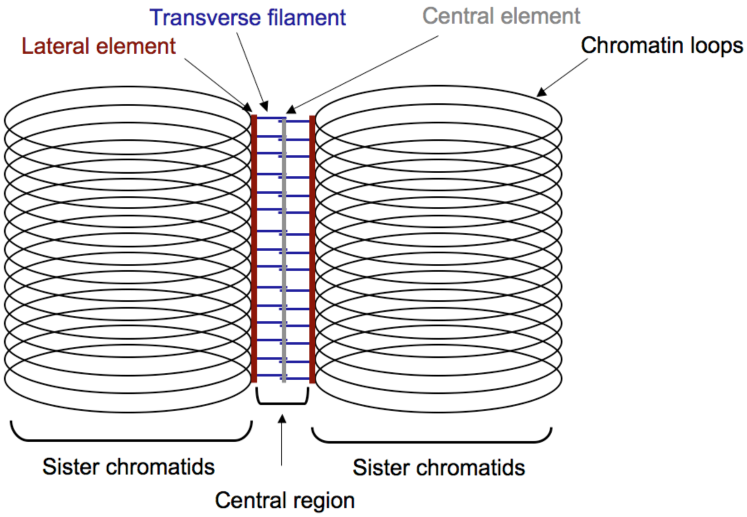


Figure shows the structure of a mature SC and is adapted from Roeder (1997).

The full spectrum of functions of the SC is still not known. It is thought to play a role in facilitating certain steps in recombination and stabilising homologue interactions (Zickler & Kleckner, 1998). It has been also been suggested that the SC is involved in maintaining the axis-association of COs (Zickler & Kleckner, 1998). Many earlier studies supported a role for the SC in mediating CO interference, the phenomenon by which the presence of a CO deters occurrence of others in its vicinity (Maguire & Riess, 1994; Egel, 1995; Sym & Roeder, 1994, 1995; Sym *et al.*, 1993). However, recent studies, mostly on *zip1* mutants show that interference persists even in the absence of a fully formed SC, thereby excluding such a role (Fung *et al.*, 2004; discussed further in section 1.2.6).

It is important to mention here that whilst events in early recombination, namely formation of DSBs and strand invasion are essential for synapsis in budding yeast, plants and mammals, this is not true for all organisms (Peoples-Holst & Burgess, 2005). Indeed, *C. elegans* and *D. melanogaster* employ recombination-independent mechanisms to initiate SC formation between homologous chromosomes (Dernburg *et al.*, 1998; McKim *et al.*, 1998). In *C.elegans*, for instance, certain sequence motifs near chromosome ends, referred to as 'pairing centres' or 'homologue recognition regions', facilitate pairing stability and formation of the SC in-*cis* and are required for chromosome segregation during meiosis (Phillips *et al.*, 2009).

From late leptotene through to pachytene, two different types of dense structures or 'nodules' also appear in association with the SC. Nodules that appear between late-leptotene/early-zygotene and early-pachytene are called 'early nodules' and are usually small and irregular in shape. These nodules are either associated with axial elements prior to formation of the SC or sit atop the tripartite structure, thus indicating an element of flexibility in their attachment to the chromosome axes (Sherman *et al.*, 1992). Some evidence supports a direct role for early nodules in the initial juxtaposition/ pairing of homologous chromosomes (Albini & Jones, 1987). They are distributed evenly and most likely correspond to sites of DSB-formation (Carpenter, 1988).

'Late nodules' or 'recombination nodules', on the other hand, appear during pachytene and tend to be larger and less frequent. They are never seen in the absence of the SC and are located either atop or inside the central region (Wettstein, Rasmussen & Holm, 1984; Zickler & Kleckner, 1998). Distributions of late nodules correlate with that of COs and exhibit the phenomenon of interference (Sherman & Stack, 1995). Both types of nodules are transient and often co-localise to the same genomic regions, suggesting that the two structures are related, with some early nodules subsequently differentiating into late nodules.

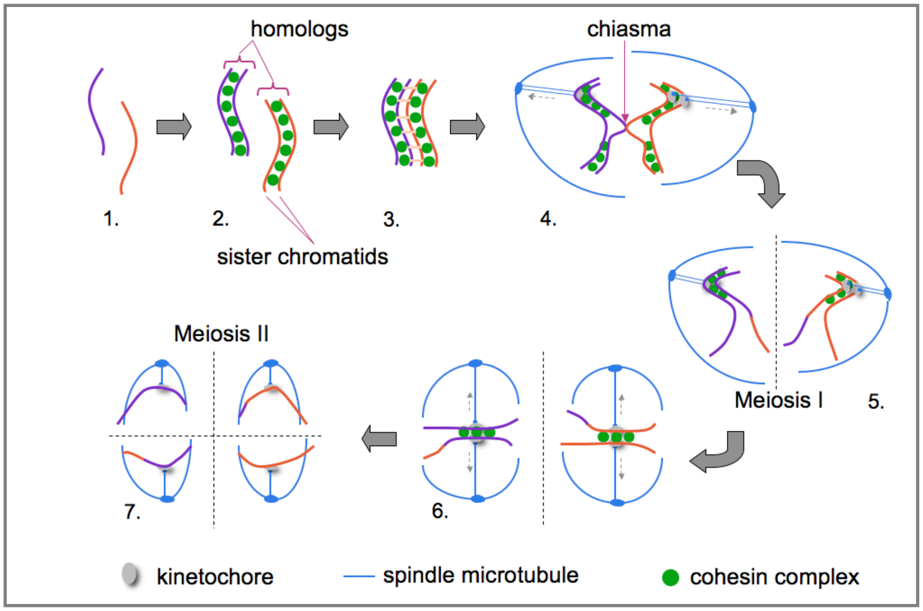
1.1.3. Events ensuing during and after late prophase

The diplotene stage of prophase commences after pachytene. This stage is characterised by a progressive disassembly of components of the SC and associated nodules. Axial elements are also lost either during or shortly after this stage and axes of homologous chromosomes are reorganised (Heyting & Dietrich, 1992). Homologues become separated at this stage, except at points of attachment called 'chiasmata'. These represent sites of crossing-over between non-sister chromatids of homologous chromosomes and are significant in keeping homologues together until anaphase I (Jones, 1984).

Diplotene gives way to the final stages of prophase, diakinesis and prometaphase, during which bivalents continue to become shorter and more compact. Thus, homologues are fully condensed by the time they align on the metaphase I spindle. Kinetochores of sister chromatids are either morphologically or behaviourally unitary at this stage to facilitate attachment of both sisters to the same spindle pole (Nicklas, 1977). Homologous chromosomes, each comprising two sister chromatids connected at the centromere, then segregate towards opposite spindle poles during anaphase I. Thereafter, events of the second meiotic division ensue, which are mostly analogous to mitotic cell division. A very simplified overview of the two meiotic divisions is shown in Figure 1.4.

Figure 1.4 Key stages of the meiotic cell division

(figure adapted from Lao & Hunter, 2010)



Homologous chromosomes in a diploid pre-meiotic cell are shown in orange and purple (1). Chromosomes replicate prior to prophase I, resulting in a pair of sister chromatids, which are held together by ‘cohesin complexes’. This close association is important for keeping sister chromatids together during the first meiotic division (2). During prophase I, homologous chromosomes pair and eventually synapse along their lengths; crossing over occurs at this stage (3). Chiasma that form at the site of crossing-over link homologous chromosomes and promote stable bipolar spindle attachment, which is important for faithful segregation during meiosis I (4). Next, homologues are pulled towards opposite spindle poles in meiosis I. Grey arrows indicate direction of pulling forces created by spindle microtubules (5). Sister chromatids still remain attached at the centromeres during this stage. This is important for the bi-polar attachment of sister-chromatid pairs to the meiosis II spindle (6). Finally, sister chromatids segregate to opposite spindle poles in meiosis II, in a way similar to mitosis. Note: dotted lines in stages 5, 6 and 7 represent the planes of cell division.

1.2. Meiotic recombination in *S. cerevisiae* - The players, processes and pathways

The initiation of meiotic recombination and major steps along the recombination pathway have been best characterised in *S. cerevisiae.* Thus, in the following paragraphs, the current understanding of mechanisms governing the initiation of recombination, the processing of recombination intermediates and their eventual resolution in *S. cerevisiae*, will be discussed in some detail.

1.2.1. Spo11 creates recombination-initiating DSBs

At the very core of meiotic recombination is the event that initiates the process, that is, formation of programmed DSBs. DSBs in *S. cerevisiae* are created by Spo11, a relative of the archaeal topoisomerase VI, which cuts the DNA in a topoisomerase-like transesterification reaction to generate a transient, covalently linked DNA-protein intermediate (de Massy *et al.*, 1995; Keeney & Kleckner, 1995; Liu *et al.*, 1995; Bergerat *et al.*, 1997, Keeney *et al.*, 1997). Spo11 is highly conserved and initiation of DSBs by Spo11 orthologs has been demonstrated in many organisms, including other fungi, nematodes and mice (Steiner, Schreckhise & Smith, 2002, Dernburg *et al.*, 1998; Baudat *et al.*, 2000). It is important to note however that in *spo11* mutants, meiotic recombination can be at least partially rescued by DSBs introduced via other DNA damage mechanisms, for example, through ionising radiation (Thorne & Byers, 1993).

Furthermore, Spo11 alone is not sufficient for the generation of meiotic DSBs and at least nine other proteins are required, namely, Ski8, Rec102, Rec104, Mer2, Mei4, Rec114, Mre11, Rad50 and Xrs2 (Keeney, 2007; Gardiner *et al.*, 1997; Sasanuma *et al.*, 2007; Henderson *et al.*, 2006; Li, Hooker & Roeder, 2006; Usui *et al.*, 1998; Alani *et al.*, 1990; Wiltzius *et al.*, 2005; Tsukamoto *et al.*, 2005). Full levels of DSB-formation also seem to require Red1, Hop1 and Mek1 (Mao-Draayer *et al.*, 1996; Xu, Weiner & Kleckner, 1997).

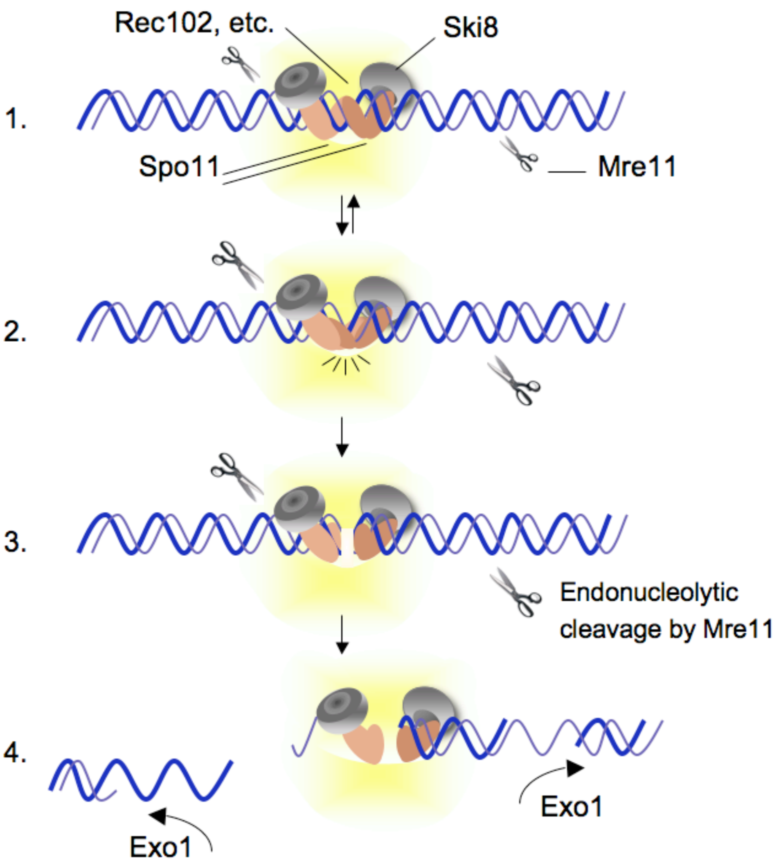
It is likely that these proteins have roles in targeting Spo11 for binding to specific sites, in activating/regulating Spo11 activity, in modifying chromatin structure for DSB formation and/or effectively coupling DSB-formation with its processing (Keeney, 2007). Whilst it is possible that all these proteins form stoichiometric subunits of a single holoenzyme, studies to date are more consistent with them forming distinct functional subunits, which somehow collaborate to form meiotic DSBs (Keeney, 2007). These subunits most likely include Spo11-Ski8, Rec102-Rec104, Mer2-Mei4-Rec114 and Mre11-Rad50-Xrs2 (Keeney, 2007).

1.2.2. Early processing of meiotic DSBs

Current data suggest that after formation of the DSB, Spo11 must be removed from the break to allow further processing. This is accomplished by an endonucleolytic mechanism, wherein nicks are introduced adjacent to the site of DSBs, thereby releasing Spo11, which is still covalently linked to a short oligonucleotide sequence (Neale *et al.*, 2005; Figure 1.5). Studies in *S. cerevisiae* and in *S. pombe* show that nucleases Sae2, Mre11, as well as Rad50 are required for Spo11 removal (Neale *et al.*, 2005). In fact, accumulation of unrepaired DSBs in *rad50s*, *mre11s* and *sae2*Δ mutants, where Spo11 remains bound to DNA, were key in the identification of Spo11 as the initiator of meiotic DSBs (de Massy *et al.*, 1995; Keeney & Kleckner, 1995; Liu *et al.*, 1995). In *S. cerevisiae*, Spo11-associated oligonucleotides are variable in length, falling into two distinct categories of either 7-12 nucleotides or 21-37 nucleotides (Neale *et al.*, 2005).

Nicks introduced by Mre11 and/or Sae2 provide the starting point for further 5′-> 3′ resection. Recent studies show that Exo1 is essential for this resection and the extent of DSB-resection is dramatically reduced, but not completely abolished, in *exo1 Δ* null mutants (Zakharyevich *et al.*, 2010; Keelagher *et al.*, 2011). Zakharyevich *et al.* (2010) suggest that Exo1 removes a minimum of ~350 nucleotides (average being ~550 nt) in a single progressive 5′-> 3′ excision reaction to expose 3′- ended single-stranded tails. It is worth noting Exo1 also appears to play a role in facilitating CO formation, independent of this nuclease activity (Zakharyevich *et al.* 2010). A model suggested for the early steps in the recombination pathway, including DSB-formation and early processing of these breaks, is summarised in Figure 1.5 (Keeney, 2007).

Figure 1.5 A model for the mechanism of formation and initial processing of meiotic DSBs (adapted from Keeney, 2007)



A model for DSB-formation, proposed by Keeney (2007) is shown. **1.** A 'pre-DSB complex', consisting of Spo11, Ski8 and at least a sub-set of other proteins required for DSB-formation (called 'Rec102, etc.'; shown as a cloud encompassing the Spo11 complex) assembles on DNA. **2.** Spo11 is reversibly activated to cleave DNA, probably through a conformational change in the Spo11 complex, but has not done so yet. **3.** The activated Spo11 complex cleaves DNA and in a concerted reaction, Mre11 nicks the DNA asymmetrically to initiate DSB-processing. This cleavage by Mre11 commits the site to undergo recombination. **4.** Exonucleolytic processing of 5´ends by Exo1 commences and is most probably coupled to cleavage by Mre11. Other models for resection by Exo1 have been suggested more recently (eg: Zakharyevich *et al.*, 2010), but are not discussed here.

1.2.3. Identifying an intact homologous partner and strand invasion

The 3´ single-stranded tails resulting from extensive 5´-> 3´ resection serve as substrates for two RecA family proteins, namely Rad51, which is ubiquitous and is required for recombination repair in mitotic cells, and Dmc1, which is meiosis-specific (Bishop *et al.*, 1992; Dresser *et al.*, 1997; Shinohara *et al.*, 1997). Both bind the resected single-stranded DNA (ssDNA) to form nucleoprotein filaments. These nucleoprotein filaments mediate the search for homology and also catalyse DNA strand invasion to generate heteroduplex DNA (hDNA). Unlike RecA however, both Rad51 and Dmc1 have little preference for ssDNA over double-stranded DNA (dsDNA) and need mediators to displace the ssDNA-binding protein RPA from the resected DNA and facilitate nucleation and filament formation (Bianco *et al.*, 1998; Li *et al.*, 1997). Thus, formation of Rad51 filaments is dependent on Rad52, Rad55-Rad57 and to a lesser extent Rad54, whereas formation of Dmc1 filaments requires Se3-Mei5, Hop2-Mnd1 and Rhd54-Tid1 (Ehmsen & Heyer, 2007). Immunofluorescence data, particularly in lilies, suggest that both Rad51 and Dmc1 form foci, which are components of early nodules discussed in section 1.1.2 (Anderson *et al.*, 1997; Bishop, 1994).

Once nucleoprotein filaments are formed, three potential partners, including a sister chromatid and two non-sister chromatids of the homologous chromosome, are available for the next step in the recombination pathway, namely, strand invasion. The sister chromatid is preferentially used in recombination repair of DSBs in mitotic cells (Kadyk & Hartwell, 1992). However, for meiotic recombination to play its role in pairing and segregation, strand exchange must take place with a homologue and not a sister chromatid. An active inter-homologue bias, which ranges from 3:1 to more than 7:1 in the budding yeast, is maintained for this purpose (Lao & Hunter, 2010). In the budding yeast, Rad51, Dmc1, as well as several of their mediators appear to be involved in creating this bias (Schwacha & Kleckner, 1997; Zierhut *et al.*, 2004; Sheridan & Bishop, 2006). In addition, components of a phosphokinase signal transduction pathway that respond to meiotic DSBs and monitor progress through prophase I, also appear to play a role in maintenance of the inter-homologue bias. This includes core DNA damage response factors, Mec1/Tel1, as well as meiosis-specific Hop1, Red1 and Mek1 (Carballo *et al.*, 2008; Schwacha & Kleckner, 1997; Thompson & Stahl, 1999; Niu *et al.*, 2005; Niu *et al.*, 2007; Callender & Hollingsworth, 2010).

Rad51/Dmc1-mediated initial strand exchange is common to all recombination interactions. However, mounting evidence suggests that thereafter, events resulting in COs and in NCOs are processed via different routes, as discussed below.

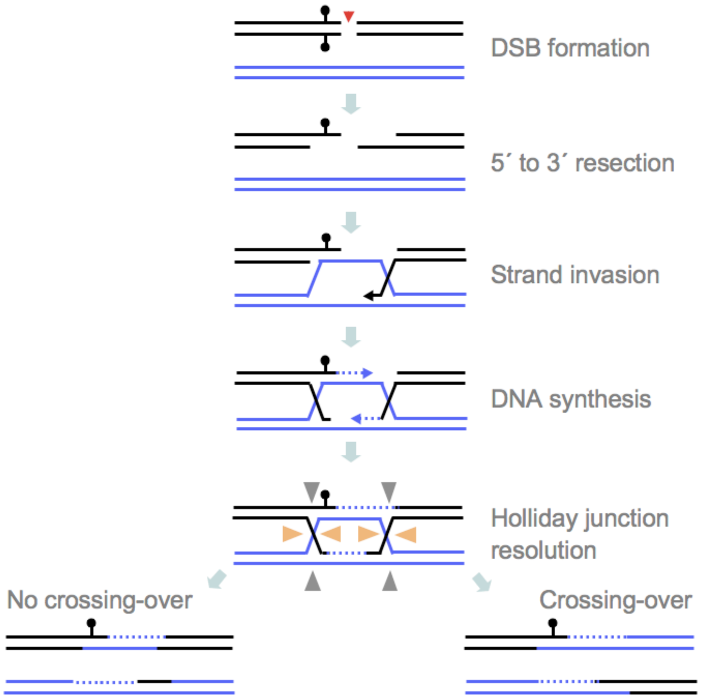
1.2.4. Multiple pathways of recombination in the budding yeast

In the yeast *S. cerevisiae*, there are two possible outcomes of meiotic recombination. A ‘crossover’ (CO) refers to the reciprocal exchange of DNA between non-sister chromatids of homologous chromosomes. COs are accompanied by a tract subject to gene conversion (GC), which refers to the non-reciprocal transfer of DNA from one molecule to its homologue (Pâques & Haber, 1999). The alternative outcome of recombination is a ‘noncrossover’ (NCO), which refers to a GC tract that is not associated with reciprocal exchange or crossing-over.

It is important to note here that usage of the term ‘gene conversion’ appears to vary in the literature and in some instances, it is used to describe the non-reciprocal transfer of DNA without exchange of flanking markers. These events are referred to as ‘noncrossovers’ in this work for purposes of clarity and to distinguish such events from gene conversions accompanying COs.

Since Robin Holliday’s seminal proposal in 1964, virtually every model for meiotic recombination, including the widely canonized ‘double-strand break repair’ (DSBR) model for meiotic recombination, has contained a definite link between CO and NCO outcomes of recombination (Resnick, 1975; Szostak *et al.*, 1983; Sun *et al.*, 1991). As shown in Figure 1.6, the DSBR model predicts that Rad51/Dmc1-mediated strand exchange is followed by second-end capture. DNA synthesis and ligation then generates a ‘double Holliday junction’ (dHJ), which contains heteroduplex DNA (hDNA) on either side of the break. The dHJ is resolved by cutting either the non-crossing (outside) strands or the crossing strands at each junction to yield CO or NCO products. According to this model, both end products of recombination contain hDNA, which upon mismatch repair yields gene conversion with and without exchange.

Figure 1.6 The double-strand break repair model of meiotic recombination


DSB formation
5´ to 3´ resection
Strand invasion
DNA synthesis
Holliday junction resolution
No crossing-over
Crossing-over

A homologous chromosome pair is shown (in blue and black). The lollipop indicates a potential mismatch. It should be noted that dHJs generated in this pathway can be resolved by cutting either the non-crossing strands (indicated using grey arrows) or the crossing strands (indicated using orange arrows). Note: only two of four possible resolution of the dHJ are shown here.

The DSBR model is supported by a number of physical studies verifying the presence of different intermediates along the pathway, including formation of hDNA and the generation of four-stranded DNA structures containing dHJs, commonly referred to as 'joint molecules' (Goyon & Lichten, 1993; Nag & Petes, 1993; Collins & Newlon, 1994; Schwacha & Kleckner, 1994, 1995). Nonetheless, whilst the DSBR model predicts that it is the orientation of resolution of dHJs that determines the outcome of recombination and that both CO and NCO products appear at the same time following resolution, several more recent studies challenge these tenets of the model. Some of these are briefly mentioned here.

First of all, a particular set of proteins, commonly referred to as 'synapsis initiation complex' (SIC) proteins or ZMMs (Zip1-Zip2-Zip3-Zip4-Spo16, Msh4-Msh5, Mer3) specifically influence the formation of COs as well as SC assembly, and various ZMM mutant yeast strains show ~80% reduction in CO levels whilst levels of DSBs and NCOs remain unaffected (Lynn, Soucek & Börner, 2007; Shinohara *et al.*, 2008). Futher, CO-formation specifically requires mismatch repair proteins Mlh1-Mlh3 (Hunter & Borts, 1997; Wang, Kleckner & Hunter, 1999).

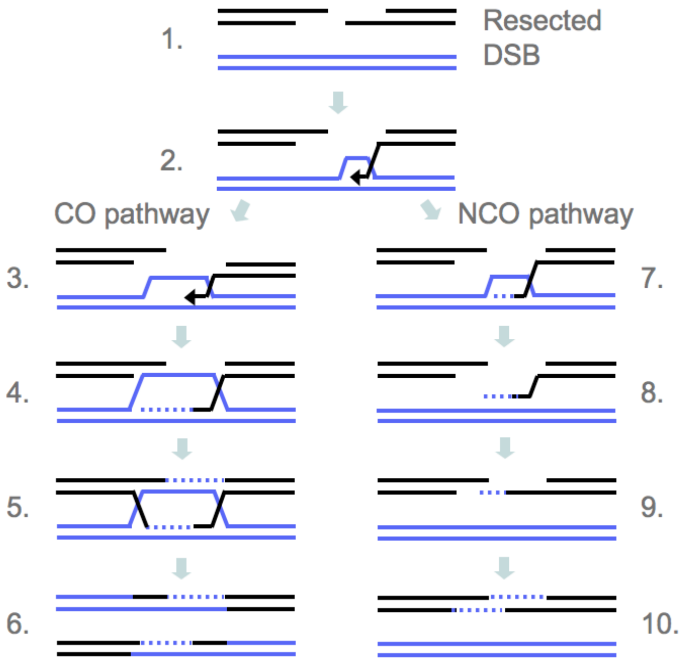
Second, the DSBR model predicts that hDNA will be present on both DNA molecules emerging from a single recombination event on opposite sides of the DSB-site. However, a study by Gilbertson & Stahl (1996) showed that this was rare. Moreover, configuration of hDNA adapted by NCOs did not conform to what was expected from the DSBR model.

Third, Allers & Lichten (2001) used physical detection methods to show that in budding yeast, NCOs appear at the same time as JMs and at least 30 minutes prior to the appearance of COs. This was contrary to the expectation from the DSBR model, which predicts that COs and NCOs are formed at the same time. The same study also investigated CO and NCO-formation in *ndt80* mutants. Ndt80 is a meiosis-specific transcription factor and is essential for the completion of prophase I (Chu *et al.*, 1998). In these mutants, JMs are not processed and accumulate to high levels. Whilst CO levels are substantially reduced in these strains, NCOs show normal levels and kinetics of appearance. Based on this, Allers & Lichten (2001) suggested that JMs predominantly give rise to COs, whilst NCOs arise from a distinct pathway that does not involve long-lived Holliday junction intermediates, for example, through synthesis-dependent strand annealing (SDSA) (as shown in Figure 1.7 on the following page).

Finally, studies conducted on yeast strains lacking ZMM proteins showed that at higher temperatures of 33ºC, as opposed to 30ºC, both DSBs and NCOs form normally. However, the formation of a prominent strand-exchange intermediate, the 'single-end invasion' (SEI), as well as dHJs, COs and the SC are adversely affected (Börner, Kleckner & Hunter, 2004). This indicates that the CO and NCO pathways in meiosis differentiate early, prior to the onset of stable strand exchange and independent of the SC. Collectively, data from these studies support an ‘early crossover decision’ (ECD) model for recombination, illustrated in Figure 1.7.

Figure 1.7 The early crossover decision (ECD) model of meiotic recombination

(adapted from Bishop & Zickler, 2004)



1. A recombination-initiating DSB is first resected to produce 3´single-stranded overhangs. **2.** According to the ECD model, a resected DSB end first forms a nascent interaction with a homologous chromosome and thereafter, CO/NCO pathways separate. **3.** En route to forming a CO, this interaction is stabilised to form a SEI intermediate (that can be visualised using 2-D gels (as shown in Hunter & Kleckner, 2001)). **4.** The second DSB end then engages the SEI by strand displacement. **5.** DNA synthesis and ligation leads to the formation of a dHJ. **6.** The dHJ is resolved such that a CO results. The NCO pathway described here is consistent with the ‘synthesis dependent strand annealing’ (SDSA) model (Nassif *et al.*, 1994). **7.** In this pathway, the invading DSB end is extended by DNA synthesis. **8.** The extended end is then ejected from the intact homologue and anneals with its partner (**9**). **10.** DNA synthesis and ligation follows, such that a NCO results. Both CO and NCO pathways, particularly in the context of correction of hDNA, will be discussed further in other parts of this work.

Whilst formation of NCOs via SDSA has been evidenced in several mitotic systems, it has not been unequivocally verified in meiosis. Thus, although the SDSA model accounts for and is consistent with the available data, it remains formally possible that NCOs form via resolution of short-lived and generally undetectable JMs (Allers & Lichten, 2001).

An additional level of complexity is added by evidence supporting a second, ZMM-independent pathway of CO formation in *S. cerevisiae* (Börner, Hunter & Kleckner, 2004). Recent studies suggest that this pathway is dependent on structure-specific endonuclease complex Mms4/Mus81 (Hollingsworth & Brill, 2004). Although Mms4/Mus81 can cleave dHJs, it is more effective cleaving other branched structures, raising the possibility that this pathway involves distinct recombination intermediates (Osman *et al.*, 2003). ZMM-dependent and Mms4/Mus81-dependent COs also show marked differences in their distribution, with the former showing interference and the latter not (discussed further in the following section).

It is worth mentioning that whilst the Mms4/Mus81 pathway is a minor pathway of CO formation in *S. cerevisiae* accounting for ~15% of COs, (Börner, Hunter & Kleckner, 2004), it is the major pathway for CO formation in *S. pombe* (Osman *et al.*, 2003). In fact, pathways used to generate COs and their relative usage varies considerably amongst organisms. For instance, *D. melanogaster* utilises an entirely different pathway dependent on Mei9 (Sekelsky *et al.*, 1995).

1.2.5. Regulation of CO-formation: The obligate CO and interference

Faithful segregation of homologous chromosomes at meiosis I require them to be physically connected at points of crossing-over, referred to as chiasmata. It is thus perhaps not surprising that the distribution of these essential COs is highly regulated. This regulation is characterised by two important features; namely the phenomenon of obligate CO/chiasmata, whereby each homologue pair receives at least one CO, and the phenomenon of interference, whereby presence of a CO decreases the likelihood of another in its vicinity (Jones, 1984).

For a long time it was suggested that interference is mediated either by the assembly of the SC and/or via signal transduction through an assembled SC (Shinohara *et al.*, 2003). However, several lines of evidence argue against a role for the SC in mediating CO designation and interference; a few of these are briefly mentioned here.

First, a study by Börner, Kleckner & Hunter (2004) showed that at higher temperatures (33ºC as opposed to 30ºC) *zmm* mutants, which show defects in CO formation and SC assembly, form normal levels of NCOs. This implies that the CO/NCO decision and thus, designation of COs occurs independently of the SC. Also, in wild-type meiosis ZMM protein Zip2 localises specifically to future sites of crossing-over and displays an interference pattern (Fung *et al.*, 2004). Interestingly, this pattern is maintained in *zip1* mutants, which are devoid of SCs (Fung *et al.*, 2004). Finally, residual COs in *spo16* and *spo22* (*zip4*) mutants, which are defective in SC formation, show a normal interference distribution, again highlighting that a fully formed SC is not required for CO interference (Shinohara *et al.*, 2008).

An attractive alternative SC-independent model for CO interference is the 'mechanical stress' model (Kleckner *et al.*, 2004). This model proposes that programmed expansion of chromatin gives rise to compression stress in the form of a pushing force along each homologue axis. This stress eventually causes the axis to buckle preferentially at 'bridges' along the axis, which represent sites wherein a DSB-end is engaged in a nascent interaction with a homologous partner and as such constitute 'flaws' within the axis (Hunter & Kleckner, 2001; Tessé *et al.*, 2003). Buckling at one end of the bridge linking the two axes causes the recombination complex involved to undergo a conformational change that commits it towards becoming a CO (Börner, Hunter & Kleckner, 2004).

Axis buckling relieves local compression stress. This relaxation then spreads outwards in both directions from the site of buckling, thus preventing additional buckling and as a consequence, CO formation at a nearby site (Börner, Hunter & Kleckner, 2004). This model thus accounts for both CO assurance, as well as interference. Additionally, it supports a role for the SC in mediating stable exchange, both at the DNA level (SEI to dHJ transition) and at the axis level, via twisting of the SC along its length under mechanical stress (Börner, Hunter & Kleckner, 2004).

1.3. Distribution of recombination events in *S. cerevisiae*

Recombination events in *S. cerevisiae* are not distributed evenly throughout the genome and various regions that are recombinationally either 'hot' or 'cold' have been identified. Recombination 'hot spots' refer to regions or loci with greater than average frequency of meiotic recombination (Goldman & Lichten, 1995). On account of their elevated activity, these hot spots allow easy experimental access to recombination events and have proved instrumental for elucidating recombination dynamics and characterising recombination intermediates. Thus, the current knowledge of yeast recombination hot spots, their features and distribution will be briefly considered here. In a broader context, the genome-wide distribution of COs and NCOs will also be discussed for a holistic picture of recombination in this important model organism.

1.3.1. Features and classification of yeast hot spots

Yeast hot spots possess certain distinctive features and can be generally classified into three sub-types. Hot spots *ARG4* and *HIS4* have been most extensively studied in budding yeast. Thus, a substantial part of the following discussion is based on data from these hot spots.

1.3.1.1. Yeast hot spots lie within chromatin that displays nuclease sensitivity

Several studies indicate that natural as well as artificial yeast hot spots are located within regions of chromatin that display DNase I and micrococcal nuclease (MNase) sensitivity (Ohta, Shibata & Nicolas, 1994; Wu & Lichten, 1994, 1995). This implies that recombination-initiating DSBs tend to occur within regions wherein nucleosomes are disrupted or absent and the DNA is exposed. At some loci, this nuclease sensitivity is further exaggerated prior to the formation of DSBs. Furthermore, hotspot sequences that alter recombination activity similarly influence nuclease sensitivity (Ohta, Shibata & Nicolas, 1994).

It is important to note however that whilst favourable chromatin structure may be necessary for hotspot activity, it alone is not sufficient to confer activity. In some cases notably at the *ARG4* hot spot, chromosomal position seems to play a role (Wu & Lichten, 1995). Thus, recombination activity at the *ARG4* hot spot varies at least four-fold depending on the chromosomal location of *ARG4* sequences. In fact, in certain chromosomal contexts, the *ARG4* promoter region receives no DSBs, despite maintaining nuclease sensitivity (Wu & Lichten, 1995).

1.3.1.2. Yeast hot spots are preferentially located within gene promoters

Many yeast hot spots are located within regions corresponding to gene promoters (Nicolas *et al.*, 1989; Detloff, White & Petes, 1992). In fact, both *HIS4* and *ARG4* hot spots display gradients of GC, referred to as 'polarity gradients', with peaks in activity corresponding to respective gene promoter regions (Nicolas & Petes, 1994). Consistently, deletions of sequences upstream of *HIS4* and *ARG4* genes reduces recombination and the gradient of GC, most probably as a result of removing and/or inactivating an element important for recombination initiation (Nicolas *et al.*, 1989; Schultes & Szostak, 1990; Detloff, White & Petes, 1992; de Massy & Nicolas, 1993).

1.3.1.3. The three classes of yeast hot spots

Given the association of hot spots and promoter regions, it is perhaps not surprising that some hotspot-loci, notably *HIS4* require the binding of specific transcription factors for activity (Arndt, Styles & Fink, 1987). Interestingly however, a high level of transcription itself is not required and deletions that substantially reduce levels of transcription have no effect on the frequency of recombination events (White *et al.*, 1992). These hot spots, whose activity depends on the binding of transcription factors but not on transcription itself, have been termed ‘α- hot spots’ (Kirkpatrick, Fan & Petes, 1999).

Another class of yeast hot spots reside within nuclease-sensitive chromatin that is not associated with transcription-factor binding. For example, insertion of a nucleosome-disrupting $(CCGNN)_{12}$ tandem repeat unit stimulates recombination, despite the fact that no known transcription factor binds this sequence (Kirkpatrick *et al.*, 1999). It therefore seems likely that this tandem repeat unit stimulates recombination via a mechanism distinct to what is used by α- hot spots; this sub-set of yeast hot spots, created by nucleosome-excluding sequences are referred to as ‘β- hot spots’ (Kirkpatrick *et al.*, 1999).

Finally, in 2000, Gerton *et al.* showed a significant association between recombination hot spots and base composition of yeast chromosomes, with many hot spots occurring near or within peaks of high ‘G+C’ content. This third sub-set of yeast hot spots is referred to as ‘γ-hot spots’ (Petes, 2001). It should be noted that despite association with specific base compositions, yeast hot spots do not correlate with any specific sequence motifs. This is in contrast to human and mouse hot spots, as discussed in later sections (Petes, 2001).

1.3.2. Global distribution of COs and NCOs in *S. cerevisiae*

Recent years have witnessed a steady progress in understanding the genome-wide distribution of recombination events in yeast, as well as in higher eukaryotes, including mammals (described in later sections). These studies are not only shedding valuable light on the relative positioning of COs and NCOs, but revealing several intriguing phenomena which would otherwise go undetected, uncovering regions of interest for further investigation and ultimately helping answer questions pertaining to mechanisms governing CO assurance, interference and alternative pathways of DSB-processing and resolution (Mancera *et al.*, 2008). Hence, this final section dedicated to a discussion of meiotic recombination in *S. cerevisiae* will focus on a recent study by Mancera *et al.* (2008), in which both COs, as well as GCs with and without exchange were mapped at a genome-wide level and at very high resolutions, with median distance between markers being just 78 bp.

Mancera *et al.* (2008) analysed the products of 51 meioses in an S288c/YKM789 hybrid yeast strain. Their findings were largely consistent with previous data obtained from studies on one or two loci or chromosomes, or from other lower-resolution genome-wide surveys. First, the numbers of COs and NCOs per meiosis, that is 90.5 and 66.1 respectively, were remarkably similar to the estimated number of DSBs per meiosis which varied between 140 and 170 (Buhler, Borde & Lichten, 2007). Second, 84% of the recombination hot spots identified overlapped a promoter region, whilst regions proximal to centromeres showed low recombination rates, as suggested previously (Baudat & Nicolas, 1997; Gerton *et al.*, 2000; Borde, Wu & Lichten, 1999). Third, median conversion-tract estimates of 2 kb for GCs with exchange and 1.8 kb for NCOs, were consistent with previous estimates from a single hot spot (Borts & Haber, 1989). Interestingly, data from *msh4* null mutants, which specifically affect COs, show that mean tract-length for GCs with exchange is increased by 479 bp compared to wildtype, whilst tract-length for NCOs is decreased by 338 bp (Ross-Macdonald & Roeder, 1994). Thus, Msh4 not only influences genome-wide levels of COs as expected, but also affects tract lengths of GC events.

Mancera *et al.* (2008) also described some previously uncharacterised aspects of genome-wide distribution of recombination events. For instance, they reported that in addition to COs, numbers of NCOs were linearly related to chromosome length, although the phenomenon was less pronounced. They also established that subtle differences between CO/NCO ratios, initially reported Martini *et al.* (2006) at two hot spots, persisted at a genome-wide level. In fact, a significant number of marker intervals collectively spanning ~100 kb show biased CO:NCO ratios ranging between 14:0 and 0:7. The authors suggested that these biases affect at least 1.4% of the genomic regions involved in formation of recombination events. It is worth mentioning that no DNA sequence motifs could be specifically associated with either bias.

Furthermore, in addition to interference between COs, Mancera *et al.* (2008) also reported interference between COs and NCOs. However, both types of interference, that is between COs as well as between COs and NCOs, disappeared in *msh4* null mutants. This was consistent with previous studies that indicated that only Msh4-Msh5-dependent or ZMM-dependent COs displayed the phenomenon of interference (Argueso *et al.*, 2004).

Finally, this study investigated GCs at a genome-wide level in unprecedented detail and suggested that as much as 1% of the genome of a yeast meiotic product may be involved in GCs in a single meiosis. Interestingly, a mismatch repair-bias favouring GC nucleotides was detected for GCs with and without exchange. Such a bias is consistent with and accounts for the association between high GC base-composition and recombination hot spots (Gerton *et al.*, 2000).

Overall, data from this and other similar studies (eg: Buhler, Borde & Lichten, 2007) will no doubt have important bearings on furthering the current understanding of recombination in *S. cerevisiae* and in opening avenues for further research. Also, in conclusion to the discussion of meiotic recombination in *S. cerevisiae*, it is worth re-emphasising the invaluable role studies in this model organism have played in elucidating numerous fundamental features and processes of meiotic recombination, which are widely conserved amongst other organisms, including mammals and thus humans.

More recently however, growing interest in modelling meiotic recombination in a genetically tractable organism whose genome size and complexity closely resembles that of humans, has sparked a plethora of research focussing on meiotic recombination in mice. Key facets and findings of these studies are briefly described in the following section.

1.4. Meiotic recombination in mouse

The pathways of meiotic recombination in *S. cerevisiae* appear to be largely conserved in mice. In this section, the distribution of recombination events, particularly COs, across the mouse genome and features of mouse hot spots and factors that influence their activity and location will be discussed in detail. Some data describing NCOs and the CO/NCO decision in mice will be discussed, although a more detailed analysis of these events is presented in Chapter 7.

1.4.1. Conservation of yeast (*S. cerevisiae*) recombination processes and pathways in mice

In yeast, Spo11-induced DSBs initiate meiotic recombination. In mouse, *Spo11* (*mSpo11*) maps to chromosome 2 and is expressed in juvenile and adult testes, and in oocytes of embryonic ovaries (Keeney *et al.*, 1999; Romanienko & Camerini-Otero, 1999). Data from Keeney *et al.* (1999) indicated that *mSpo11* expression commences in the early stages of prophase I prior to pachytene and steadily increases throughout the pachytene stage, consistent with the gene encoding the functional homologue of budding yeast Spo11. Numerous lines of evidence support the idea that recombination initiation between *S. cerevisiae* and mice is conserved. Mahedevaiah *et al.* (2001) studied phosphorylation of histone H2AX (γ-H2AX), which marks sites of DSB formation in mice and confirmed that this signal, which appeared at leptotene and was thus concomitant with formation of meiotic DSBs, was in fact *Spo11*-dependent.

Second, two populations of SPO11-associated oligonucleotides were detected in wild-type mouse testes by Neale *et al.* (2005), with sizes ranging between ~12-26 and ~24-28 nucleotides respectively. These SPO11-associated oligos are reminiscent of those identified in *S. cerevisiae* and highlight similarities in the initial processing of meiotic- DSBs in the two organisms. Indeed, meiotic DSBs with 3′-ended single-stranded overhangs have been detected in mouse spermatogenesis, similar to those seen in *S. cerevisiae* (Zenvirth *et al.*, 2003). These were not detected in *Spo11*$^{-/-}$ spermatocytes, supporting the requirement of SPO11 for DSB-formation.

Third, several studies showed that like in budding yeast meiosis, strand-exchange proteins RAD51 and DMC1 co-localise into foci during the leptotene and zygotene stages of meiosis in mice (Plug *et al.*, 1996; Moens *et al.*, 1997; Tarsounas *et al.*, 1999). Mice carrying *Rad51* null mutations are embryonic lethal, whilst *Dmc1* null mutations result in failure of homologue synapsis and meiotic arrest at zygotene (Lim & Hasty, 1996; Pittman *et al.*, 1998). Interestingly, *Dmc1* mutants occasionally show synapsis between non-homologous chromosomes, indicating that the role of DMC1 in mediating homologue interactions is probably conserved from *S. cerevisiae* to mice (Yoshida *et al.*, 1998). RAD51/DMC1 foci most probably represent most, if not all, early DSB-repair events and range between 200-400 in number during leptotene (Plug *et al.*, 1996; Tarsounas *et al.*, 1999; Oliver-Bonet *et al.*, 2005).

It appears that the pathways of CO formation in *S. cerevisiae* and mice are also largely similar and involve many homologous players. Thus, MLH1-MLH3 are required for formation of most COs (see below) and co-localise into foci corresponding to positions of chiasmata formation (Marcon & Moens, 2003). Further, several mammalian homologues of *S. cerevisiae* ZMM proteins have been identified, including SYCP1, MSH4 and MSH5 (Sage *et al.*, 1995; Kneitz *et al.*, 2000; Her *et al.*, 1999). Finally, residual CO activity in *Mlh1*$^{-/-}$ mice, corresponding to 5-10% of wild-type levels, suggests presence of an MLH1-independent CO pathway similar to *S. cerevisiae* (Woods *et al.*, 1999; Guillon *et al.*, 2005). Indeed, Holloway *et al.* (2008) reported significant meiotic defects in mice carrying *Mus81* null mutations and proposed at least two pathways of CO formation in mice, with MUS81 responsible for processing at least a sub-set of events. It should be mentioned however, that the precise mechanism of this control, including the stage at which it is imposed, remains to be elucidated.

1.4.2. Properties and distributions of COs in mice

Numerous approaches have been used to study numbers and distributions of COs in mouse. Cytological studies have relied upon detection of chiasmata directly in mouse spermatocytes. The total number of chiasmata in mouse is around 23 per meiosis, indicating that numbers of COs are substantially fewer than numbers of initiating DSBs (see above; Polani 1972; Speed 1977; Jagiello & Fang, 1979; Lawrie, Tease & Hultén, 1995). Immunocytogenetic approaches investigating localisation of MLH1 have also been used to study COs, although these studies have yielded consistently lower estimates of numbers of CO than genetic data (Kolas *et al.*, 2005; Lynn *et al.*, 2005; Koehler *et al.*, 2006; Baudat & de Massy, 2007). These cytological studies highlighted the presence of one CO per chromosome arm in mice, indicating that the phenomenon of obligatory crossing-over is maintained from yeast. These COs also display the phenomenon of interference (Lawrie, Tease & Hultén, 1995; Broman *et al.*, 2002).

COs in mouse are unevenly distributed, clustering into 1-2 kb wide hot spots of varying intensities. The first such hot spot was characterised in 1982 by Steinmetz *et al.* in the H2 region of chromosome 17. Over the last decade or so, high-resolution mapping of recombination events among the progeny of crosses between inbred strains of mice, studies of patterns of haplotype diversity, molecular analysis of recombination hot spots using sperm approaches and genome-wide mapping of recombination-initiating DSBs have facilitated a detailed characterisation of CO events and allowed insights into the levels of regulation imposed on the numbers and distribution of such events (Guillon & de Massy, 2002; Yauk *et al.*, 2003; Kauppi, Jasin & Keeney, 2007; Paigen *et al.*, 2008; Billings *et al.*, 2010; Smagulova *et al.*, 2011). Some important findings from these studies, particularly where parallels can be drawn with humans, are briefly discussed below.

1.4.2.1. Variation in CO activity along chromosomes

Two studies analysing COs on chromosomes 1 and 11, which differ in size and gene density, in crosses between C57BL/6J (B6) and CAST/EiJ (CAST) strains of mice, showed that recombination events along both chromosomes were unevenly distributed (Paigen *et al.*, 2008; Billings *et al.*, 2010). Indeed, recombinationally active regions were separated by intervals (often >1 Mb-long) showing substantially lower rates. Moreover, centromere and telomere-proximal regions usually showed lower and higher-than average recombination activity respectively (Billings *et al.*, 2010).

1.4.2.2. Influence of genetic background on CO activity

Differences in CO frequency and distribution associated with variation in genetic background have been noted by several studies. For instance, comparisons of numbers of MLH1 foci between various strains of inbred mice show subtle but consistent differences (Koehler *et al.*, 2002). Data from mouse crosses further validates this. Thus, a study of CO activity along chr1 showed increased overall rates in heterogeneous stock (HS) mice, which combines genetic information from 8 strains, relative to B6xCAST hybrid mice (Paigen *et al.*, 2008). Interestingly however, patterns of CO distribution were conserved in both strains at a regional (Mb) level, despite usage of different sets of hot spots. This in turn indicated that the distribution of CO events was controlled both at local and region levels. Variation in hot spot usage/activity with strain background is demonstrated well at the *Psmb9* hot spot, discussed in section 1.4.2.5.

1.4.2.3. Sex-specific differences in CO activity

In mice, the female genetic map is 1.3 times longer than the male equivalent (Shifman *et al.*, 2006). Sex-specific differences in CO activity are also observed at the chromosome-level, with females showing higher overall rates (Paigen *et al.*, 2008; Billings *et al.*, 2010). It is likely that these differences are intimately linked with shorter interference distances (measured in megabases of DNA-sequence) in females than in males, allowing female chromosomes to undergo multiple COs (Petkov *et al.*, 2007). The two sexes also show differences in distributions of COs, with males showing elevated recombination in telomere-adjacent regions, whilst females showed a more uniform distribution, with regions alternating in high and low activity from centromere to telomere (Paigen *et al.*, 2008; Billings *et al.*, 2010). In addition, hot spot usage varies between the two sexes, with some hot spots active in one sex only (this is discussed further in section 1.4.2.5). Moreover, females seem to use more hot spots than males, with lower activity per hot spot (Billings *et al.*, 2010).

1.4.2.4. Hot spot locations and genomic features

Data from a genome-wide survey of recombination indicated that mouse hot spots tend to overlap with genes and identified significant correlations between hot spot locations and GC-content, density of short and long interspersed nuclear elements and long terminal repeats (Smagulova *et al.*, 2011). The same study also showed that sequences 5′ of hot spot centres are rich in purines, whilst those 3′ show a pyrimidine bias, a feature that was shown to apply to human hot spots as well.

1.4.2.5. Factors influencing CO activity at a hot spot- Lessons from *Psmb9*

Regulation of recombination activity at a mouse hot spot is perhaps best understood at the *Psmb9* hotspot, located in the mouse major histocompatibility complex (MHC) region. Overall recombination activity and sex-specificity at this hot spot is influenced both by local DNA sequences acting *in cis*, as well as by genetic factors located away from the hot spot acting *in trans*. These factors are briefly described below.

First, enhanced activity at the hot spot is associated with specific haplotypes, namely, the *wm7* MHC haplotype derived from *M.m. molossinus* or the *CAS3* MHC haplotype derived from *M.m. castaneus* (Shiroishi *et al.*, 1990). Second, the genetic locus responsible for the increased activity observed in these haplotypes falls outside of the hot spot and affects CO frequencies *in trans* (Shiroishi *et al.*, 1991; Baudat & de Massy, 2007). Additional studies showed that the locus necessary for hotspot activity in mice carrying the *wm7* haplotype lies at least 17.5 Mb away from the *Psmb9* hot spot, between 10.1 and 16.8 Mb on chromosome 17. This locus was referred to as *Dsbc1* for double strand break control 1 (Grey *et al.*, 2009). Third, overall activity is also influenced by a second haplotype-specific element within the *Psmb9* hot spot that represses initiation *in cis* (Baudat & de Massy, 2007).

Finally, the *Psmb9* hot spot shows sex-specific differences in activity. It is female-specific in mice carrying the *wm7* haplotype, but is active in both sexes in mice carrying the *CAS3* haplotype (Shiroishi *et al.*, 1990). Interestingly however, mice carrying a shorter *wm7* fragment show elevated recombination activity in both sexes. This observation led to the identification of a genetic locus distal to the hot spot, which is responsible for suppressing activity, specifically in males (Shiroishi *et al.*, 1991).

These studies have shown that overall activity at the *Psmb9* hot spot is the result of intricate interplay between various *cis*- and *trans*-acting factors and provided valuable insights into the delicate and complex mechanisms that shape the recombination profile in mammals.

1.4.2.6. Global regulation of mouse hot spots

In addition to regulating CO activity at the *Psmb9* hot spot, the *Dsbc1*^{wm7} allele also influences recombination on other chromosomes, including chr15 and chr18. CO frequencies at the *Hlx1* hot spot on chr1 are also affected (Grey *et al.*, 2009). At the same time as these findings were reported, Parvanov and colleagues (2009) reported that a locus *Rcr1*, which lies between 11.74 and 17.04 Mb on chromosome 17, regulated the activity of at least three hot spots on chr1 *in trans*, including the *Hlx1* hot spot.

Interestingly, *Rcr1* was found to overlap with *Dsbc1* and since both affected activity at the hot spot *Hlx1*, it seemed likely that both loci represented the same gene. Further, the two loci influence the frequency of COs, as well as NCOs, at hot spots *Hlx1* and *Esrrg-1*, indicating that they regulate the frequency of initiating DSBs rather than the CO/NCO decision (Parvanov *et al.*, 2009). Finally, the loci are similar in that their effects are specific, only affecting some hot spots, but not others (Parvanov *et al.*, 2009). In 2010, Parvanov *et al.* showed that *Rcr1/Dsbc1* represented the effects of the gene *Prdm9* (PR domain containing 9). The protein product of this gene has since been identified as a global regulator of hot spots in mammals; these studies will be described in detail in Chapter 6.

1.4.3. NCOs in mice

The large excess of DSBs, as estimated from RAD51/DMC1 foci, over COs predicts that ~90% of DSBs get resolved as NCOs. Detecting and analysing NCOs in mammals has proved a considerable challenge since these events have very localised signatures. Most of the current data on NCOs comes from molecular analysis at a few mouse hot spots. These studies suggest that NCOs co-localise with COs, supporting the idea that both outcomes result from initiating DSBs over the same region (Guillon & de Massy, 2002). Interestingly, a study by Guillon *et al.* (2005) reported that appearance of NCOs shows similar kinetics to that COs, appearing between mid and late-pachytene. This is contrary to data in yeast, which shows that NCOs are formed earlier than COs (Allers & Lichten, 2001).

Nonetheless, consistent with yeast, some evidence does point towards separate CO and NCO pathways in mice. For instance, conversion tracts associated with COs and NCOs show differences in length, with CO associated conversion tracts being longer (~500 bp in size) and NCO tracts being considerably shorter, ranging between 5- 300 bp (Guillon *et al.*, 2005). Also, mismatch repair proteins MLH1 and MLH3 have been shown to be specifically required for CO formation, with *Mlh1*$^{-/-}$ mice showing normal levels of NCOs (Guillon *et al.*, 2005). These and other features of NCOs, including distributions and relative ratios of COs and NCOs at recombination hot spots are discussed in more detail in Chapter 7.

In conclusion, studies in mice have highlighted that major players, pathways and processes of meiotic recombination are largely conserved from yeast to mammals. In addition these studies have shed light on previously unidentified aspects of meiotic recombination in mammals, including mechanisms governing the regulation of recombination activity at mammalian hot spots. More recently discoveries made in mice have also provided important clues regarding the regulation of meiotic recombination in humans. These studies will be discussed in more detail in various parts of this work.

1.5. Meiotic recombination in humans

Studying meiotic recombination in humans has proved rather challenging. In yeast, the ability to introduce genetic modifications and to analyse all four products of meiosis has greatly facilitated the understanding of meiosis and meiotic recombination. Mice too, are genetically malleable to a certain extent. Moreover, it is possible to conduct large-scale *controlled* genetic crosses in this organism in order to map numbers and distributions of recombination events. None of this is possible in humans and for a long time, the understanding of meiotic recombination was limited to information from low-resolution cytogenetic studies and from pedigree data based on polymorphic loci in a limited number of families (discussed below). Nonetheless, recent years have witnessed major advances in the elucidation of distributions and processes of meiotic recombination in humans, as resolution and coverage of studies on recombination have vastly improved. In this section, the different methods used to investigate recombination in humans will be described and information and insights from respective strategies will be discussed in detail.

1.5.1. Cytogenetic studies

Traditionally, positions of COs have been directly investigated by analysing the distribution of chiasmata in diakinesis-stage human meiocytes (Hultén, 1974; Laurie *et al.*, 1981; Laurie & Hultén, 1985). These studies have revealed that numbers of chiasmata in male meiosis vary between 49.6 and 53.7 (McDermott, 1973; Laurie & Hultén, 1985). Analysis of chiasmata has generally proved useful for investigating patterns of crossing-over on individual chromosomes and for estimating genome-wide frequencies of such events. Indeed, these studies have identified an interference distribution of chiasmata in male meiocytes, as well as a preference for CO-formation in telomere-proximal regions (Laurie & Hultén, 1985). Nonetheless, difficulties in obtaining cells in this stage of meiosis from fetal oocytes and spermatocytes together with the sub-optimal, highly condensed morphology of these chromosomes, have imposed serious limitations on information gleaned from this technique (Vallente, Cheng & Hassold, 2006).

Subsequently immunolocalisation studies investigating patterns of CO-associated proteins, most prominently MLH1, in pachytene-stage meiocytes have become increasingly popular. This is partially due to the success of such studies in providing accurate estimates of COs in mice (see section 1.4), on account of better resolution of these studies compared to ‘conventional’ cytogenetic approaches and considering the abundance of pachytene-stage meiocytes in fetal oocytes and testicular biopsies (Vallente, Cheng & Hassold, 2006). Thus, several studies have analysed numbers of MLH1 foci in human males as surrogates for CO-events. These studies indicate that there are ~50 MLH1 foci per spermatocyte, which translates to ~50 COs genome-wide and a genetic map length of 2,500 cM[1,2] (Barlow & Hultén, 1998; Lynn *et al.*, 2002; Gonsalves *et al.*, 2004; Tease & Hultén, 2004). These estimates are consistent with those obtained from analysis of distributions of chiasmata and from genetic linkage analysis (see below), thus validating this methodology for investigating CO distribution.

Studies on immunolocalisation of MLH1 foci in males have confirmed several similarities in CO distribution between mice and humans, including presence of interference and distal locations of COs (Anderson *et al.*, 1999). Moreover, they have showed that each chromosome arm, with the exception of short arms of acrocentric chromosomes, nearly always displays one or two MLH1 foci and rarely more (Vallente, Cheng & Hassold, 2006). Finally, numbers of MLH1 foci seem to linearly correlate with length of the SC, implicating a role for the SC in mediating levels of COs, at least in human males (Lynn *et al.*, 2002).

Relatively little immunolocalisation data is available for human females. This is mostly on account of difficulties obtaining the relevant samples, that is ovaries from female fetuses. The limited data available is intriguing on several accounts. First, several studies reported that numbers of MLH1 foci in females were similar to those observed in males (Lenzi *et*

---

[1] CentiMorgan (cM) is a measure of genetic distance, such that 1 cM is equivalent to a recombination frequency of 1%.

[2] 1 CO equals a genetic distance of 50 cM. Therefore, 50 COs correspond to a map length of 2,500 cM (Vallente, Cheng & Hassold, 2006).

al., 2005; Tease, Hartshorne & Hultén, 2006; Cheng *et al.*, 2009). This contradicts genetic linkage data, which suggests that recombination levels in females is ~1.6 times greater than that in males (Kong *et al.*, 2002; see below). The precise reason for this discrepancy is unclear, but several possible explanations have been suggested, such as selection favouring oocytes with higher levels of recombination for ovulation (Cheng *et al.*, 2009).

Furthermore, in contrast to human males and mouse males and females, where MLH1 foci appear at pachytene, these foci usually appear early in zygotene in females, sometimes even as early as leptotene (Lenzi *et al.*, 2005; Tease, Hartshorne & Hultén, 2006; Cheng *et al.*, 2009). This indicates a wider temporal window of CO-formation in females and in turn implies that not all COs may be detected at the same time. More importantly, it suggests a significant difference in the kinetics of recombination events between males and females.

Despite these contradictions, male and female distributions of MLH1 foci do show some fundamental similarities, including a distribution consistent with the phenomenon of interference (Cheng *et al.*, 2009). Moreover, variation in numbers of MLH1 foci and thus COs has been observed amongst both males and females (Lynn *et al.*, 2002; Sun *et al.*, 2006). Further, consistent with data in mice, MLH1 foci in human females tend to be interstitial. This is further confirmed by linkage data (Cheng *et al.*, 2009), as discussed below.

Aside from MLH1, immunolocalisation studies have been used to study several other recombination proteins, including RAD51, γH2AX, RPA, MSH4 and MSH5; synaptonemal complex proteins SCP1 and SCP3, as well as cohesins REC8 and STAG3 (Vallente, Cheng & Hassold, 2006). These studies have made it possible to directly analyse the chronology and kinetics of significant structures and complexes in meiosis and meiotic recombination in humans and highlighted important similarities in these processes between humans, mice and yeast. Broadly speaking, these studies are reinforcing the idea that humans follow the yeast paradigm when it comes to meiotic recombination (Vallente, Cheng & Hassold, 2006).

1.5.2. Insights from analysis of recombination in human pedigrees

A seminal paper by Botstein *et al.* (1980), which described an effective methodology for constructing genetic maps using inheritance patterns of DNA rather than protein polymorphisms, marked the beginning of the era of genome-wide linkage-based genetic maps. Characterisation of large pedigrees, such as the Centre d'Etude du Polymorphisme Humain (CEPH) reference families, the availability of good quality genome-wide sequence data, the identification of a large number of single nucleotide polymorphisms (SNPs) through projects such as HapMap[3] and finally the advent of high-throughput SNP-typing technologies, have all contributed towards the increasingly better resolution of such studies (Dausset *et al.*, 1990; International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001; The International HapMap Consortium, 2005, 2007). Indeed, linkage analyses in pedigrees have shed light on the sex-specific genome-wide distribution of COs and differences therein, enabled identification of Mb-scale ‘recombination jungles’ and kb-scale ‘recombination hot spots’ and genomic features associated with these, as well as revealing substantial inter-individual differences in levels of COs and aided the identification of genetic factors influencing this. These studies have marked crucial progress in the understanding of meiotic recombination in humans and are described in detail in the following sections.

1.5.2.1. General features of sex-specific linkage-based genetic maps

Many genome-wide genetic maps have been published over the last two decades (see Table 1.1). Estimates of male and female map lengths from these studies appear largely similar to one another and are consistent with female maps being ~1.6 times longer, indicating higher levels of crossing-over in this sex. It is worth noting however, that significant variations in recombination rates exist within each sex; this is discussed further in section 1.5.2.4. Also, males show more COs than females in telomere-proximal regions, whilst the reverse is true for centromeric regions (Broman *et al*., 1998; Yu *et al*., 2001; Kong *et al*., 2002).

---

[3] The International HapMap project was created for the purpose of developing a haplotype map of the human genome, in order to facilitate identification of genetic variants that confer susceptibility to common diseases. Three complete datasets (phases I, II and III) from the project have been released since 2005.

Table 1.1 Summary of genetic map lengths obtained from several recent studies

Study reference	Male genetic map length (cM)	Female genetic map length (cM)
Matise *et al.*, 1994	2,625	3,799
Broman *et al.*, 1998	2,730	4,435
Kong *et al.*, 2002	2,590	4,460
Matise *et al.*, 2003	2,642	4,414
Kong *et al.*, 2004	2,813	4,600
Jorgenson *et al.*, 2005	2,654	4,320

Beyond sex-specific differences, Kong *et al.* (2002) also showed variations in levels of recombination according to chromosome size, such that shorter chromosomes showed higher levels of crossing-over than longer chromosomes. For example, the numbers of COs on chr21 and chr22 are nearly double the levels on chr1 or chr2.

Finally, it is worth mentioning that linkage-based genetic maps for different populations are largely similar. Jorgenson *et al.* (2005) specifically addressed this in a study wherein genetic maps were compared between four population groups, namely, Europeans, African Americans, Mexican Americans and East Asians. The study showed good consistency amongst the maps, aside from a slight increase in length of the African American map that the authors suggest is in part owing to presence of genotyping errors. Other differences amongst the maps are mainly local and exclusively affect small parts of the genome.

1.5.2.2. Mb-scale recombination 'deserts' and 'jungles'

Until recently, resolution of linkage analysis from pedigrees was limited to Mb-levels. It is not possible to detect or study hot spots, which are typically 1-2 kb wide, at this resolution. Nonetheless, larger regional variations in CO activity could be well documented.

Thus, Yu *et al.* (2001), in their study based on 188 meioses in 8 CEPH reference families, showed that recombination rates can vary from 0 cM/Mb to 8.8 cM/Mb along individual chromosomes. Further, they identified regions of up to 6 Mb, which showed greater than average CO activity (> 3 cM/Mb), referred to as recombination 'jungles', as well as regions of up to 5 Mb showing less than average activity (< 0.3 cM/Mb), referred to as recombination 'deserts'. Data from Kong *et al.* (2002), based on 1,257 meioses in 869 individuals from 146 Icelandic families, was consistent with all 12 jungles reported by Yu *et al.* (2001). However, they could only identify 8 of the 19 deserts reported by the earlier study, a likely consequence of the small sample size of the Yu *et al.* (2001) study.

A more recent study by Chowdhury *et al.* (2009) based on 1,295 two-generation families with two or more children per family, from the Autism Genetic Research Exchange (AGRE) and Framingham Heart Study (FHS) collections, mapped genome-wide recombination jungles at higher resolutions. Thus, 308 and 155 maternal and paternal recombination jungles were reported in total and mapped to average intervals ranging between 1.5 - 3.7 Mb. Recombination jungles identified were preferentially located in the 5% most telomeric regions analysed. This was especially true for males, where 70% of jungles were present in such regions, as opposed to the 18% that resided in these regions in females.

1.5.2.3. Correlations between genomic features and recombination rates

Analysis of recombination in pedigrees has highlighted a number of genomic features associated with elevated CO activity. First, Yu *et al.* (2001) reported a correlation, albeit weak, between $(AC)_n$ tandem repeats and CO rates. The same study also reported a significant positive association between recombination frequencies and GC content, as previously shown by Eisenbarth *et al.* (2000). At higher resolutions however, it appeared that regions with the highest frequencies of COs contained high CpG fractions, but relatively low GC contents and poly(A)/ poly(T) fractions (Kong *et al.*, 2002).

More recent high-resolution (kb-level) linkage analyses in Hutterite and Icelandic pedigrees show that recombination rates are generally low in regions containing genes (Coop *et al.*, 2008; Kong *et al.*, 2010). In fact, recombination rates initially increase with distance from the nearest gene and are at their highest tens or hundreds of kb away from transcription start sites, before decreasing again (Coop *et al.*, 2008).

1.5.2.4. Inter-individual differences in recombination

Several pedigree-based studies have reported polymorphisms in recombination activity amongst females (Broman *et al.*, 1998; Kong *et al.*, 2002; Kong *et al.*, 2004). However, a corresponding level of variation was not reported for males until very recently (Broman *et al.*, 1998; Kong *et al.*, 2002). The first linkage-based study that showed significant variation in recombination levels amongst males was by Cheung *et al.* (2007) and was based on 38 CEPH pedigrees. Subsequently, Chowdhury *et al.* (2009) and Coop *et al.* (2008) confirmed highly significant levels of variation in recombination rates in both males and females in AGRE/FHS and Hutterite pedigrees. Additionally, Coop *et al.* (2008) reported significant variation in numbers of COs on individual chromosomes in males.

Genome-wide association analysis utilising numbers of COs as a quantitative trait showed that six different loci were associated with the observed variation in recombination activity in AGRE/FHS families (Chowdhury *et al.*, 2009). Two of these had been reported previously from studies in Icelandic pedigrees. The first is an inversion on chromosome 17q21.31, which is associated with increased recombination rates in females (Stefansson *et al.*, 2005). The second locus is *RNF212*; previous work by Kong *et al.* (2008) had shown significant associations between *RNF212* haplotypes and recombination rates in both males and females, but with opposite effects in the two sexes (discussed further in Chapter 6).

Chowdhury *et al.* (2009) showed that in addition to the inversion on 17q, recombination rates in females were also associated with variants on chr10, near *K1AA1462* and on chr1, near *PDZK1*. Together, these loci accounted for an estimated 6% of the total variation in recombination rates in AGRE and FHS females. Recombination in males, on the other hand, was associated with variants on chr9, near *UGCG* and on chr7, near *NUB1*. Expression profiles for mouse homologues of *Ugcg* and *Rnf212* are very similar and along with *Nub1*, they are expressed during meiotic prophase I, with peak levels at diplotene. Together, these loci accounted for 5.4% of the variation in recombination activity seen in males, leaving ample scope for additional factors to exercise a significant influence on overall recombination rates. Indeed, such a major regulator of recombination in humans was identified in 2010 and is described in detail in Chapter 6.

1.5.2.5. Recombination levels and reproductive success

Oocytes from a single mother have been known to show substantial variation in recombination rates (Kong *et al*., 2002). Several studies have investigated correlations between such variations and maternal age. Whilst earlier studies provided ambiguous results, a consensus now seems to be emerging with the advent of very high-resolution pedigree analysis.

Thus, a study by Kong *et al.* (2004), based on a substantially larger sample size of >14,000 maternal meioses, showed a significant positive correlation between maternal age and recombination frequencies of their offspring. The authors suggested that this was the likely result of selection favouring chances of an egg with greater number of recombination events to result in a live birth, in an attempt to protect against certain non-disjunction events, an increase in which has been associated with increase in maternal age (Hassold *et al*., 2000).

A subsequent study by Coop *et al.* (2008), based on Hutterite pedigrees of European descent, also confirmed a similar effect of maternal age, such that children born to mothers aged 35 or over, carried on average, additional 3.1 maternal recombination events than children born to mothers under 35 years of age. Kong *et al.* (2004) also showed that mothers whose oocytes show increased recombination frequencies tend to have more children. It is worth noting however, that no significant associations were reported between recombination levels and paternal age in any of these studies.

1.5.2.6. High-resolution mapping of COs in Hutterite pedigrees

Recent pedigree-based studies have taken advantage of the availability of high-density SNP-typing technologies, in order to detect and map COs at very high (kb-level) resolutions. Data from two such studies, Coop *et al.* (2008) and Kong *et al.* (2010) are briefly described here.

Coop *et al*. (2008) analysed genome-wide recombination profiles in 725 related Hutterites using the Affymetrix GeneChip Mapping 500K Array Set, which provided a SNP density nearly 100-fold higher than that of the Kong *et al.* (2002) study. Kong *et al.* (2010) on the other hand, used Illumina SNP arrays to study 15,257 meioses in Icelandic two-generation pedigrees at a resolution effectively down to 10 kb. At these resolutions, it is possible to study kb-scale human recombination hot spots, rather than Mb-scale jungles. Thus, Coop *et al.* (2008) reported that 60% of COs in Hutterite pedigrees, which could be resolved to sufficiently high resolutions, occurred within LD-based hot spots identified from Phase II HapMap data (2007) (discussed in section 1.5.3.3). Most of these hot spots were used by both sexes, but it appeared that males used a smaller proportion of hot spots compared with females. Kong *et al.* (2010) also showed that generally, males and females tend to use the same hot spots and only ~15% of hot spots observed seemed to be sex-specific. Data from this study are further discussed in Chapter 8.

Coop *et al.* (2008) noted significant differences in genome-wide hot spot usage within each sex. In other words, some individuals tended to use these ‘LD-hot spots’ more than others and this variation was heritable, pointing to genetic differences in at least one aspect of the recombination machinery. Later studies linked these variations in hot spot usage to the meiosis-specific protein PRDM9, as discussed in Chapter 6 (Baudat *et al.*, 2010).

In conclusion, it is worth emphasising that high-resolution linkage-analyses in large pedigrees overcome the limitations of cytogenetic and immunolocalisation studies with respect to sample sizes and resolution. Moreover, unlike studies based on LD (section 1.5.3) which provide a population and sex-averaged picture of recombination, linkage-based studies allow sex-specific, as well as, inter-individual aspects of recombination to be investigated at kb-level resolutions. Indeed, these studies have played a prominent role in unravelling genetic factors that regulate both locations and frequencies of meiotic COs, which was a major focus of my work (see Chapters 6 and 8).

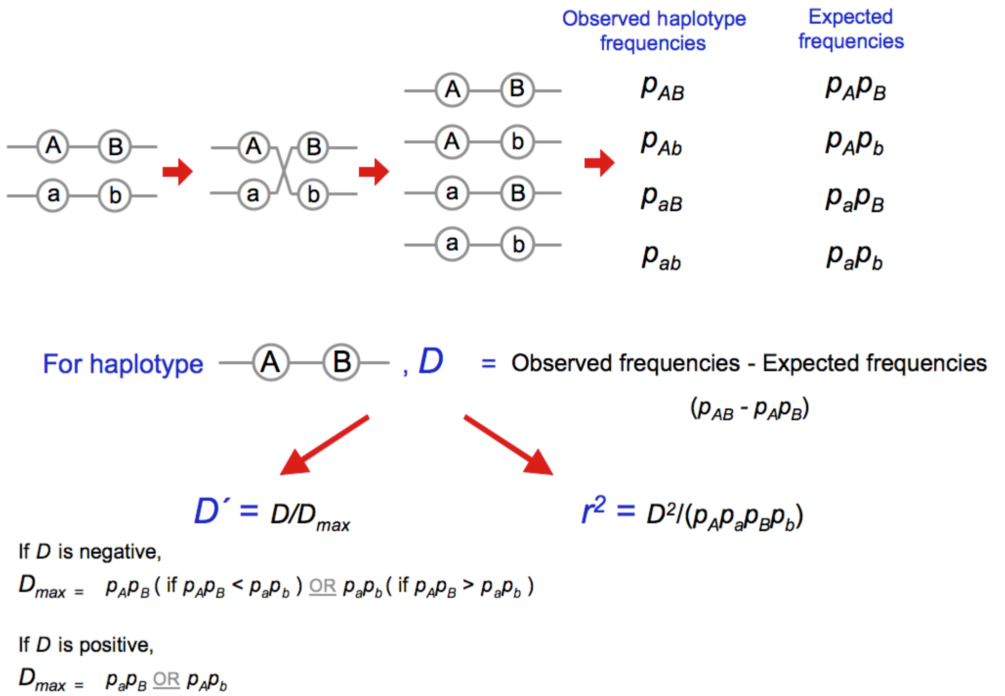
1.5.3. Insights from linkage disequilibrium studies

Linkage disequilibrium (LD) refers to the non-random association of alleles between loci. LD between a pair of markers decreases as the distance between them increases. The main force responsible for this breakdown of LD is recombination, such that regions displaying intense LD breakdown are likely to show very high recombination activity and regions showing uninterrupted and very strong LD are likely to be recombinationally inactive, although it is possible that these regions underwent recombination events which are not represented in contemporary populations. Albeit being an indirect measure of recombination, descriptions of genome-wide LD profiles greatly furthered an understanding of the fine-scale distribution of recombination events in the human genome, at a time when pedigree-studies were still limited in resolution to Mb-levels (eg: Broman *et al.*, 1998; Kong *et al.*, 2002). Also, recombination signals are more readily detectable from analysis of LD patterns since these provide information across several generations of the population. This section is dedicated to a description of studies investigating genome-wide recombination levels using LD and insights gained from these. Prior to this however, the most commonly used measures of LD will be described briefly.

1.5.3.1. Commonly-used measures of LD

The most widely used measures of pair-wise LD are $|D'|$ (Lewontin, 1988) and r^2 (Hill & Robertson, 1968), which essentially calculate the degree of association between a pair of bi-allelic markers as described in Figure 1.8.

Figure 1.8 Measuring LD (illustration reproduced from May, Slingsby & Jeffreys, 2007)


Observed haplotype frequencies
Expected frequencies
A B
a b
A B
a b
A B
A b
a B
a b
p_{AB}
p_{Ab}
p_{aB}
p_{ab}
$p_A p_B$
$p_A p_b$
$p_a p_B$
$p_a p_b$
For haplotype A B, D = Observed frequencies - Expected frequencies
$(p_{AB} - p_A p_B)$
$D' = D/D_{max}$
$r^2 = D^2/(p_A p_a p_B p_b)$
If D is negative,
$D_{max} = p_A p_B$ (if $p_A p_B < p_a p_b$) OR $p_a p_b$ (if $p_A p_B > p_a p_b$)
If D is positive,
$D_{max} = p_a p_B$ OR $p_A p_b$

The scenario considers alleles (A, a and B, b) at two nearby loci. In the presence of recombination, all four possible haplotypes for the two loci will be observed in a population. 'D', the simplest measure of LD, essentially represents the difference between observed frequencies of a particular haplotype and those expected from frequencies of individual alleles, as shown in the illustration. Another measure of LD that is independent of allele frequencies is D', which equals D/D_{max}. Usually, the absolute value of D', that is, $|D'|$ is used. At complete linkage equilibrium (LE), the observed frequencies of a haplotype are equal to the frequencies expected; thus, $|D'| = 0$; at complete LE. However, where only two or three of the four possible haplotypes are observed, the markers are in complete LD and $|D'| = 1$. An alternative measure of LD to $|D'|$ is r^2, which equals D^2 divided by the product of frequencies of alleles A, a, B, and b. $r^2 = 0$ indicates complete LE, whilst $r^2 = 1$ implies absolute LD, where only two of the four possible haplotypes exist in a population.

In contrast to $|D'|$, where values between 0 and 1, merely indicate incomplete or partial disruption of LD, intermediate values of r^2 are more informative due to a relationship between r^2 and the effective population size (N_e), distance between loci/ markers (d) and the rate of recombination per unit distance (R), such that $r^2 = 1/(1 + 4N_eRd)$ (Sved, 1971).

1.5.3.2. Inferring recombination levels from genome-wide analysis of LD

Over the last decade, many studies have used genome-wide patterns of LD to infer the fine-scale distribution of historical recombination events in human populations. It is important to note however that in addition to recombination, LD profiles are influenced by mutation, selection, genetic drift and the demography of populations (Stumpf & McVean, 2003). Detecting the signal of recombination against this backdrop has posed quite a challenge and a lot of emphasis has been placed on modelling the impact of such stochastic evolutionary processes on genetic variation, in order to tease out information on recombination.

In this regard, coalescent approaches, which are based on re-constructing the genealogy of sampled haplotypes by tracing them backwards in time to where they meet or 'coalesce', that is, to their most recent common ancestor, have proved very effective. Under the coalescent model, the amount of recombination needed to produce observed levels of LD is a function of ρ, the population recombination rate. ρ depends on the effective population size (N_e) and the per generation recombination rate r, the probability of a recombination event occurring during meiosis, such that $\rho = 4N_er$ (Stumpf & McVean, 2003). Various approaches have been used for calculations of ρ (reviewed in Stumpf & McVean, 2003).

One commonly used method is LDhat (McVean *et al.*, 2002). McVean *et al.* (2004) used LDhat to study patterns of recombination and validated the method using extensive simulation to test susceptibility of data to the effects of population bottlenecks, growth, as well as gene conversion. They also compared their estimates of recombination rates with previous data from pedigree and sperm analyses, both of which highlighted strong correlations between datasets. Moreover, recombination profiles obtained using LDhat were largely consistent between populations and were robust to variations in SNP spacing and ascertainment. LDhat has since been applied to infer genome-wide recombination rates for a variety of populations and datasets, including HapMap. Findings from these and other studies are briefly described below.

1.5.3.3. Distribution of LD hot spots

Subsequent to the validation of LDhat (McVean *et al.*, 2004), Myers *et al.* (2005) applied this method to study genome-wide patterns of recombination in a previously published dataset wherein ~1.6 million SNPs were genotyped in 71 Americans of either African, Asian or European ancestry (Hinds *et al.*, 2005). They identified ~25,000 hot spots in total and suggested that these features were ubiquitous and present every 50-100 kb in the human genome. LDhat was also subsequently applied to the HapMap datasets (The International HapMap Consortium, 2005, 2007). HapMap Phase II data, which was based on genotype information on over 3.1 million SNPs in 270 individuals from four different populations, showed evidence for presence of 32,996 human recombination hot spots (The International HapMap Consortium, 2007).

These studies showed that at high resolution, the recombination landscape in humans was largely dominated by narrow recombination hot spots, such that ~80% of detectable recombination typically occurred within 10-20% of the sequence analysed (Myers *et al.*, 2005; The International HapMap Consortium, 2005). Furthermore, whilst regions showing relatively low density and intensity of hot spots have been identified using these approaches, true Mb-scale recombination deserts, as reported by pedigree studies were not identified (Yu *et al.*, 2001; Kong *et al.*, 2002). Thus, Myers *et al.* (2005) reported that no regions greater than 200 kb, with the exception of centromeres as well as non-recombining parts of the Y chromosome, were completely devoid of recombination. Fluctuations in the density and intensity of recombination hot spots also seemed to contribute towards other large-scale variations in recombination rates, previously reported in pedigree studies, for instance, increased activity in telomeres compared to centromeres (Myers *et al.*, 2005).

Furthermore, data from Myers *et al.* (2005) showed that hot spots in humans tend to occur within ~50 kb of genes, but preferentially avoid transcribed regions, a finding further validated by high-resolution linkage analysis in Hutterite and Icelandic pedigrees (see section 1.5.2.3). There is however, evidence of asymmetry in the distribution of recombination 3′ and 5′ of transcribed regions, with the former being elevated compared with the latter (The International HapMap Consortium, 2007). Intriguingly, correlations also have been reported between gene functions and recombination, with genes generally associated with external/environmental interactions (*eg.* immune responses) and cell surface functions showing higher rates, compared with those involved in core cellular activities (*eg.* cell cycle regulation).

1.5.3.4. Sequence features associated with recombination hot spots

High-resolution population-based analyses of recombination have brought to light many sequence features associated with recombination hot spots. First, Myers *et al.* (2005) compared recombination hot spots with 'cold' regions of similar size and SNP density, in order to find sequence features specifically associated with the former. This revealed that long terminal repeats of THE1A and THE1B retrotransposons were both enriched within hot spots. Comparisons of sequences of these long terminal repeats within and outside of hot spots showed striking differences, the most prominent being a 6-fold over-representation of the sequence motif CCTCCCT within THE1A/B elements present in hot spot regions. The authors further suggested that the 7-mer motif resulted in a hot spot 60% of the time within the context of THE1A/B elements, but this was reduced to 2-3% outside the context of these repeats. A direct role for the motif in determining hot spot activity has been confirmed independently from sperm analysis at the *DNA2* hot spot (Jeffreys & Neumann, 2002) (see also Chapters 3 and 6).

Subsequently, a longer 13-mer motif CCTCCCTNNCCAC, which contained the previously identified CCTCCCT motif, was identified from analysis of HapMap Phase II data, wherein 22,699 autosomal and 608 X-chromosome hot spots were localised to within <5 kb of this sequence (Myers *et al.*, 2008). This motif was enriched amongst hot spots present within repetitive DNA, specifically in the context of THE1A/B and L2 elements, as well as within unique, single-copy DNA. Further testing for motifs carrying single mismatches to this 13-mer ‘core’ motif revealed degeneracy at positions 3, 6 and 12, such that motifs containing mismatches to the core motif at these positions were still associated with some, albeit reduced, hot spot activity (Myers *et al.*, 2008). The consensus motif was thus modified to CCNCCNTNNCCNC to reflect degeneracy at the aforementioned positions. Myers *et al.* (2008) showed that this motif was associated with 41% of 22,699 well-defined autosomal hot spots reported in HapMap. The motif was also associated with other recombination-associated processes and is further discussed in Chapter 6. It is worth noting that additional sequence features are also associated with hot spots, including enrichment of CT- and GA- rich repeats and depletion of $(TA)_n$ repeats, GC-rich repeats as well as L1s, with the effect appearing stronger the closer these LINEs (long interspersed nuclear elements) are to their full lengths (Myers *et al.*, 2005).

1.5.3.5. LD analysis as a tool for investigation hot spot evolution

Comparisons of fine-scale LD profiles indicate that hot spots are not conserved between humans and our closest evolutionary relatives, the chimpanzees, with whom we share ~99% sequence identity. Thus, estimates of recombination activity at six orthologous loci between western chimpanzees and humans showed that hot spot signals in humans were not conserved in chimpanzees (Winckler *et al.*, 2005). This is suggestive of rapid evolution of recombination hot spots over the last six million years, certainly at rates much faster than those of DNA sequence evolution. Interestingly however, further analysis of an extended 14 Mb-long region indicated that whilst hot spots were scarcely shared between the two species, recombination activity was conserved, albeit to very limited extents, at large (50 kb) scales (Ptak *et al.*, 2005). Ptak *et al.* (2005) proposed that this was a likely result of similar background recombination rates in the two species.

In summary, population-based approaches have proved an effective, albeit indirect, tool for investigating the fine-scale recombination landscape in humans. They have helped identify thousands of putative recombination hot spots and provided important insights into their distribution and evolution in human populations. They represent a happy medium between linkage analysis in human pedigrees, which have provided direct information on genome-wide patterns of recombination, but have, until recently been limited in resolution and sperm-based approaches which enable very detailed characterisation of hot spots at extremely high resolution, but are technically challenging and limited in coverage. These studies have highlighted important sequence motifs, especially the 13-mer CCNCCNTNNCCNC motif, which is associated with a substantial proportion of hot spots and has recently been implicated as the binding-site for a global *trans*-regulator of meiotic recombination (discussed in Chapter 6). Finally, in conjunction with sperm-based studies, population data provide insights into dynamics of hot spot evolution (see section 1.5.4.2).

Nonetheless, it is important to bear in mind that recombination rates estimated from population-based approaches are sex-averaged, although attempts have been made to extract sex-specific information from this data. These have mostly involved comparisons of recombination profiles on autosomes, which recombine in both sexes and the X-chromosome, which only recombines in female meiosis and analyses of regions that show sex-specific differences in activity from pedigree analysis (Myers *et al.*, 2005). Moreover, population studies do not provide insights into inter-individual variation in levels and distributions of recombination. Finally, these profiles represent inferred historical recombination events and thus, may not accurately reflect true contemporary recombination activity. The latter is best studied using sperm-typing assays, as described in the following section.

1.5.4. Insights from sperm-typing assays

A human sperm cell represents one of the four products of a single male meiosis. The large number of sperm that can be recovered from a single ejaculate ($\sim 10^8$) offers the scope for studying male meiotic recombination at extremely high resolutions, not only at the level of individual men, but at the level of constituent haplotypes in a man (Li *et al.*, 1988). Traditionally, sperm typing involved analysis of single sperm cells, but these studies were limited on account of the typically small numbers of sperm assayed per study. More recently, however, batch sperm typing, which allows screening of multiple pools of sperm DNA rather than single molecules, has been increasingly used. Both types of sperm typing approaches will be discussed presently, although the main emphasis will be on batch sperm typing, since this technique has been used to study meiotic recombination throughout the work described in the following Chapters.

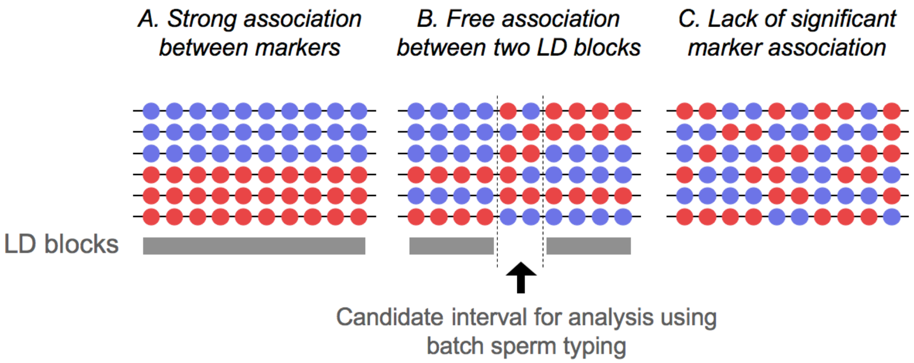
1.5.4.1. Single-sperm typing

Individual sperm cells for single-sperm typing can be isolated either manually or through florescence-activated cell sorting and then lysed to expose the DNA (Li *et al.*, 1988; Lien *et al.*, 1993; Cui *et al.*, 1989). In order to detect recombination events, whole genome amplification (WGA) is carried out on individual cells using degenerate primers, comprising random 15-nucleotide long oligos (Zhang *et al.*, 1992). Following this, loci of interest are PCR amplified and relevant markers typed in order to detect and map exchange intervals. This strategy was notably used by Cullen *et al.* (2002) to study recombination over a 3.3 Mb interval in the major histocompatibility (MHC) region in ~20,000 sperm cells from 12 men. 325 crossovers were mapped in total and 6 marker intervals showing hot spot activity identified. Single sperm typing has also been used for studying recombination across the human Xp/Yp pseudoautosomal regions (see section 1.6). Nonetheless, studies based on this technique have often been limited in terms of numbers of markers typed per sperm. Low numbers of sperm analysed per study have also imposed further restrictions in resolution. In fact, on account of these limitations, batch sperm typing (described below) has been used more extensively in the identification and characterisation of human recombination hot spots.

1.5.4.2. High-resolution batch sperm typing

An in-depth analysis of meiotic recombination, including definition and characterisation of hot spots at the sub-kb level, has become possible with the advent of batch sperm typing, wherein pools containing hundreds or even thousands of sperm molecules can be simultaneously screened for recombination events. Characterising hot spots via batch sperm typing is essentially a two-tiered process. First, a target interval is identified using patterns of LD as a guide (Figure 1.9). Regions showing high breakdown of marker association flanked on either side by LD blocks (wherein markers are in strong association with one another) are generally good targets for further investigation. However, a true picture of contemporary recombination is only obtained from the second stage of analysis, which involves batch sperm typing.

Figure 1.9 Possible LD profiles over a region of interest


A. Strong association between markers
B. Free association between two LD blocks
C. Lack of significant marker association
LD blocks
Candidate interval for analysis using batch sperm typing

The figure shows examples of LD profiles across a hypothetical genomic region. Blue and red circles represent polymorphic markers. Note that SNPs are the markers of choice for these studies on account of their abundance and low frequencies of recurrent mutation. 'A' represents a scenario wherein markers are in strong association with one another and constitute a continuous LD block, which has most likely been recombinationally inactive in the history of the population (Note: gradual breakdown of this LD over extended distances implies random distribution of recombination events). 'C' represents an example of a region wherein there is extensive breakdown of marker association; it is likely that this entire region has been very active recombinationally. 'B' on the other hand, shows a region wherein two LD blocks are separated by a region of sudden breakdown of marker association. This sudden and localised breakdown of LD hints at the presence of an interval wherein recombination events may cluster to form a hot spot. 'B' is thus a good target for sperm assays.

In a typical batch sperm-typing assay, subsequently referred to as a 'CO assay', recombinant molecules are selectively amplified from pools of sperm DNA and crossover breakpoints mapped by typing informative markers within the interval assayed. This strategy has been used for high-resolution analysis of meiotic CO breakpoints throughout this work and is described in further detail in several chapters of the thesis, notably Chapter 3. CO assays allow recombination to be investigated at resolutions of 0.0001 cM or less and whilst these studies have only covered a very small fraction of the human genome, they have provided unprecedented insights into the features and characteristics of recombination hot spots (Jeffreys *et al.,* 2004).

For instance, it is now well accepted that human CO hot spots are generally 1-2 kb wide. The first indication of this indeed came from CO assays at the MS32 minisatellite (Jeffreys *et al.,* 1998). Moreover, these studies have shown that in addition to a more or less constant width, CO hot spots also share symmetrical and quasi-normal distribution of CO breakpoints across the hot spot interval. A notable exception to this is skewed CO-distribution at the *NID1* hot spot, wherein exchange events seem to avoid a palindromic AT-rich minisatellite, thus creating a ‘cold spot’ (or a CO-refractory zone) within a hot spot (Jeffreys & Neumann, 2005). Interestingly, a recent study at the mouse *A3* hot spot on chr1 has shown abundant formation of NCOs within a CO-refractory zone, which contains an ~140 bp long imperfect inverted repeat, suggesting that heteroduplex containing repetitive DNA may preferentially be resolved as NCOs rather than COs (Cole, Keeney & Jasin, 2010). Other features of CO hot spots, including polymorphism in activity, the preponderance of ‘simple’ COs involving a single switch between haplotypes and co-localisation COs with NCOs, as well as factors influencing the frequency and distribution of recombination events within hot spots, will be discussed in other, more relevant parts of this thesis.

Presently however, it is worth highlighting that estimates of recombination activity at CO hot spots determined using the two most commonly used strategies, namely, population and sperm–based approaches, are not always consistent with one another. This has been well documented over a 206 kb region on chr1, wherein hot spots were identified within regions of strong marker association, including hot spot *NID3*, which was the most intense of the eight hot spots investigated in this study (Jeffreys *et al.*, 2005). Conversely, sperm analysis over a region within the *DPB1* gene did not reveal an active hot spot, despite presence of intense LD breakdown (Kauppi, Stumpf & Jeffreys, 2005). Given that population-based approaches infer historical activity, whilst sperm assays detect contemporary rates, it is likely that the former scenario indicates presence of young hot spots, which have failed to leave their mark on population diversity data, whilst the latter likely represents an extinct hot spot or one that is specifically active in females. Collectively, these discrepancies between historical and contemporary CO activity not only highlight the complexities of inferring recombination activity from population data, but also provide clues regarding likely time-scales of hot spot evolution. Indeed, along with evidence showing a lack of conservation of hot spots between humans and chimpanzees, these data support rapid turnover of recombination hot spots within the time-scales of existing human populations.

It is important to mention here that further evidence supporting the transient nature of hot spots comes from observations of ‘meiotic drive’ at recombination hot spots, a phenomenon by which hot spots effectively extinguish themselves by over-transmission of hot-spot disrupting alleles (described in detail in Chapters 3 and 6). This raises several questions as to how hot spots arise and persist in human populations, in the face of such hot-spot attenuating drive. Some very recent work, including data presented in this thesis, has provided important answers pertaining to mechanisms governing the appearance, maintenance and eventual extinction of hot spots from human populations and will be discussed at length in subsequent Chapters. Whilst sperm studies have described several important characteristics of hot spots and shed light on the underlying recombination processes in the male germline, they have largely been based on autosomes, with the notable exception of one hot spot.

1.6. Human X-Y pseudoautosomal regions

In contrast to autosomes and X chromosomes in females, recombination between X and Y-chromosomes in males is limited to two regions of homology, referred to as pseudoautosomal regions (PARs) (Cooke, Brown & Rappold, 1985, Simmler *et al.*, 1985; Freije *et al.*, 1992). These regions are unique in terms of their evolutionary histories and genomic contexts and play very significant roles in male meiosis (see 1.6.1). In fact, recent data from mice suggest that meiotic recombination in such regions is under distinct genetic control, further highlighting their importance in meiosis in males (Kauppi *et al.*, 2011; see section 1.6.4). An understanding of recombination in human PARs is thus not only interesting in its own right, but also important for a holistic understanding of mechanisms and processes underlying meiotic recombination in the human genome.

The two PARs lie at the ends of the short and long arms of X and Y chromosomes. PAR1, the larger of the two PARs, lies at the end of the short arms and is thus sometimes referred to as the Xp/Yp or 'major' PAR. It is estimated to be ~2.7 Mb-long, however, the preponderance of repetitive DNA has made it very difficult to determine the exact physical length of PAR1. In fact, large gaps still exist in the PAR1 sequence data, the combined size of which is estimated at 370 kb (Flaquer *et al*., 2008). PAR2, on the other hand, is much smaller at ~ 330 kb. It lies at the ends of long arms of X and Y and is sometimes referred to as the Xq/Yq or 'minor' PAR. Although both PARs represent regions of homology between X and Y chromosomes, they show important differences with regards to their evolution, sequence composition and roles in male meiosis, as discussed hereafter.

First, PAR1 and PAR2 have evolved separately. PAR1 has a longer evolutionary history and is homologous to PARs present in several other species, including Old World monkeys (Blaschke & Rappold, 2006). PAR2, on the other hand, is specific to humans and has been shown to have resulted from an L1-mediated ectopic recombination event subsequent to the human-chimpanzee divergence, thus leading to the translocation of an ~320 kb sequence from the X-telomere onto the Y chromosome (Kvaløy *et al*., 1994).

Second, PAR1 and PAR2 show differences with regards to sequence composition. PAR1 represents a region with a GC- content significantly higher than the rest of the X chromosome and increased density of repetitive DNA (Cooke, Brown & Rappold, 1985; Simmler *et al*., 1985; Rouyer *et al*., 1990). Also, in contrast to PAR1, which does not show highly significant differences in sequence composition along its length, PAR2 is clearly comprised to two distinct 'zones'. Zone 1, which relates to the proximal 295 kb of PAR2, shows an extremely low GC-content of 34.5%, while, zone 2, which comprises the distal 35 kb of PAR2, is marked by a sharp elevation in GC-content to >51% (Ciccodicola *et al*., 2000). Zones 1 and 2 also differ in LI and Alu contents, with these repeats extending over 67% of the total sequence in zone 1 and only over 29% in zone 2 (Ciccodicola *et al*., 2000).

Table 1.2 highlights some differences in sequence composition between PAR1 and PAR2, in context of the X-chromosome average.

Table 1.2 Comparisons of sequence compositions between PAR1 and PAR2

The table compares prevalence of various sequence features (as % of total sequence) in PAR1 and PAR2. Corresponding values are shown for the X chromosome (average) as well. Data taken from Blaschke & Rappold (2006).

Sequence Feature	% X-chromosome	% PAR1	% PAR2
GC-content	39.5	48.1	40.5
L1	28.8	6.9	37.5
Alu	4.9	20.3	3.9

The two PAR2 zones also differ in terms of patterns of gene expression. Thus, genes residing in zone 1 (namely, *VAMP7* and *SPRY3*) undergo X-inactivation, in contrast to all other pseudoautosomal genes which escape this process (D'Esposito *et al*., 1996, 1997; Ciccodicola *et al*., 2000). It has been suggested that differences in features and regulation of the two zones are correlated with their evolutionary histories, with zone 1 having translocated from an autosome to the X chromosome subsequent to development of marsupials, but prior to the development of mice and zone 2 having translocated much later in evolution through a multi-step process (Ciccodicola *et al*., 2000).

In addition to differences in evolutionary histories and genetic context, the two PARs show marked differences in terms of recombination. Exchange within PAR1 is essential in male meiosis and it constitutes a male-specific recombinationally 'hot' domain (Burgoyne, 1982; see below). In contrast, exchange within PAR2 does not appear to be required in male meiosis and as such relatively little is known about recombination in this human-specific X-Y homology region.

In fact, initial cytogenetic data suggested that whilst pairing occurred at the distal ends of X and Y, that is at PAR2, in ~50% of early pachytene spermatocytes, this did not persist and only associations at PAR1 were maintained by late pachytene (Chandley *et al*., 1984;

Armstrong, Kirkham & Hultén, 1994). Nonetheless persistent pairing at the distal ends of X and Y, even extending into metaphase I, resulting in the so-called ‘ring bivalent’ have been reported in infertile males and in at least one study formation of a short SC has been shown, raising the possibility of reciprocal genetic exchange in this region (Chandley *et al.,* 1987; Speed & Chandley, 1990). Further evidence for exchange within PAR2 was obtained from data on three-generation CEPH pedigrees (Freije *et al*, 1992; Li & Hamer, 1995). These and other studies leading up to the high-resolution analysis of recombination in this region are discussed in detail in Chapter 5. Presently however, the unique role of PAR1 in male meiosis and key, unusual features of recombination in this region are briefly described.

1.6.1. Recombination within PAR1 is essential in male meiosis

The possibility of genetic exchange within homologous regions of X and Y was first suggested by Kohler and Darlington (1934) from a study on male meiosis in rats. Since then, genetic exchange between X and Y-chromosomes has been validated using several different strategies, including cytogenetic and immunolocalisation studies, linkage analysis in pedigrees and sperm assays (*eg.* Pearson & Bobrow, 1970; Chandley *et al*., 1984; Cooke, Brown & Rappold, 1985; Simmler *et al*., 1985; Rouyer *et al*., 1986; Schmitt *et al*., 1994). Furthermore, it has become clear that there is at least one obligatory CO within PAR1 in male meiosis (Burgoyne, 1982; Rappold *et al*., 1994).

Crossing-over in PAR1 is crucial for X-Y pairing and synapsis. In the absence of X-Y pairing, germ-cell development largely arrests between metaphase-I and II (Mohandas *et al*., 1992). Pairing failure at PAR1 is thus associated with sterility or at least reduced fertility in human males (Chandley & Edmond, 1971; Gabriel-Robez *et al*., 1990; Hassold *et al*., 1991; Mohandas *et al*., 1992). Reduced recombination in PAR1 has also been linked with increased X-Y non-disjunction and production of aneuploid sperm (Jacobs *et al*., 1988; Hassold *et al*., 1991; Shi *et al*., 2001).

1.6.2. PAR1 constitutes a male-specific recombinationally ‘hot’ domain

It is estimated that, in males, recombination activity within PAR1 occurs at 10-20 times the genome average, whilst in females, wherein COs can take place along the entire lengths of the X chromosome, recombination within PAR1 occurs at levels scarcely above the genome average of ~1 cM/Mb (Petit, Levilliers & Weissenbach, 1988; Brown, 1988; Renwick, 1969; Weissenbach *et al*., 1987). PAR1 thus constitutes a male-specific recombinationally hot domain.

Genetic maps published for PAR1 (see Table 1.3), established using a variety of methods, are all consistent with highly elevated recombination activity in males. Interestingly however, these maps often show marked differences in PAR1 map lengths for males. In fact, the latest published male genetic map shows highly significant differences in overall length compared to the Schmitt *et al*. (1994) and Kong *et al*. (2004) maps (Flaquer *et al*., 2009). It is likely that the shorter lengths reported in these maps is a consequence of lack of markers within the first 750 kb from the telomere, which could result in missing COs occurring in this region (Flaquer *et al*., 2008). Shorter map length for the HapMap (2005) data, on the other hand, could be a result of the complexities inherent in estimating sex-specific information from sex-averaged population data. Minor fluctuations in other maps may well be a consequence of limited numbers of markers, data errors or analysis of insufficient numbers of meioses.

Fluctuations in map lengths occur to a lesser extent in female PAR1 genetic maps, all of which show substantially reduced lengths (on average ~12-fold) compared to the male maps. It is worth noting however, that the Flaquer *et al.* (2009) map, based a large number of markers, reports the most modest difference in male and female rates. Moreover, it reports equal rates of recombination in males and females for the first 100 kb from the telomere. Further studies are warranted to clarify these issues and to reach more robust estimates for genetic map lengths in the region.

Table 1.3 Estimates of PAR1 genetic map lengths (adapted from Flaquer *et al*. (2008)

Estimates of PAR1 genetic map lengths (cM) in males and females, estimated using 3-generation pedigrees, single-sperm typing and unrelated individuals are shown. Sex-specific map lengths from sex-averaged HapMap data were calculated assuming a male to female map ratio of 10:1. Also estimates shown are based on the Phase I dataset (The International HapMap Consortium, 2005) since PARs have not been analysed in the latest (Phase III dataset). Also, Schmitt *et al.* (1994) used 'sex' as a marker for estimation of map lengths; this is not included in the number of markers shown in table for this study.

Reference	Samples used	No. of markers	Male (cM)	Female (cM)
Rouyer *et al*., 1986	8 CEPH families	3	50	4
Page *et al*., 1987	44 CEPH families	5	49.9	4.18
Henke *et al*., 1993	38 CEPH families	11	49	4
Schmitt *et al*., 1994	2 men, 903 sperm	4	38	-
Lien *et al*., 2000	4 men, 1912 sperm	9	55.3	-
Kong *et al*., 2004	40 CEPH & 146 deCODE	6	11.7	0.8
HapMap, 2005	269 unrelated individuals	1400	38	3.8
Flaquer *et al*., 2009	28 CEPH families	23	55	6

Whilst initial studies suggested a uniform distribution of recombination events in PAR1 (Petit, Levilliers & Weissenbach, 1988; Brown, 1988), more recent studies seem to suggest otherwise. Thus, Lien *et al.* (2000) showed some evidence of heterogeneity in CO activity across PAR1 albeit at relatively low resolution. Further, more compelling evidence for a non-random distribution of recombination in PAR1 and clustering of COs into a hot spot was provided by May *et al.* (2002). These studies are discussed in further detail in Chapter 3.

1.6.3. Double-COs in PAR1 and implications on interference mechanisms

Burgoyne (1982) proposed that X and Y-chromosomes exchange genetic material through a 'single obligatory' CO within the Xp/Yp PAR (or PAR1) and that the small size of the pairing region should prevent more than one CO. Earlier genetic maps, including that of

Rouyer *et al.* (1986), based on analysis of 3 RFLP (restriction fragment length polymorphism) loci in 8 CEPH pedigrees and that of Page *et al*. (1987), based on studying the inheritance of 3 markers, including sex, in 44 CEPH families, did not detect any double-COs, consistent with a single obligatory exchange in PAR1 during male meiosis.

Since then several studies have reported double-COs in PAR1 during male meiosis. Such an event was first detected by Henke *et al.* (1993) and further characterised by Rappold *et al.* (1994), who analysed recombination using 11 markers in 330 CEPH offspring and detected 127 single COs and 1 double-CO, with maximum and minimum distances between the two events corresponding to 780- and 230-kb respectively. Flaquer *et al.* (2009) also reported a double-CO in their analysis, which is likely to be a replication of the aforementioned event reported by Henke *et al.* (1993) and Rappold *et al*. (1994).

Double-COs have also been reported in PAR1 from sperm analysis. Indeed, Schmitt *et al*. (1994) reported 3 double-COs from 555 sperm, whilst Lien *et al.* (2000) reported 21 such events from 1,912 sperm. The presence of more than one CO in PAR1, which is only ~2.7 Mb long, when more than 2-3 events are seldom detected on 50-250 Mb long autosomes (see section 1.5.1), raises important questions regarding mechanisms of interference in PAR1. This is discussed further in Chapter 4, section 4.3.2.

1.6.4. Distinct temporal and structural regulation of recombination in the mouse PAR

It is tempting to speculate that pairing and genetic exchange within the rather limited region of homology between the X and Y is more challenging than that on autosomes, wherein multiple interactions along chromosomes stabilise pairing interactions (Weiner & Kleckner, 1994). Nonetheless, X and Y manage to pair and recombine in the large majority of male meioses, suggestive of presence of distinct mechanisms that exist to ensure pairing within this small region of homology. Indeed, a very recent study by Kauppi *et al*. (2011) showed direct evidence supporting such mechanisms.

First, chromatin structure in the mouse PAR is modified, such that the DNA is organised into smaller loops tethered to a disproportionately long axis; this modification presumably allows increased scope for DSB-formation[4]. Second, DSB-formation and pairing in the mouse PAR seem to be under distinct temporal controls. Thus, Kauppi *et al.* (2011) found that pairing between X and Y rarely occurred before pachytene and was delayed relative to pairing on autosomes, which was accomplished by mid to late-zygotene. Further, it appeared that delayed X-Y pairing was a consequence of late formation of meiotic DSBs in the mouse PAR. Thus, PAR RAD51/DMC1 foci, which mark sites of DSB-formation, only started appearing during late-zygotene, when genome-wide numbers of these foci were declining.

Further, *Spo11*α and *Spo11*β are the two major mRNA splicing isoforms present in humans and in mice; expression of *Spo11*β is concomitant with global formation of DSBs, whereas *Spo11*α is expressed later (Romanienko & Camerini-Otero, 1999, 2000; Keeney *et al.*, 1999; Bellani *et al.*, 2010). Kauppi *et al.* (2011) showed that delayed PAR1 DSB-formation during late-zygotene requires SPO11α, as opposed to SPO11β, which is required for DSB-formation on autosomes and for the formation of a small fraction of PAR DSBs that form earlier (during leptotene). Thus, SPO11β fully sustains genetic exchange, pairing, synapsis, meiotic progression and chromosome segregation in *Spo11*β-only females, wherein exchange within PAR is not essential. In *Spo11*β-only males, on the other hand, X and Y-chromosomes fail to synapse in ~70% of the spermatocytes. Kauppi *et al.* (2011) suggest that failure of unsynapsed X and Y-chromosomes to orient themselves correctly on the metaphase I spindle triggers spindle-checkpoint induced apoptosis, evidenced by presence of apoptotic metaphase I cells in testis sections. Moreover, whilst some *Spo11*β-only males produce offspring, most are infertile as a consequence of absence of later-forming PAR DSBs.

---

[4] It has been previously suggested that DNA organised into smaller chromatin loops has greater propensity for DSB-formation (Kleckner, Storlazzi & Zickler, 2003) and since loop density per μM of the chromosome axis remains constant (Revenkova, 2004), loop size becomes inversely related to axis length (Zickler & Kleckner, 1999).

Since *Spo11* isoforms are conserved between mice and humans, these findings have important implications on the genetic and temporal control of meiotic exchange between X and Y chromosomes in human males and imply that variations in splicing patterns of *Spo11* may have vital bearing on susceptibility to X-Y non-disjunction, as seen in Turner and Klinefelter's syndromes.

1.7. This work

High-resolution analyses of recombination in humans have so far been largely based on autosomes and as such, little is known about the fine-scale distribution of recombination events in the human PARs or the factors influencing these. The broad aims of this work were thus to gain a better understanding of the processing and regulation of meiotic recombination in the human PARs, using a combination of linkage disequilibrium analysis and high-resolution batch sperm typing. More specific aims of each study carried out as part of this thesis are described in detail in individual Chapters.

Chapter 2. Materials and Methods

2.1. Materials

2.1.1. Chemical reagents

2M trisma base, 2-mercaptoethanol, ammonium sulphate, dithiothreitol, glycerol, sodium chloride, sodium lauryl sulphate, spermidine trichloride, tetramethyl-ammonium chloride and water were obtained from Sigma-Aldrich (Poole, U.K.). Bovine serum albumin (BSA) was obtained from Ambion (Huntington, U.K.). Deoxy-ribonucleoside triphosphate(s) (dNTPs) were obtained from Promega (Hampshire, U.K.). Ethidium bromide was obtained from BioRad (Hemel Hempstead, U.K.). Ethylenediamine tetra-acetic acid (EDTA), magnesium chloride, polyvinylpyrrolidone (PVP) and sodium hydroxide were obtained from Fisher Scientific (Loughborough, U.K.). Ficoll 400 was obtained from GE healthcare (Little Chalfont, U.K.). Details on final concentrations of reagents in various solutions are listed in the next section.

2.1.2. Plastic-ware

Most plastic-ware (including PCR tubes, 96-well plates, self-adhesive seals) were obtained from ABgene (Surrey, U.K.). PCR master mixes were set up in Protein LoBind Tubes obtained from Eppendorf Scientific (Hamburg, Germany). Plastic pipette tips were obtained from Starlab Ltd. (Milton Keynes, U.K.) and Sarstedt (Germany).

2.1.3. Oligonucleotides

Oligonucleotides were mostly supplied by Sigma-Aldrich (Poole, U.K.). Others were ordered from Protein and Nucleic Acid Laboratory (University of Leicester).

2.1.4. Enzymes

Taq polymerase was obtained from Kapa Biosystems (distributed by Anachem, U.K.) and *Pfu* polymerase from Stratagene/Agilent Technologies. Exonuclease I and T4 polynucleotide kinase were supplied by New England Biolabs (Hitchen, U.K.) and shrimp alkaline phosphatase (SAP) was supplied by Roche Applied Science (Sussex, U.K.).

2.1.5. Molecular weight markers

λ DNA digested with *Hind*III and ϕx174 digested with *Hae*III were supplied by ABgene (Epsom, U.K.).

2.1.6. Standard solutions

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

2.1.7. Human DNA

Sperm DNA was prepared from semen samples donated by anonymised donors. Samples subsequently referred to as 'European' were collected by Jane Blower (Assisted Conception Unit, Leicester Royal Infirmary), from U.K. donors of North European descent. Additional samples were also collected from members of the Department of Genetics, University of Leicester. All of these samples were obtained with informed consent and with approval from the Leicestershire Health Authority Research Ethics Committee. Samples subsequently referred to as 'African' samples were donated by A.D. Nakomo and S.B. Kanoyangwa (Forensic Science Laboratory, Causeway, Zimbabwe), with approval from the local ethical committee. All donor samples were re-named donor 1 (d1) through to d291 to ensure complete anonymity and organised into panels depending on the population origin of the samples.

Semen samples, as well as sperm DNAs extracted from these, were stored at -80 ºC. Sperm DNAs were quantified using a Nanodrop ND1000 spectrophotometer (Thermo Scientific). On account of limited availability, these samples were whole genome amplified prior to use in routine genotyping and optimisation experiments. This was done using the multiple displacement amplification (MDA) method developed by Deans *et al.* (2002). All MDA reactions were carried out using the GenomiPhi^{TM} DNA Amplification Kit (GE Healthcare), as per the manufacturer's instructions. As such, use of genomic DNA was limited to sperm crossover/conversion assays.

2.1.8. Computers

This thesis was produced using Macintosh OS X (version 10.4.11). Data was analysed and compiled using software packages Microsoft Word, Microsoft Excel, Microsoft Powerpoint and Cricket Graph III (version 1.5.2). Programs Factura and Autoassembler were used for DNA sequence trace analysis. Programs used for analysis of crossover hot spot centre, width and peak activity were written by Alec J. Jeffreys in True Basic 4.0.

Patterns of linkage disequilibrium (LD) were analysed using LD unit (LDU) mapping. These maps allow LD to be represented as a metric map, with additive distances. LDU maps are based on the Malécot model, which describes the decay of LD with distance, governed by levels of recombination (Malécot, 1948). 1 LDU corresponds to 1 swept radius, that is, the extent of LD useful for association mapping (Collins & Morton, 1998). LDU maps generally comprise a series of plateaus and steps, corresponding to LD blocks and regions of breakdown of marker association respectively. LDU maps used throughout this work were generated using an online program, available at http://143.210.151.197/cgi-bin/LDMAP/ldumapping.pl.

2.2. Methods

2.2.1. PCR amplification of DNA

DNA was amplified in 10μL reaction volumes, unless otherwise stated, using polymerase chain reaction (PCR). Reactions were set up in 200μL tubes or 96-well high-profile non-skirted plates with self-adhesive film. Suppliers of these and all other plastic-ware used are listed in Appendix 2.1. PCRs were either run on an MJ Research PTC-225 Tetrad DNA Engine supplied by GRI, Braintree, Essex, U.K. or on an Applied Biosystems Veriti® Thermal Cycler. All PCR reactions were set up in a category II laminar flow hood to minimise contamination. All plastic-ware, as well as reagents and pipettes used in PCRs were used exclusively in this 'PCR-clean' laminar flow hood.

Each 10μL PCR reaction typically contained 6.7mM 2-mercaptoethanol, 11mM ammonium sulphate, 113μg/mL bovine serum albumin (BSA), 1mM dATP, 1mM dCTP, 1mM dGTP, 1mM dTTP, 4.4μM ethylenediaminetetraacetic acid disodium salt (EDTA-Na) pH 8.0, 4.5mM magnesium chloride, 4.5mM Tris/HCl pH 8.8, 5-10ng template DNA, 12.5mM Trisma base, 0.2μM forward and reverse primers, 0.03U/μL *Taq* polymerase and 0.0015U/μL *Pfu* polymerase per reaction. Trisma base was added to PCRs as it raises the pH of PCR reactions, thus enabling denaturation at lower temperatures and reducing the risk of template depurination that occurs at higher temperatures. *Pfu* polymerase, on the other hand, is a proofreading 3′ to 5′ exonuclease, which removes base mismatches that potentially cause *Taq* polymerase to stall during amplification, resulting in shorter products or PCR jumping.

In PCR reactions where less than 5ng of template DNA was used for amplification, ~7.8μg/mL salmon sperm DNA was added per reaction as a carrier. The latter was supplied by Sigma-Aldrich, Poole, U.K. High-molecular weight carriers such as salmon sperm DNA coat the inside of plastic-ware used in PCRs and prevent target DNA sequestration.

Cycling conditions varied for different PCRs, depending on the DNA sequence context, amplicon length and primers used. However, generally, all PCRs started with an initial denaturing step (96°C for 1 min), followed by 28-36 cycles at 96°C for 20 sec (denaturing), 52-68°C for 30 sec (annealing) and 56-68°C for several minutes (extension). Extension times were set such that at least one minute was allowed per kilobase of DNA to be amplified. A final extension step was used in some PCRs, but not always. PCRs were typically held at 12°C after the run finished.

2.2.2. Agarose gel electrophoresis

Lonza SeaKem LE agarose was used to make gels of concentrations ranging between 0.8-2% (w/v). Higher concentration gels were typically used for smaller PCR amplicons. Sizes of gels varied between 5-20cm in length, depending on number of samples, amplicon sizes and degree of resolution required to separate fragments. Gels were run in horizontal submarine format in 0.5x TBE buffer (44.5 mM Tris-borate pH 8.3, 1 mM EDTA). 0.5μg/ml ethidium bromide was used in the electrophoresis buffer, to aid visualisation of PCR products. A quarter volume of loading dye (30% glycerol, 0.5x TBE and bromophenol blue to colour) was added to PCR products before gel loading. The amount of product loaded on a gel varied, typically ranging between 1-10μL. Samples were run alongside molecular weight markers, namely λ DNA digested with *Hind*III and ϕx174 digested with *Hae*III.

Electrophoresis tanks were made in house and power packs supplied by BioRad. PCR products were visualised using either a UV wand (Chromato-vue UVM-57, UVP Life Sciences) or a UV transilluminator (Syngene). Gel photographs were obtained, analysed and presented using Gene Snap and Gene Genius analysis systems (Syngene). Photographs were obtained using a Sony digital graphic printer UP-D895.

2.2.3. DNA re-sequencing

2.2.3.1. PCR product purification

PCR products were treated with Exonuclease I (ExoI) and shrimp alkaline phosphatase (SAP). ExoI and SAP degrade excess primers and dNTPs respectively. 10U of ExoI and 1.5U of SAP were added to 5μL of PCR products and incubated at 37°C for 1 hour, 80°C for 15 minutes and finally at 4°C for 2 minutes. Once purified, an accurate estimation of PCR product concentration was obtained via electrophoresis on an agarose gel, alongside size markers of known concentration.

2.2.3.2. DNA sequencing reactions

Automated DNA sequencing was carried out using the ABI PRISM Big Dye^{TM} Terminator Cycle Sequencing Ready Reaction Kit (version 3.1). Each sequencing reaction (10μL volumes) contained 1μL Big Dye^{TM} terminator ready reaction mix, 1.5μL 5x Big Dye^{TM} terminator buffer, 20-30 ng/kb purified PCR product template and 3.2pM sequencing primer. Sequencing reactions were cycled at 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes, between 25-30 times in an MJ Research PTC-225 Tetrad DNA Engine.

2.2.3.3. SDS/ heat treatment

10μL water and 0.2% SDS were then added to sequencing reactions. These were then mixed thoroughly in order to disperse any dye blobs. Reactions were then incubated at 98°C for 5 minutes and then at 25°C for 10 minutes in an MJ Research PTC-225 Tetrad DNA Engine.

2.2.3.4. Final purification using gel filtration

Sequencing reactions were purified using Performa spin columns (Edge Bio). These columns use a hydrated gel matrix to separate low molecular weight materials from DNA fragments greater than 20 bases. The eluate obtained as per manufacturer's instructions was submitted to the Protein and Nucleic Acid Chemistry Laboratory, University of Leicester for sequencing on an Applied Biosystems 3730 sequencer.

2.2.3.5. DNA sequence analysis

DNA sequence traces were edited and assembled using Factura and Autoassembler software respectively on an Apple Macintosh computer. Limits on sequence traces were identified by eye and were typically between 300-800 base pairs in length. Potential polymorphisms were highlighted by Autoassembler and also checked by eye from sequence traces.

2.2.4. SNP typing by allele specific nucleotide (ASO) hybridisation to dotblots

2.2.4.1. Dot blotting

PCR amplification was carried out such that at least 3ng (ideally >100ng) DNA per kilobase of PCR product was available for each dot. A quarter of a volume of loading dye (30% glycerol, 0.5x TBE, bromophenol blue to colour) and ~5x volume of denaturing solution (0.5M sodium hydroxide, 2M sodium chloride, 25mM EDTA) was added to each PCR product. A 96-well dot blot manifold was assembled with one sheet of Hybond –NX nylon transfer membrane (supplied by GE Healthcare Life Sciences), on top of two sheets of 3MM Whatman chromatography paper. All membranes were soaked in water prior to assembling on the manifold. A vacuum was then applied to the dot blot manifold and at least 30μL of denatured PCR products were applied to each filter. All wells were then washed with 200μL 2x SSC to neutralise the DNA. Filters were dried at 80°C for 10 minutes. Finally, the DNA was covalently linked to the membrane by exposure to UV light (7×10^4 Jcm^{-2}) in an RPN 2500 UV cross-linker (supplied by GE Healthcare).

2.2.4.2. ASO hybridisation to dot blots

Allele specific oligonucleotides (ASOs) were typically 18-nucleotide long and designed such that the SNP site lay at position eight from the 5′ end of the oligonucleotide. In some cases, the SNP site was shifted to a different position (usually position eleven) in order to boost specificity and/or efficiency of hybridisation. 9.6ng of each ASO was labelled in a reaction containing 1μL 10x kinase mix (700mM Tris-HCl pH 7.5, 100mM magnesium chloride, 50mM spermidine trichloride, 20mM dithiothreitol), 7.8μL water, 0.40 μL 10U/μL T4 polynucleotide kinase and 0.20μL 10mCi/mL γ-^{32}P-ATP for 1-3 hours at 37°C. Following incubation, 20μL of kinase stop solution (25mM di-sodium EDTA, 0.1% SDS, 10μM ATP) was added to the labelling reaction.

Dot blot filters were first soaked in 3x SSC and then pre-hybridised in 2.5mL TMAC hybridisation solution (3M tetramethylammonium chloride or TMAC, 0.6% SDS, 1mM EDTA pH 8.0, 10mM sodium phosphate pH 6.8, 0.1% Ficoll 400, 0.1% polyvinylpyrrolidone, 0.1% BSA and 0.4μg/mL yeast RNA) at 56°C in a Hybaid Shake 'n' Stack hybridisation oven. TMAC stops the preferential melting of AT base pairs versus GC base pairs and allows hybridisation with ASOs of varying GC-contents at the same temperature. The hybridisation solution was discarded after 10 minutes and replaced with 3mL fresh hybridisation solution, as well as 80μL unlabelled competitor ASO (at 8μgmL^{-1}) and incubated further for 10 minutes at 56°C. Following this, 30μL of α-^{32}P- labelled ASO and stop solution mix was added. Incubation was then continued at 50°C for 45 minutes-2 hours.

The hybridisation solution was then discarded and filters were washed in 2-3mL of TMAC wash solution (3M TMAC, 0.6% SDS, 1mM EDTA pH 8.0, 10mM sodium phosphate pH 6.8). Typically, filters were washed three times at 50°C for 5 minutes, followed by a final wash at 53°C for 15 minutes. Fresh TMAC wash solution was used for each wash. Membranes were then rinsed thoroughly in 3 x SSC at room temperature, wrapped in Saran wrap and exposed to an X-ray film (Fuji RX100) at -80°C, with an intensifier screen. Duration of exposure varied with intensity of signal, ranging between 2- 48 hours.

2.2.4.3. Removing probe and re-using membranes

Membranes were washed in 4-6 changes of 200mL boiling 0.1% SDS. A Geiger counter was used to monitor probe removal and to ensure that counts were below 5cps after washing. Membranes were then rinsed extensively in water at room temperature, followed by 3x SSC. Membranes were finally stored damp in Saran Wrap at 4°C.

Chapter 3. Characterisation of a second recombination hot spot near the Xp/Yp pseudoautosomal *SHOX* gene

3.1. Introduction

The Xp/Yp pseudoautosomal region (or PAR1) plays an essential role in male meiosis and represents a male-specific recombinationally 'hot' domain, with activity ~10-20 times greater than the genome average (Burgoyne, 1982; Petit, Levilliers & Weissenbach, 1988; Brown, 1988). Preliminary indication of uneven distribution of recombination events along PAR1 came from single-sperm typing experiments (Lien *et al*., 2000). Soon afterwards, a highly active PAR1 sperm CO hot spot, located ~515 kb from the Xp/Yp telomere and in many ways analogous to its autosomal counterparts, was identified using high-resolution LD analysis and batch sperm typing (May *et al*., 2002). These studies have marked important advances in understanding the features and distributions of recombination events in this important and unusual genomic region and are described briefly below.

3.1.1. Heterogeneity in recombination activity across PAR1

Just over a decade ago, Lien *et al*. (2000) used single-sperm typing to investigate variation in recombination activity within PAR1 in 1,912 sperm from 4 men of similar age and ethnicity. They showed substantial variation in CO activity across PAR1 in all men analysed, with two regions in particular showing highly elevated rates. The first included the interval between markers DXYS201 and GGAT2F08, which showed the highest rates in PAR1, at 26-38 times the genome average. These markers are located ~816 kb-apart, extending from 490 to 1307 kb (NCBI36/hg18). Elevated recombination per unit length of DNA was also noted close to the pseudoautosomal boundary (PAB1). This was consistent with previous data from Schmitt *et al*. (1994), who reported a 31-fold increase in CO activity in a 25 kb-long interval extending between marker DXYS77 and the boundary. It is worth mentioning that such high rates close to boundary suggest that CO frequencies immediately proximal to the boundary must be sufficiently low, in order to maintain the significantly different X- and Y-specific regions (Ellis *et al.*, 1989).

It is important to note that whilst the resolution of the Lien *et al*. (2000) study was sufficient to detect heterogeneity in the distribution of recombination events across PAR1, it did not provide any indications of whether COs clustered into 1-2 kb wide 'classical' hot spots. Indeed, the first direct evidence for clustering of COs in PAR1 into highly localised, intense hot spots came from high-resolution LD and sperm analysis in the *SHOX* region, as described below (May *et al.,* 2002).

3.1.2. Identification of the first pseudoautosomal hot spot

At the time of the May *et al*. (2002) study, published sequence data were not available for most of PAR1; the *SHOX* region was however an exception in this regard. This was on account of the medical significance of the *SHOX* gene, which had been implicated in idiopathic short stature as well as in Turner, Leri-Weill and Langer syndromes (Blaschke & Rappold, 2006). May *et al*. (2002) thus chose the *SHOX* region for study because of the availability of sequence data for this region and considering that this interval had previously been reported to show elevated recombination activity by the Lien *et al*. (2000) study.

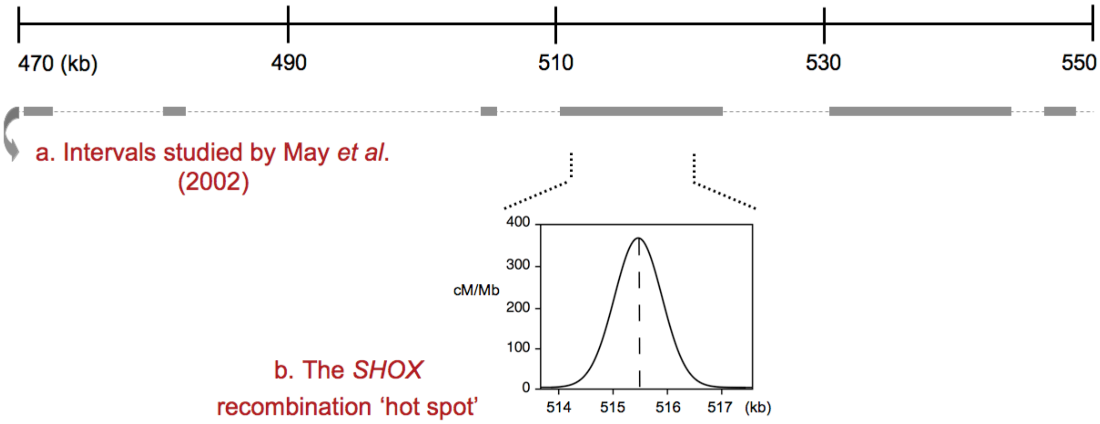
In total, May *et al.* (2002) studied six regions in and around the *SHOX* gene, collectively encompassing an interval between 470271 – 548368 bp (NCBI36/hg18; as shown in figure 3.1a). Analysis of patterns of marker association in 99 semen donors of North European descent revealed extreme break-down of LD with the largest LD block extending over just 3 kb. This presented a stark contrast with autosomes, wherein blocks of LD extend over tens of kilobases or even more (Jeffreys, Kauppi & Neumann, 2001).

In order to exclude the possibility that these observed patterns of LD in PAR1 were the chance result of population history, the same region was also studied in two additional populations, which are expected to show greater extents of allelic association. These included the Saami, who had a constant population size as well as the Vlax Roma of Bulgaria, who are a young founder population (Sajantila *et al.*, 1995; Slatkin, 1994). However, both populations showed rapid breakdown of LD, to similar extents as seen in North Europeans.

Sperm assays in three North European men over a ~9 kb-interval of LD-breakdown (510860- 520036 bp, NCBI36/hg18), revealed that COs in all three men clustered into a 1.9-2.5 kb-wide hot spot, wherein CO breakpoints were normally distributed. The mean CO frequency at the hot spot was 0.3% and peak activities ranged between 190 cM/Mb to 370 cM/Mb, making this hot spot, at the time of its identification, not only the first PAR hot spot, but also the most active that had been identified in the human genome.

Identification of the *SHOX* hot spot and characterisation of its morphology highlighted important similarities between this and autosomal hot spots, in terms of hot spot width and distribution of events. It thus provided the first indications that similar or indeed the same underlying recombination processes are at play in PAR1 as on autosomes, despite potentially distinct regulation of recombination initiation in the former (Chapter 1; Kauppi *et al.*, 2011).

Figure 3.1 High-resolution analysis of recombination in PAR1 by May *et al.* (2002)



a. Relative positions of the six intervals in and around the *SHOX* gene, subjected to LD analysis by May *et al*. (2002). b. Morphology of the *SHOX* hot spot in a North European semen donor. The centre of the hot spot, indicated in the figure using a dotted line, maps to 515,467 bp. All co-ordinates are as per NCBI36/hg18.

3.1.3. An extended survey of recombination in and around the *SHOX* gene

Identification of the *SHOX* hot spot, against the backdrop of a region that shows very little LD over extended intervals, raised important questions pertaining to the relative spacing of hot spots in PAR1 and inter-hot spot recombination rates and how these compared to autosomes. In order to address these issues and to better understand the processes of meiotic recombination in PAR1, the systematic survey of recombination was extended to ~30 kb of sequence flanking the *SHOX* hot spot (C.A. May, unpublished data; Figure 3.2).

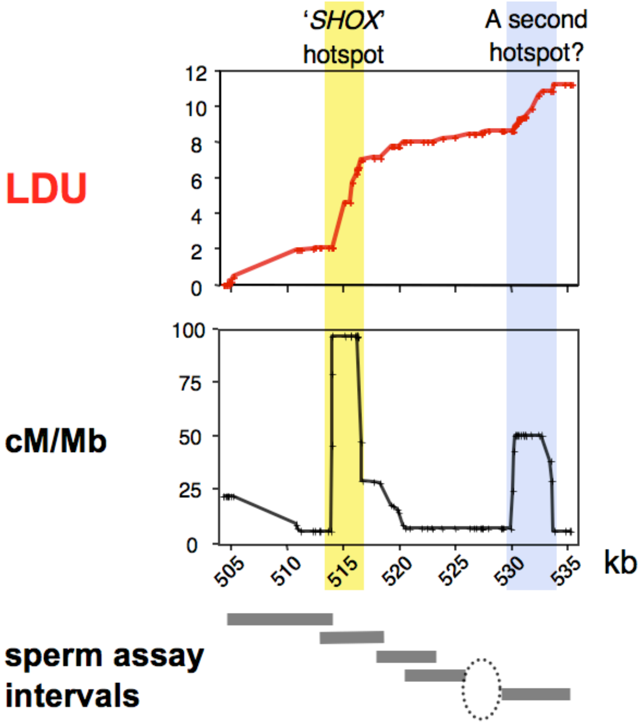
It is worth noting that similar surveys of recombination on autosomes have allowed a detailed understanding of the general recombination landscape in these regions. They have also provided in-depth information on the distributions, morphologies and relative intensities of sperm CO hot spots and also indicated how well these correspond to the observed patterns of haplotype diversity.

Thus, analysis of recombination over a 216 kb- wide interval in the class II MHC region and across a 206 kb region on chr1q42.3 showed remarkably similar landscapes, with clusters of hot spots located within 1-7 kb of each other, separated by 60-90 kb intervals of little or no recombination (Jeffreys, Kauppi and Neumann, 2001; Jeffreys *et al*., 2005). Hot spots identified from the two surveys also showed very similar peak activities, ranging between 0.9 and 70 cM/Mb on chr1 and between 0.4 and 140 cM/Mb in the MHC. Overall recombination activities across both intervals were also very similar and corresponded to 0.9 and 1.1 cM/Mb for the MHC and chr1 regions respectively.

It will be interesting to see how these data correspond to that from the extended *SHOX* survey in PAR1. Sperm data from three assay intervals around the *SHOX* hot spot (first, third and fourth intervals from the right in Figure 3.2) are already revealing important difference between PAR1 and autosomes in terms of inter-hot spot recombination rates (C.A. May, unpublished data). Indeed, recombination activities at such intervals ranged between 5 cM/Mb to 12 cM/Mb and were thus considerably higher compared to the 0.04

cM/Mb estimated for equivalent inter-hot spot intervals on autosomes (Jeffreys, Kauppi and Neumann, 2001; Jeffreys *et al*., 2005). This implies that there is a greater degree of 'leakage' of recombination activity outside of hot spots in PAR1 compared with autosomes, which may well be a consequence of the obligatory CO required in this region.

Figure 3.2 Patterns of inferred historical recombination activity at intervals in and around the *SHOX* gene



Graphs shown are different representations of patterns of haplotype diversity obtained from genotyping 133 markers in a panel of 99 North European semen donors. The top graph shows the LDU profile over the interval, whilst the one below represents historical recombination activity inferred using coalescent approaches. It should be noted that this analysis is based on twice as many markers compared to the May *et al.* (2002) study and is thus of higher resolution. Both graphs are consistent in that they both indicate high recombination activity in the interval encompassing the *SHOX* hot spot (highlighted in yellow), plus evidence of a second region of high-inferred activity (highlighted in blue). This is particularly pronounced in the graph representing data from coalescent analysis and represents a good candidate for sperm analysis. Intervals analysed using sperm assays, including the region described in this work, are indicated using grey bars. The dotted circle indicates a gap, which could not be assayed owing to a highly repetitive sequence context. Note: All positions are as per NCBI36/hg18 (chr X). LDU and coalescent analysis were both carried out by C.A. May.

3.1.4. This work

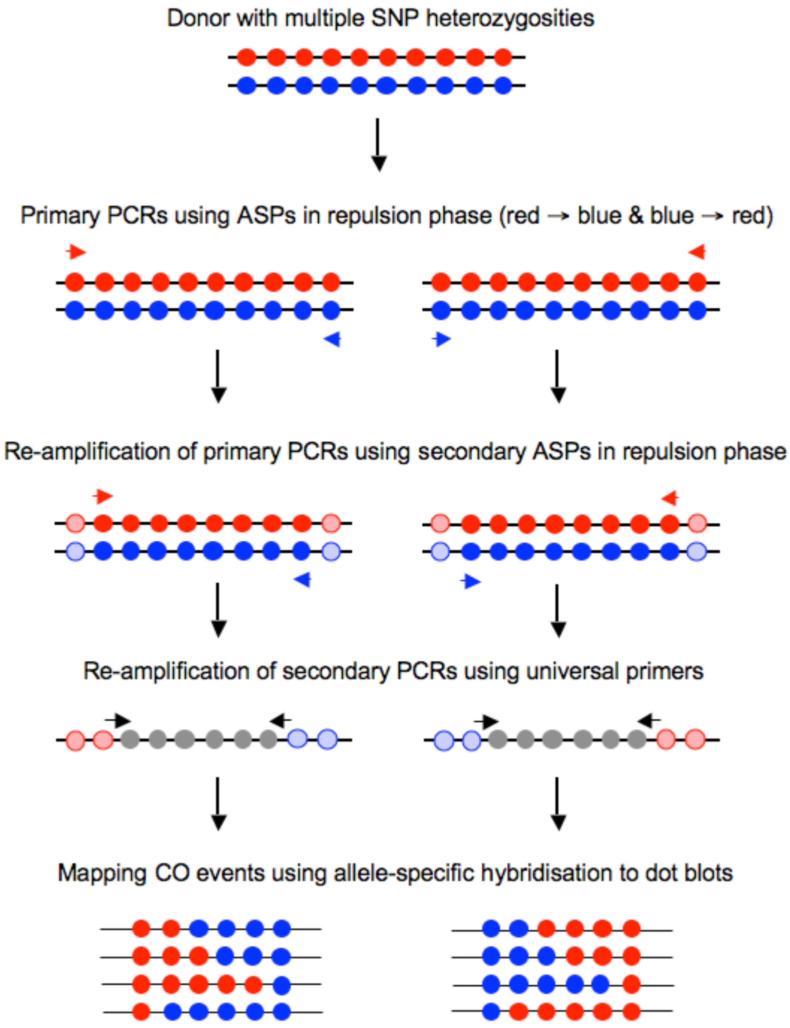
Data from May *et al.* (2002) showed the first evidence of clustering of COs into recombination hot spots in PAR1. This work represents an attempt to address the issue of hot spot spacing in PAR1 and how this compares to estimates from autosomes. This study will also further the current understanding of how well patterns of haplotype diversity reflect contemporary recombination rates in PAR1.

Thus the second region showing evidence for high inferred recombination activity, ~15 kb-proximal to the *SHOX* hot spot (highlighted in blue in Figure 3.2) was investigated using high-resolution sperm CO assays. A brief overview of the assay is provided here, but various stages of the assay are described in detail in section 3.2.

Effective sperm CO assays rely on the availability of sperm DNA from donors who carry multiple SNP heterozygosities across the chosen interval of LD breakdown. Next, pairs of selector sites 5′ and 3′ of this interval are chosen and allele-specific primers (ASPs) directed towards these sites are designed and optimised. Parental haplotypes (or linkage phase) are determined next using various combinations of internal and external allele-specific primers.

Recombinant molecules are selectively amplified using two rounds of repulsion-phase allele-specific PCR, wherein primers specific for alleles on one haplotype are used in combination with those on the other haplotype, as shown in Figure 3.3. Reciprocal exchange events (red-blue and blue-red combinations in Figure 3.3) can be recovered separately using suitable allele-specific primer combinations. CO breakpoints can be mapped by typing internal SNPs.

Figure 3.3 Strategy for a sperm CO assay


Donor with multiple SNP heterozygosities
Primary PCRs using ASPs in repulsion phase (red → blue & blue → red)
Re-amplification of primary PCRs using secondary ASPs in repulsion phase
Re-amplification of secondary PCRs using universal primers
Mapping CO events using allele-specific hybridisation to dot blots

Informative SNPs are indicated using circles, whilst red and blue colours indicate parental haplotypes. Red/blue arrows represent allele-specific primers (ASPs) and are positioned above corresponding selector sites. ASPs are specific to SNP alleles; thus, the ASP directed towards the red allele of the primary selector SNP should, in theory, only amplify the haplotype carrying this allele and not the opposite blue allele. In a CO assay, nested ASPs are used in repulsion-phase (that is, in red → blue and blue → red combinations), such that only molecules containing COs should be amplified. Secondary PCRs are re-amplified using universal primers (shown in black). Universal primers are designed within sequences that do not contain polymorphisms and hence, do not distinguish between haplotypes. Tertiary PCR products thus obtained are dot blotted. Allele-specific oligonucleotide hybridisation to dot blots is used to determine the haplotype status of internal SNPs and to infer CO-breakpoints. Various steps of the assay are described in greater detail in subsequent sections of this Chapter.

3.2. Results

3.2.1. Choice of donors

Ninety-nine North European men were genotyped for the interval encompassing the peak in inferred historical recombination activity (C.A. May and M. Denniff, unpublished data). Of these, only six men contained sufficient informative markers for an effective sperm CO assay. Table 3.1 shows genotypes for these men across the potential assay interval.

Table 3.1 Genotypes of six North-European men suitable for sperm assays

SNPs with ‘rs’ numbers were identified through dbSNP, whilst the rest were identified through re-sequencing (C.A. May and M. Denniff). Since the time of this study, many of these markers were genotyped in dbSNP (see Appendix 3.2 for details). The approximate centre of LD breakdown is highlighted in orange. Green highlights indicate potential selector sites. ‘H’ in the table is abbreviation for heterozygous. Out of the six short-listed men, d10 and d62 collectively had the best distribution of markers and sufficient sperm DNA for CO assays and were therefore selected.

Marker	Location (chrX)	Relative distance (bp)	Genotypes					
			d10	d62	d14	d85	d88	d93
35/34C/G	530213		C	C	C	H	C	C
rs28465428G/T	530291	78	G	G	G	G	G	H
3/341A/T	530392	101	H	H	H	H	H	H
rs28592040C/T	530569	177	C	C	C	C	C	C
rs28433655G/T	530670	101	T	T	T	G	T	T
3/663C/G	530715	45	H	H	H	H	H	H
3/849C/T	530901	186	T	T	H	T	T	H
rs3995663(COMPLEX)	531072	171	H	H	-	-	H	-
3/1034G/T	531086	14	T	T	T	H	T	T
3/1035A/G	531087	1	H	H	G	G	H	G
3/1076A/G	531129	42	H	G	H	G	G	H
3/1239C/G	531291	162	G	G	G	H	G	G
3/1248A/G	531299	8	G	H	G	H	H	G
35/1652C/G	531829	530	C	H	C	H	G	H
35/2287A/G	532465	636	H	G	G	G	G	H
35/2637A/G	532815	350	H	G	G	H	H	G
35/3390C/G	533567	752	G	G	G	G	G	G
INDEL103	533621	54	H	H	+	+	+	+
rs28374757A/G	533714	93	G	G	H	G	G	G
rs34995914A/G	533877	163	H	H	H	H	H	H
35/4673A/G	534840	963	H	H	H	H	H	H
rs28459931C/T	535272	432	H	H	H	H	H	H
rs28412402C/T	535328	5	H	H	H	H	H	H

Having short-listed two sperm donors, d10 and d62, for CO assays, pairs of 3′ and 5′ selector sites were identified next. It is imperative that both 3′ and 5′ selector sites fall outside of the putative hot spot. SNPs 3/341A/T, 3/663C/G, 35/4673A/G and rs28459931C/T were chosen as candidate selector sites for both men and ASPs directed towards them were designed. rs28412402C/T was not chosen as it is present within an AT-rich region and an ~50% CG content is preferred for ASPs. rs34995914A/G was not used since it is located within an Alu element and ectopic mispriming at multiple Alus risks amplification of non-specific products.

3.2.2. Allele-specific primer design and optimisation

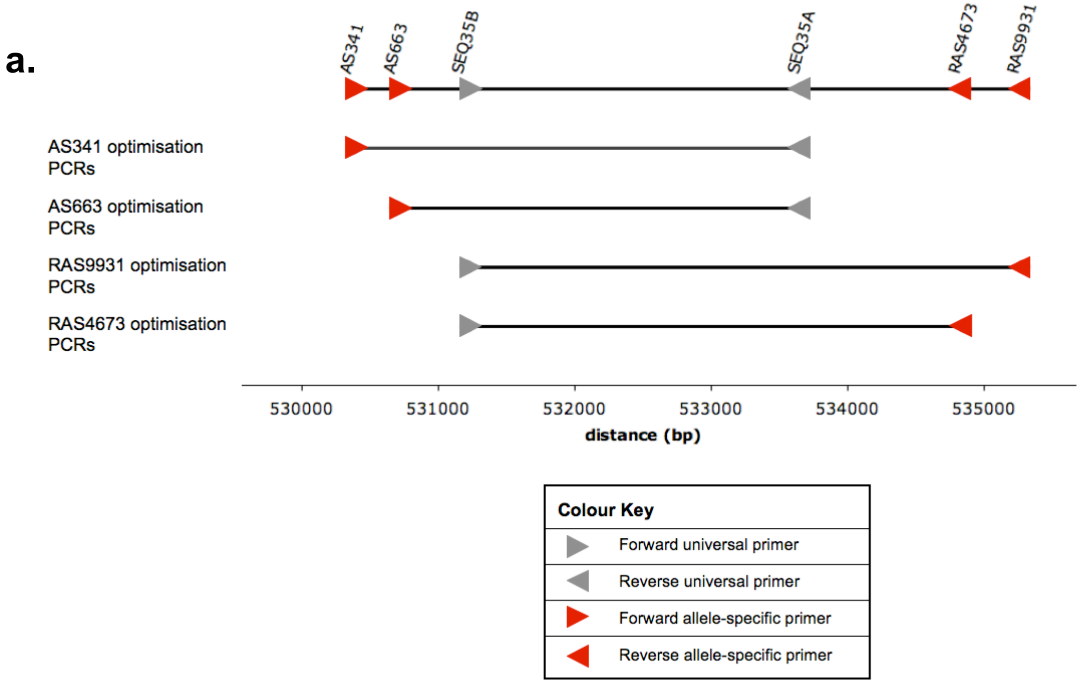
The specificity of an ASP is determined by its last 3' base. ASPs are usually 18bp long, although very GC-rich primers may be 1-2 bases shorter. The GC-content of primers is mainly determined by the genomic context of the SNP towards which the primer is designed, although 5' G/C-tails can be added to boost the GC-content of relatively AT-rich primers. This is often helpful for improving primer efficiency.

Overall, it is difficult to predict the behaviour of ASPs; therefore, it is necessary to test them over a range of temperatures so that optimum specificity and efficiency is achieved for when they are used in a CO assay. ASPs were paired with universal primers and tested in individuals homozygous for one or the other allele for this purpose. Figure 3.4 shows the strategy for optimisation of ASPs used in this assay.

3.2.3. Linkage phasing of semen donors

5′ and 3′ selector sites are likely to be in free association and linkage phase will thus be unknown. Haplotypes were therefore determined for both donors, using all four possible combinations of various ASP pairs, as shown below.

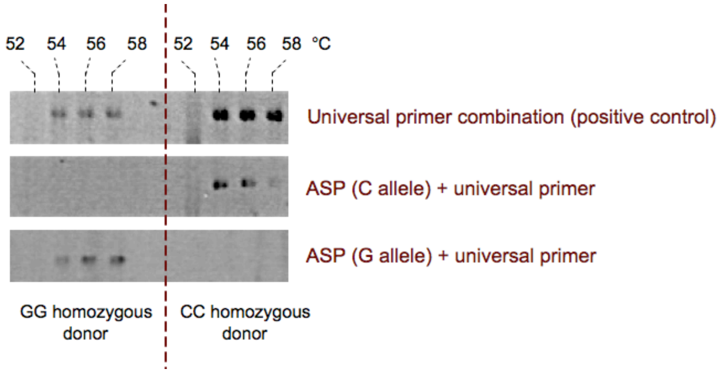
Figure 3.4 Allele-specific primer optimisation


a.
AS341
AS663
SEQ35B
SEQ35A
RAS4673
RAS9931
AS341 optimisation PCRs
AS663 optimisation PCRs
RAS9931 optimisation PCRs
RAS4673 optimisation PCRs
530000
531000
532000
533000
534000
535000
distance (bp)
Colour Key
Forward universal primer
Reverse universal primer
Forward allele-specific primer
Reverse allele-specific primer

a. Different primer combinations used to optimise individual ASPs. AS341A/T, AS663C/G, RAS4673A/G and RAS9931C/T are ASPs designed against SNPs 3/341, 3/663, 35/4673 and rs28459931 respectively. In order to determine primer efficiency and specificity, each ASP was paired with a universal primer and tested over a range of temperatures. In all cases, a universal primer combination was used as positive controls (not shown). For example, to optimise AS341, both versions of the ASP were paired with universal primer SEQ35A and tested over various temperatures. All primer sequences are listed in Appendix 3.1.

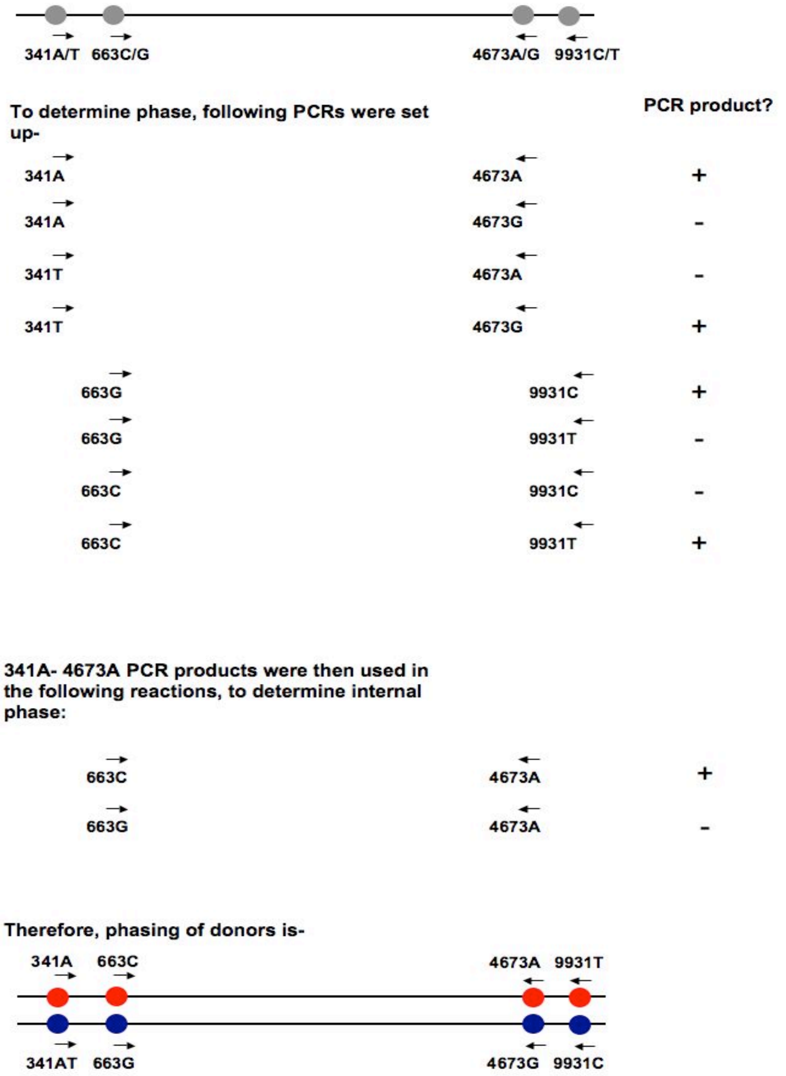
b. Primer sequence:

- GAGGAAAGCTGGGGGTAG - 3 AS663C
- GAGGAAAGCTGGGGGTAC - 3' AS663G


52
54
56
58
52
54
56
58 °C
Universal primer combination (positive control)
ASP (C allele) + universal primer
ASP (G allele) + universal primer
GG homozygous donor
CC homozygous donor

b. Gel electrophoretic analysis of PCR products from ASP optimisation assays. As an example, gel electrophoresis images for AS663C/G optimisation PCRs are shown, with PCRs at different annealing temperatures, as indicated. Both AS663C and AS663G are specific between 54 and 58°C, although product yield is best at 54°C. Other allele-specific primers were similarly optimised for specificity and efficiency.

Figure 3.5 Phasing donors for CO assay



To determine phase, following PCRs were set up-

		PCR product?
341A	4673A	+
341A	4673G	-
341T	4673A	-
341T	4673G	+
663G	9931C	+
663G	9931T	-
663C	9931C	-
663C	9931T	+

341A- 4673A PCR products were then used in the following reactions, to determine internal phase:

663C	4673A	+
663G	4673A	-

Therefore, phasing of donors is-

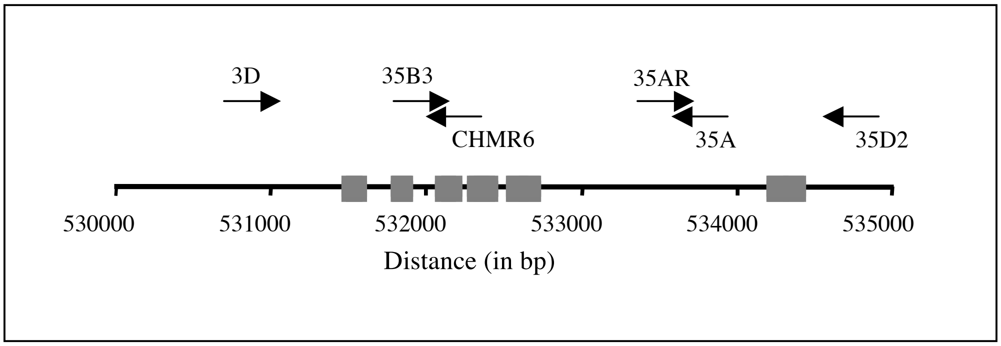
341A/T and 9931C/T are external (primary) ASPs used in the assay, while 663C/T and 4673A/G are internal (secondary) ASPs. PCRs were set up to determine linkage phase of d10 and d62 used in the assay. Both had the same phasing, as shown above. Red and blue colours were assigned arbitrarily in order to distinguish the two haplotypes. Note: Primer sequences and PCR conditions are listed in Appendix 3.1.

Thus, haplotypes can be inferred from positive PCRs generated from various ASP combinations. It is important to note that under no circumstances were phasings done using external primer combinations, as PCR products can contaminate subsequent CO assays. ASPs for alleles on one haplotype could now be used in combination with those specific for the other haplotype, to selectively amplify recombinant molecules.

3.2.4. Re-sequencing d10 and d62 over the assay interval

Prior to setting up sperm CO assays in d10 and d62, both individuals were re-sequenced over the assay interval to identify any polymorphisms that had not been previously reported, but which would allow CO-breakpoints to be mapped more precisely. Separated haplotypes obtained from phasing PCRs were re-amplified for this purpose and subsequently sequenced, as described in Figure 3.6.

Figure 3.6 Re-sequencing donors over the assay interval



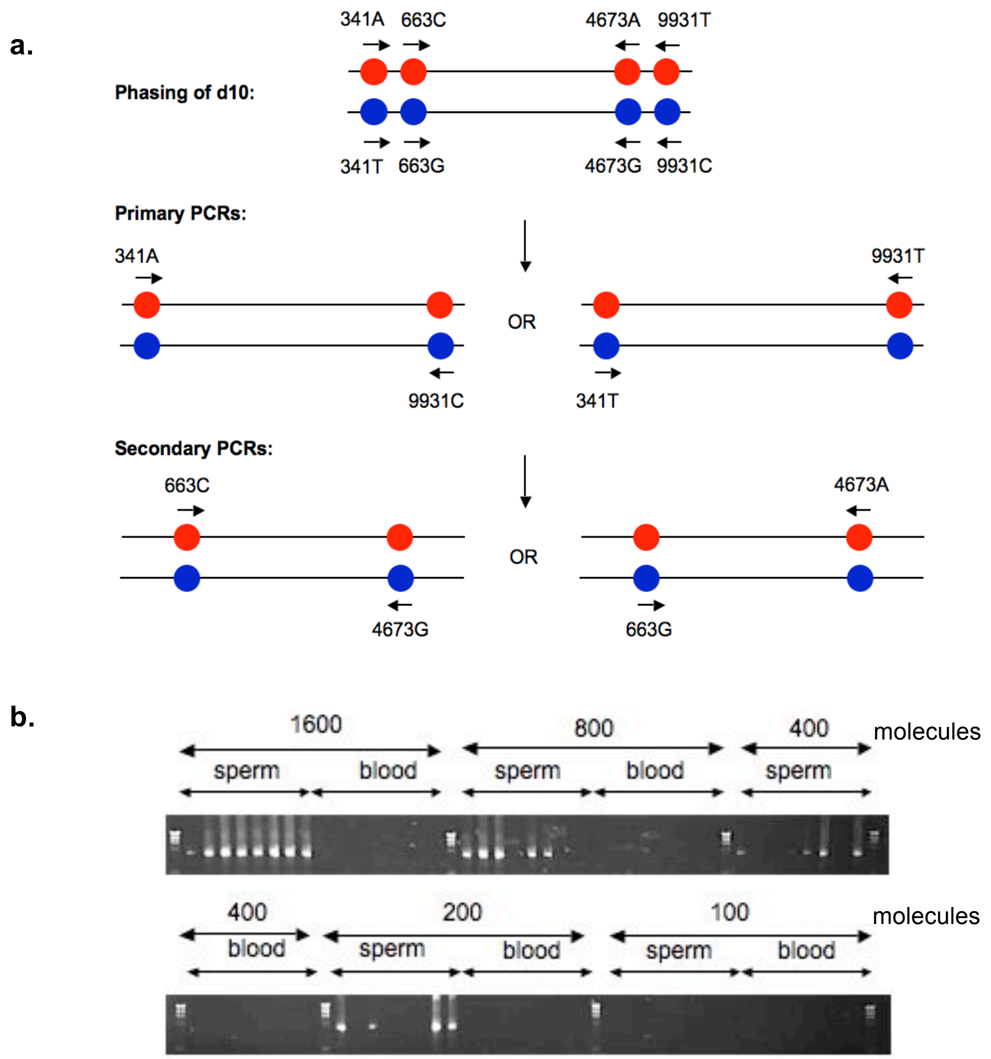
Constituent donor haplotypes were re-amplified using primers 3D (forward) and 35D2 (reverse) and sequenced using all six primers indicated (sequences listed in Appendix 3.1). The protocol used is as described in Chapter 2. Grey boxes indicate gaps that could not be re-sequenced accurately due to the presence of poly-A/T tracts.

No additional polymorphisms were identified in d10. However, a previously unreported C/T polymorphism was identified in d62. This is referred to as marker 35/1400 henceforth and maps to 531400 (NCBI36/hg18).

3.2.5. Pilot CO assay

Human CO hot spots are extremely variable in their recombination frequency (0.0004% to 1% (Kauppi *et al.*, 2009). A pilot assay using a wide range of DNA inputs was therefore set up first, to optimise DNA inputs for the subsequent bulk recovery of CO molecules (Figure 3.7).

Figure 3.7 Pilot CO assay for d10


a.
Phasing of d10:
341A 663C 4673A 9931T
341T 663G 4673G 9931C
Primary PCRs:
341A
9931C
OR
9931T
341T
Secondary PCRs:
663C
4673G
OR
4673A
663G
b.
1600
800
400
molecules
sperm
blood
sperm
blood
sperm
400
200
100
molecules
blood
sperm
blood
sperm
blood

A. Strategy for the pilot CO assay. ASPs 341A/T and 9931C/T were used in repulsion phase in primary PCRs. These PCR products were then used to seed secondary PCR reactions, in which ASPs 663C/G and 4673A/G were used in repulsion phase.

B. Gel electrophoretic data from secondary PCRs for the d10 pilot CO assay. Data from only one orientation, that is, one combination of repulsion-phase ASPs, are shown. DNA inputs for the pilot assay were first calculated as follows. Given that a haploid genome contains 3pg DNA, 6pg of genomic DNA should contain, on average, one molecule of each haplotype. However, not all DNA molecules amplify during allele-specific PCR and 50% efficiency is routinely assumed for these assays (Kauppi *et al.*, 2009). Therefore, 12pg genomic DNA should contain on average, one amplifiable molecule of each haplotype. Based on this, five DNA inputs ranging from 1600 molecules - 100 molecules per PCR were analysed. Eight reactions (for sperm and blood) were set up for each input. After two rounds of allele-specific PCR, secondary PCR products were analysed using agarose gel electrophoresis to detect positive and negative pools. Authentic CO events have never been detected in blood and therefore, blood DNA were used as a negative control. Note: DNA ladder is as described in section 2.1.5).

For each orientation, that is each combination of ASPs from opposite haplotypes (red to blue and blue to red in Figure), two rounds of allele-specific PCR were used to selectively amplify recombinant molecules from pools of sperm DNA. Secondary PCR products were then analysed using agarose gel electrophoresis, as shown in Figure 3.7. This showed increasing numbers of positive signals, with increasing sperm DNA inputs, consistent with a Poisson distribution. No positive signals were however detected from blood DNA, which was used as a negative control. Numbers of positive and negative PCR reactions for each sperm DNA input were used to estimate the mean number (m) of CO molecules per PCR for that input, using Poisson approximation, as follows:

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

Maximum likelihood methods were then used for a more precise estimate of CO frequency by combining data across all inputs. The mean CO frequency calculated for the pilot assay shown was 0.32%. A similar pilot CO assay for d62 gave a CO frequency estimate of 0.045% (data not shown).

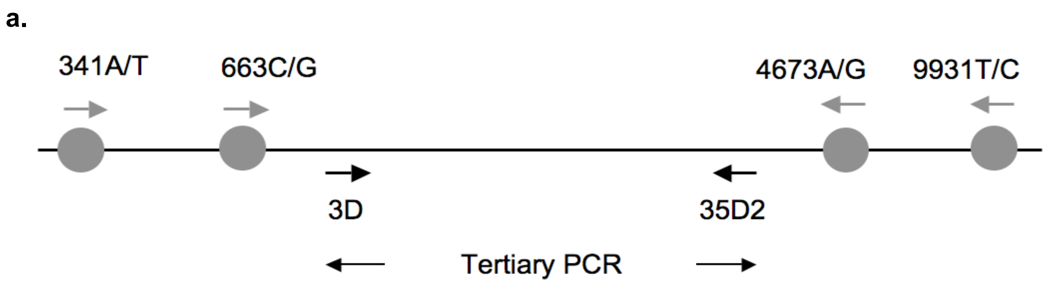
3.2.6. Bulk recovery of CO molecules

CO frequencies estimated from the pilot assay were used to optimise DNA inputs for the bulk recovery of CO molecules using reciprocal CO assays, such that most pools contained, at most, one CO-positive molecule. Primary and secondary PCRs were set up as described in Figure a. Thus, ASPs 341A/T and 9931A/T were used in repulsion-phase in primary PCRs. Products obtained from these reactions were then used to seed secondary PCRs, in which 663C/T and 4673A/G were used as forward and reverse ASPs respectively. In total, ~103,000 and ~234,000 sperm molecules were screened per orientation for d10 and d62 respectively. Gel electrophoretic analysis of secondary PCR products (as shown in **Figure** b) revealed 196 and 138 CO-positive pools for d10 and d62 respectively. This implied a CO frequency of ~0.2% (95% C.I.= 0.16-0.22%) and 0.05% (95% C.I.= 0.04-0.07%) for d10 and d62 respectively. True CO frequencies are however, likely to be higher, since only one CO-event can be detected per pool from analysis of gel electrophoresis data. This is discussed further in the next section.

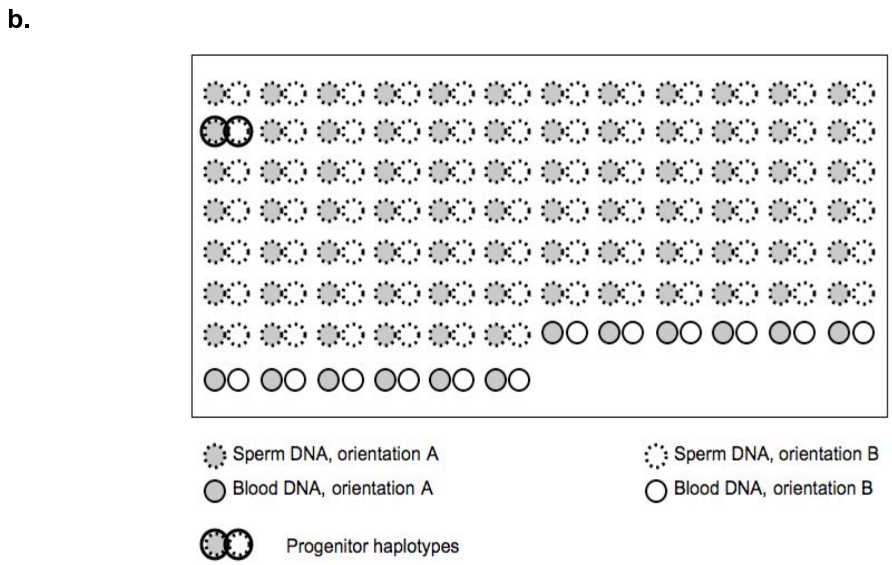
All secondary products were re-amplified using universal primers (Figure 3.8a; see Appendix 3.1 for details on primer combinations and PCR conditions). Universal primers are not haplotype-specific; thus, the same primers were used to amplify secondary products from either orientation. Tertiary PCRs were carried out in order to explore the nature of negative secondary PCRs and to determine whether these contained CO events that were too weak to be detected on an agarose gel or bleed-through from one or the other parental haplotype, which can result from ASPs mis-priming from the unselected haplotype. Tertiary PCRs were also important to ensure that enough PCR products were available per reaction for CO mapping using allele-specific oligonucleotide (ASO) hybridisation to dot blots.

Tertiary products for both men assayed were dot blotted as shown in Figure 3.8b. The same dot blots also contained progenitor haplotypes, in order to determine internal SNP haplotypes and infer CO breakpoints. Allele-specific oligonucleotide hybridisation was then used to determine the status of internal SNPs (as shown in Figure 3.8c). Details of ASOs used are listed in Appendix 3.2.

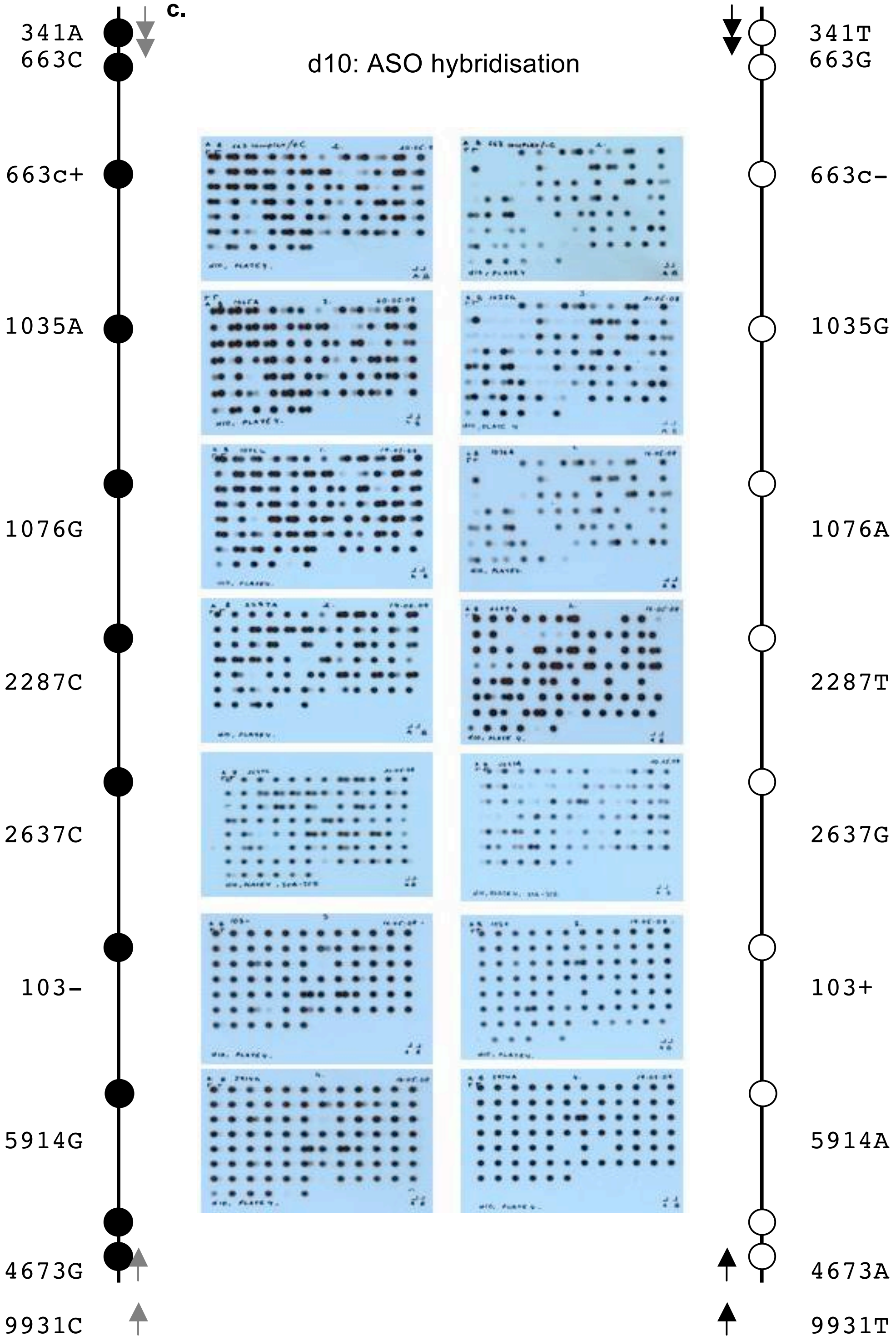
Figure 3.8 Mapping CO breakpoints in d10 using ASO hybridisation to dot blot


a.
341A/T
663C/G
4673A/G
9931T/C
3D
35D2
Tertiary PCR

a. Relative positions of universal primers used in tertiary PCRs for d10 and d62 CO assays. Tertiary primers used are indicated using black arrows. Grey circles represent primary and secondary selector sites used in the CO assays and grey arrows represent ASPs designed against these. Note: The figure is not to scale.


b.
Sperm DNA, orientation A
Sperm DNA, orientation B
Blood DNA, orientation A
Blood DNA, orientation B
Progenitor haplotypes

b. Dot blot positions of various tertiary PCR products (different orientations, DNAs and progenitors). CO assay dot blots were probed for both alleles of informative SNPs located within the assay interval, in order to determine the haplotype status of these markers and thus, to infer CO breakpoints (as shown in part 'c'). Orientation names, 'A' and 'B', were arbitrarily assigned.


c.
d10: ASO hybridisation
341A
663C
663c+
1035A
1076G
2287C
2637C
103-
5914G
4673G
9931C
341T
663G
663c-
1035G
1076A
2287T
2637G
103+
5914A
4673A
9931T

c. Typing of internal SNPs. Constituent haplotypes in d10 are shown on either side (relative distances are not to scale). Black and white circles represent the two alleles at each informative SNP on each haplotype. CO assay products were first hybridised with one allele of an informative SNP, followed by the other. CO breakpoints were inferred where markers switch from one haplotype to the other. Note: ASOs are referred to using the last three/four digits of their 'rs numbers' or 'local numbers'. For more details on ASOs, see Appendix 3.2.

3.2.7. Scoring COs for d10 and d62

CO-positive reactions were either ‘simple’ or ‘mixed’. ‘Simple’ reactions contained one class of CO only, and whilst there might be more than one CO molecule mapping to the same interval, these cannot be detected; nonetheless, they can be estimated by Poisson analysis. ‘Mixed’ reactions contain at least two different COs and therefore give positive signals upon hybridisation with both alleles at one or more SNP sites. These can be scored as two COs, one with a breakpoint 5′ of where mixed signals start, and the other 3′ of where the mixed signals end. **Error! Reference source not found.** illustrates these steps using a sample data set from d10.

Figure 3.9 Scoring COs

a. CO data from a subset of d10 PCRs. All PCRs shown were amplified using 341A-5914T as primary ASPs and 663C-4673A as secondary ASPs (underlined in the figure). Locations refer to positions on a 96-well plate (which were used to set up PCR reactions). Yields of secondary PCR products were recorded from gel photos after the second-round of allele-specific PCR (as shown in Figure 3.7) Varying numbers of ‘+’ symbols reflect the strength of PCR signals. Positive pools were expected to contain at least one recombinant molecule. ‘-’ indicates the absence of a PCR signal; ‘ +/- ’ are assigned where PCR products are barely detectable on the gel. Although ‘-‘ pools are not expected to contain COs, this was not always true in practice. Thus, CO mapping revealed the presence of COs in a small subset of these pools. False negatives such as these most likely result as a consequence of the low efficiency of ASPs, coupled with inadequate levels of amplification. Haplotype statuses of internal SNPs were determined as shown in Figure 8.3c. Note: ‘663’ refers to secondary selector site 3/663C/G whilst ‘663c’ refers to internal SNP rs3995663 (COMPLEX). ‘M’ indicates mixed pools.

a.

Haplotypes		341	663	663c	1035	1076	2287	2637	103	5914	4673	5914
		A	*C*	-	G	A	G	A	+	A	*G*	*C*
		T	*G*	+	A	G	A	G	-	G	*A*	*T*
Location	PCR signal											
1	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
2	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
3	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
4	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
5	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
6	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
7	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
8	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
9	+	*A*	*C*	M	M	M	M	M	-	G	*A*	*T*
10	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*

11	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
12	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
13	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
14	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
15	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
16	+	*A*	*C*	M	M	G	A	G	-	G	*A*	*T*
17	+/-	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*
18	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
19	+/-	*A*	*C*	-	G	A	G	M	-	G	*A*	*T*
20	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
21	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
22	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
23	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
24	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
25	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
26	+	*A*	*C*	-	G	M	A	G	-	G	*A*	*T*
27	+	*A*	*C*	-	G	M	A	G	-	G	*A*	*T*
28	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
29	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
30	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
31	+/-	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
32	+	*A*	*C*	M	M	M	M	G	-	G	*A*	*T*
33	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
34	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
35	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
36	+	*A*	*C*	M	M	G	A	G	-	G	*A*	*T*
37	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
38	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
39	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
40	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
41	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
42	+	*A*	*C*	M	M	G	A	G	-	G	*A*	*T*
43	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
44	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
45	+++	*A*	*C*	M	M	M	M	M	-	G	*A*	*T*
46	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
47	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
48	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
49	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
50	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
51	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
52	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
53	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
54	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
55	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
56	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
57	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
58	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
59	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
60	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
61	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
62	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
63	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
64	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
65	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
66	+	*A*	*C*	-	G	M	A	G	-	G	*A*	*T*
67	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
68	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
69	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
70	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
71	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
72	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
73	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
74	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
75	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
76	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
77	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
78	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*

b. Resolution of mixed pools. Mixed pools give positive signals with both alleles of a given SNP, indicating the presence of more than one recombinant molecule in the same pool. A marker interval flanked on one side by a mixed site can still be scored for COs. However, a marker interval, flanked on both sides by mixed sites cannot be scored for COs. For example, in pool 5, the intervals between markers 663c and 1035 and between 1035 and 1076, cannot be scored for COs. Haplotypes obtained from the resolution of mixed pools are referred to as 'a' and 'b'.

b.

Haplotypes		341	663	663c	1035	1076	2287	2637	103	5914	4673	5914
		A	*C*	-	G	A	G	A	+	A	*G*	*C*
		T	*G*	+	A	G	A	G	-	G	*A*	*T*
Location	PCR signal											
1	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
2	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
3	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
4	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
5	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
5a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
5b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
6	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
7	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
8	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
9	+	*A*	*C*	M	M	M	M	M	-	G	*A*	*T*
9a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
9b	+	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*
10	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
11	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
12	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
13	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
13a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
13b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
14	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
15	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
16	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
16a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
16b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
17	+/-	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*
18	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
19	+/-	*A*	*C*	-	G	A	G	M	-	G	*A*	*T*
19a	+/-	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*
19b	+/-	*A*	*C*	-	G	A	G	G	-	G	*A*	*T*
20	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
21	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
21a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
21b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
22	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
22a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
22b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
23	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
24	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
25	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
26	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
26a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
26b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
27	+	*A*	*C*	-	G	M	A	G	-	G	*A*	*T*
27a	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
27b	+	*A*	*C*	-	G	G	A	G	-	G	*A*	*T*
28	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
28a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
28b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*

29	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
29a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
29b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
30	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
31	+/-	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
31a	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
31b	+/-	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
32	+	*A*	*C*	M	M	M	M	G	-	G	*A*	*T*
32a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
32b	+	*A*	*C*	-	G	A	G	G	-	G	*A*	*T*
33	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
34	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
35	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
35a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
35b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
36	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
36a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
36b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
37	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
38	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
39	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
40	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
41	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
41a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
41b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
42	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
42a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
42b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
43	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
44	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
44a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
44b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
45	+++	*A*	*C*	M	M	M	M	M	-	G	*A*	*T*
45a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
45b	+++	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*
46	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
47	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
47a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
47b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
48	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
49	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
50	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
50a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
50b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
51	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
51a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
51b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
52	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
52a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
52b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
53	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
54	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
55	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
56	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
56a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
56b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
57	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
58	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
58a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
58b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
59	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
60	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
61	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
62	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
63	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
64	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
65	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
66	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
66a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
66b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*

67	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
67a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
67b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
68	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
69	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
70	+++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
70a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
70b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
71	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
72	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
73	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
74	++	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
74a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
74b	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
75	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
76	+	*A*	*C*	M	M	M	A	G	-	G	*A*	*T*
76a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
76b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
77	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
78	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*

c. COs arranged according to the distribution of breakpoints. Classifying all d10 COs (including those not shown in the figure) according to the interval to which they map, showed that ~46% of COs mapped to the interval between the forward secondary selector site and internal marker 663c (rs3995663) and ~1% between marker 5914 and the reverse secondary selector site. These terminal events cannot be verified by CO mapping without extending the assay interval and may represent bleed-through from one or the other parental haplotype (as shown in part ‘d’).

c.

Haplotypes		341	663	663c	1035	1076	2287	2637	103	5914	4673	5914
		A	*C*	-	G	A	G	A	+	A	*G*	*C*
		T	*G*	+	A	G	A	G	-	G	*A*	*T*
Location	PCR signal											
1	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
2	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
3	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
4	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
5a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
6	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
7	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
8	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
9a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
10	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
11	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
12	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
13a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
14	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
15	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
16a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
18	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
20	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
21a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
22a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
23	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
24	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
25	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
26a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
28a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*

29a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
30	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
31a	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
32a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
33	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
34	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
35a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
36a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
37	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
38	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
39	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
40	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
41a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
42a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
43	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
44a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
45a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
46	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
47a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
48	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
49	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
50a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
51a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
52a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
53	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
54	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
55	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
56a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
58a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
59	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
60	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
61	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
62	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
63	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
64	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
65	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
66a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
67a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
68	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
69	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
70a	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
71	-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
72	+++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
73	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
74a	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
75	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
76a	+	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
77	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
78	+/-	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
5b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
13b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
16b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
21b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
22b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
26a	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
27a	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
28b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
29b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
31b	+/-	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
35b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
36b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
41b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
42b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
44b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
47b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
50b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
51b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
52b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*

56b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
57	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
58b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
66b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
67b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
70b	+++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
74b	++	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
76b	+	*A*	*C*	-	G	A	A	G	-	G	*A*	*T*
19b	+/-	*A*	*C*	-	G	A	G	G	-	G	*A*	*T*
32b	+	*A*	*C*	-	G	A	G	G	-	G	*A*	*T*
9b	+	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*
17	+/-	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*
19a	+/-	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*
45b	+++	*A*	*C*	-	G	A	G	A	-	G	*A*	*T*

d. COs mapping to the terminal intervals. COs mapping to the end intervals (that is, between a selector sites and a terminal SNP) cannot be verified by ASO hybridisation and may represent either a genuine event (option 1 in figure) or bleed-through of the 'blue' parental haplotype, as a result of forward ASPs mis-priming with the 341T-663G haplotype (option 2). Bleed-throughs can artificially inflate estimates of CO frequencies. Thus, in order to maintain an accurate dataset, these events were excluded from further analysis.

Haplotypes		341	663	663c	1035	1076	2287	2637	103	5914	4673	5914
		A	*C*	-	G	A	G	A	+	A	*G*	*C*
		T	*G*	+	A	G	A	G	-	G	*A*	*T*
Location	PCR signal											
Option 1-												
1	++	*A*	*C*	+	A	G	A	G	-	G	*A*	*T*
Option 2-												
1	++	*T*	*G*	+	A	G	A	G	-	G	*A*	*T*

The method described above is a conservative way of scoring COs, since COs mapping to terminal intervals are excluded from analysis. This end interval problem was particularly pronounced at the 5′ end of the assay interval and may either imply that a significant number of 5′ COs are being missed using the current assay interval or that the forward ASPs are relatively less specific, allowing bleed-through of the 'blue' parental haplotype in primary and secondary PCRs.

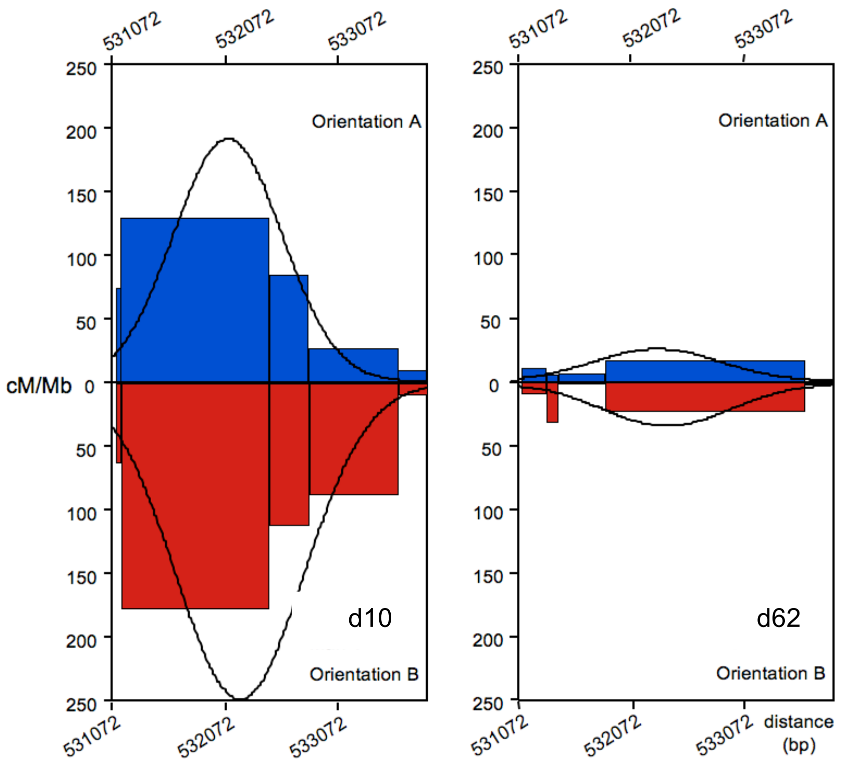
Further, ambiguous intervals flanked on either side by mixed sites are not scored for COs using the described method. In order to account for events masked by such mixed pools, a ‘Poisson correction’ is applied. According to this, the corrected number of COs = -Nln[(N-R)/N], where N= number of PCR reactions scorable for a given type of CO and R= number of reactions positive for a particular class of recombinants. Corrected numbers thus obtained were used to calculate recombination frequencies, as discussed below.

3.2.8. Comparison of recombination activity and distribution of COs in d10 and d62

In total, ~85,600 molecules were assayed for d10 and ~470,600 molecules for d62. 253 and 195 corrected-COs were analysed for the two men respectively. There were no significant differences in numbers of COs recovered in either orientation, indicating that COs were reciprocal in rate. Comparison of overall CO frequencies between the two men however, revealed a 7-fold difference, with COs frequencies in d10 and d62 at 0.29% and 0.041% respectively. In order to study the distribution of COs across the assay interval, CO frequency per unit length of DNA (cM/Mb) was calculated per marker interval for both men (as shown in Appendix 3.3) and values plotted against distance (Figure 3.10). For both men, the distribution of events was consistent with presence of a ‘conventional’ recombination hot spot, wherein CO breakpoints are normally distributed. This hot spot will henceforth be referred to as hot spot SHOX-5, since this is the fifth interval analysed as part of the extended *SHOX* survey.

It should be mentioned here that the 5′ end of the hot spot in d10 was not captured in full, as is evident from the skewed normal distribution shown in Figure 3.10. In light of this, it is likely that at least some of the 5′ end interval events identified in d10 were representative of real COs, rather than bleed-through of the ‘blue’ parental haplotype. Unfortunately however, this cannot be fully verified without extending the current assay, such that additional 5′ markers can be typed for COs. Extending the assay is a considerable challenge given the absence of informative markers in close proximity to the current selector sites (Table 3.1) and in light of the difficulties associated with carrying out effective allele-specific PCRs in such a repeat-rich genomic region. Given this and considering time constraints, the assay was not modified to type additional markers. Terminal events in the current assay were nonetheless excluded from analysis largely for the purposes of maintaining a conservative dataset.

Figure 3.10 Distribution of reciprocal COs in d10 and d62



Peak CO activity for each interval in cM/Mb is plotted against distance in bp. Specific numbers of COs recovered from each interval are listed in Appendix 3.3. Bell-shaped curve shows the underlying least squares best-fit distributions of COs, assuming that they are normally distributed across the hot spot.

Assuming a normal distribution of COs across the hot spot, maximum-likelihood approaches were used to calculate the least-square best-fit values for hot spot centre, width and peak recombination activity (Table 3.2).

Table 3.2 Least-square best-fit values for hot spot centre, 95% hot spot width and peak recombination activity

All values were obtained using a crossover analysis program written in True Basic 4.0 by A.J. Jeffreys. This program uses information on the position of each informative marker and the cumulative frequency of COs at each marker to estimate the least-square best fit values of hot spot centre, width and peak activity. Note: Inferences of hot spot centre and width for d10 should be treated with caution since the inability to recover all COs from the 5′ end of the hot spot will have an effect on these calculations.

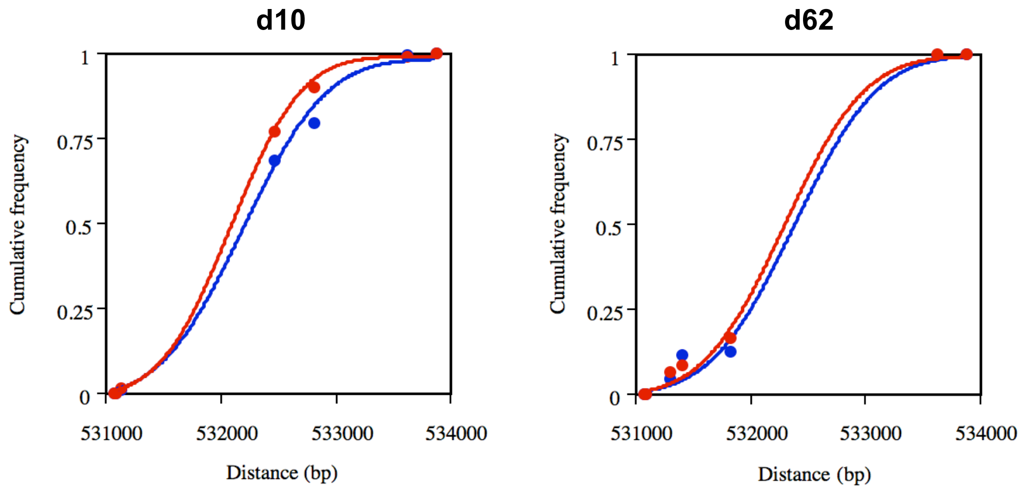
d10:	Hot spot centre (bp)	Hot spot Width (bp)	Peak activity (cM/Mb)
Orientation A	532103	1930	190
Orientation B	532207	2280	250
A and B combined	532150	2110	220

d62:	Hot spot centre (bp)	Hot spot Width (bp)	Peak activity (cM/Mb)
Orientation A	532311	2200	25
Orientation B	532398	2280	34
A and B combined	532356	2260	29

Values for hot spot centres calculated for each orientation differed by 104 bp in d10 and 87 bp in d62. This disparity is the likely consequence of subtle shifts in the distribution of reciprocal COs, which are more readily detectable in corresponding cumulative frequency distributions (shown in Figure 3.11).

The location of the hot spot (using data combined for both orientations) is however, consistent between the two donors, with the hot spot in d62 mapping to a slightly wider interval of 2254 bp, as opposed to 2114 bp in d10. This was however, not significant and probably a result of lack of additional informative markers for more accurate mapping of COs. Missing COs from the 5′ end of the hot spot in d10 would have also affected estimates of the hot spot centre in this donor and contributed to the disparity observed between the two men.

Figure 3.11 Cumulative frequency distribution of CO breakpoints in d10 and d62



Cumulative frequencies of COs for each orientation (red and blue circles), assuming normal distribution of COs across the hot spot, are plotted against distance (in bp). S-shaped curves for each dataset show the underlying least squares best-fit cumulative distributions for each orientation. In d62, the distributions of reciprocal COs are very similar and no significant differences are detected. In d10, however, there is some asymmetry (shift) in the distribution of reciprocal COs. This was however, only of borderline significance; Table 3.3 describes this in more detail.

3.2.9. Transmission distortion in reciprocal COs

When reciprocal COs map to the same locations, ~50% of the recombinant progeny will carry the same allele at a given SNP. However, when reciprocal events map to different locations causing shifts in reciprocal CO distributions or 'asymmetry' (perhaps seen in d10), there is a deviation from a 50:50 transmission ratio and alleles from a particular haplotype are over-transmitted relative to the other. This 'transmission distortion' (TD) accompanying COs has been observed previously in autosomal hot spots and is discussed in section 3.3. Therefore, data at this hot spot were tested for any evidence of significant TD. Transmission frequencies of hot spot alleles in d10 and d62 are shown in the following table.

Table 3.3 Transmission frequencies of hot spot alleles in d10 and d62

Transmission frequencies of marker alleles, normalised to equal numbers of orientation A and B COs, are shown. Orientations were named A and B arbitrarily and were represented in red and blue respectively in Figure 3.10 and Figure 3.11. P values were calculated using two-tailed Fishers Exact tests. Marker 35/2637 showed significant TD in d10 (*P*=0.025). However, after Bonferroni correction, the TD observed was only of borderline significance (*P*=0.05).

d10:

Marker	Allele	Orientation A COs	Orientation B COs	T.F.	95% C.I.	P value
rs3995663	+	0	153	0.500	0.489-0.517	1
	-	100	0	0.500		
3/1035	A	0	153	0.500	0.489-0.517	1
	G	100	0	0.500		
3/1076	G	1	152	0.502	0.487-0.522	1
	A	99	1	0.498		
35/2287	A	77	49	0.545	0.498-0.590	0.153
	G	23	104	0.455		
35/2637	G	90	32	0.555	0.517-0.591	0.025
	A	10	121	0.445		
INDEL103	-	99	1	0.498	0.478-0.512	1
	+	1	152	0.502		
rs34995914	G	100	0	0.500	0.489-0.517	1
	A	0	153	0.500		

d62:

Marker	Allele	Orientation A COs	Orientation B COs	T.F.	95% C.I.	P value
rs3995663	+	0	113	0.500	0.484-0.522	1
	-	82	0	0.500		
3/1035	A	0	113	0.500	0.484-0.522	1
	G	82	0	0.500		
3/1248	G	24	72	0.465	0.412-0.520	0.357
	A	58	41	0.535		
1400	C	25	66	0.444	0.39-0.503	0.133
	T	57	47	0.556		
35/1652	C	30	65	0.471	0.416-0.528	0.46
	G	52	48	0.529		
INDEL103	-	82	1	0.504	0.484-0.523	1
	+	0	112	0.496		
rs34995914	G	82	0	0.500	0.484-0.522	1
	A	0	113	0.500		

Thus, in contrast to d62, who showed no instances of significant TD, d10 showed over-transmission of the ‘G’ allele at marker 35/2637 (*P*=0.025). However, since two men were assayed at the hot spot, the probability of seeing a P-value of 0.025 in one or the other man was 2*0.025, that is 0.05. Given this borderline significance, it seemed most likely that the observed TD was a sampling artefact, as discussed in the following section.

3.3. Discussion

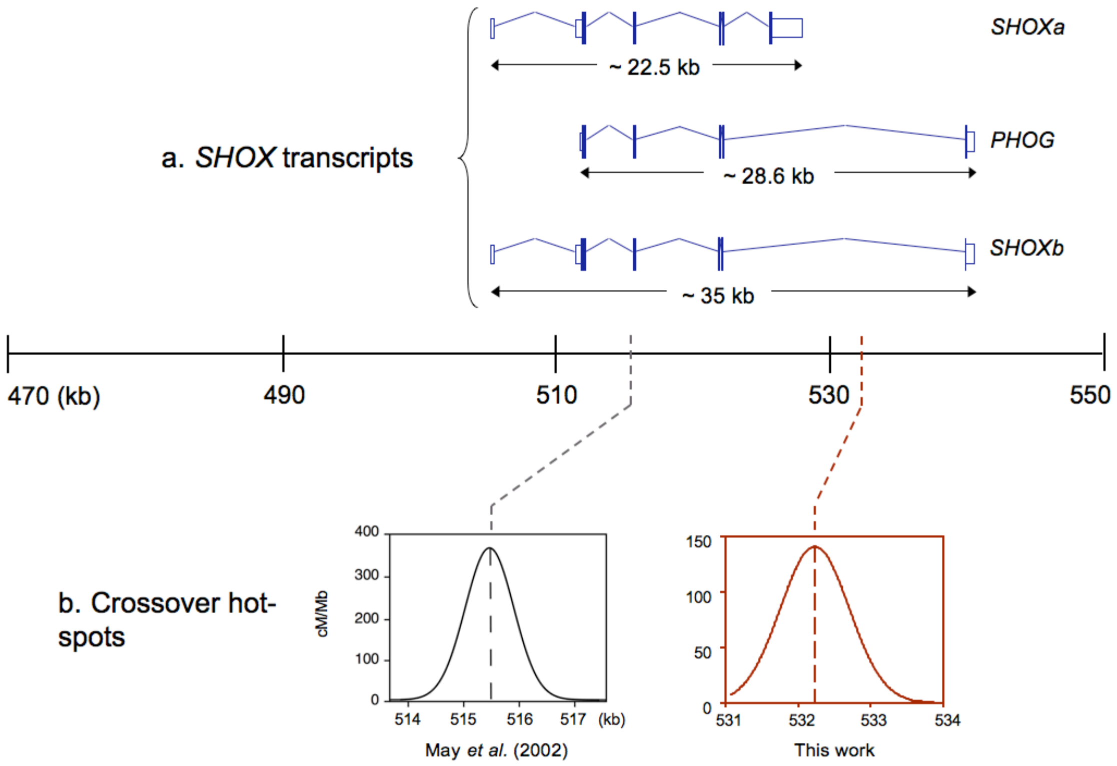
This study confirms the presence of an active sperm CO hot spot, ~18 kb proximal to the previously characterised *SHOX* hot spot (Figure 3.12). This hot spot, named SHOX-5, also falls within the *SHOX* gene, but lies in an intron (see Figure 3.12). The SHOX-5 hot spot is less GC-rich compared with the *SHOX* hot spot (43% and 55.2% GC-content respectively) but both hot spots lie within regions considerably more GC-rich compared to the average for the X-chromosome (~39.5% ; Blaschke & Rappold, 2006).

In terms of morphology, this newly characterised hot spot was slightly wider than the *SHOX* and previously characterised autosomal hot spots. Thus, average width of the hot spot at 2.2 kb is at least 300 bp wider than the 1-1.9 kb-range seen at other hot spots (discussed in Chapter 1). The distribution of events, roughly approximating a normal distribution, is nonetheless similar to previously identified hot spots and is consistent with the progressive decay of recombination activity on either side of a central initiation zone (see Figure 3.12). Thus, in addition to the previously characterised *SHOX* hot spot, data from this hot spot further suggest that similar underlying recombination processes operate in PAR1 and on autosomes, despite the need for an obligatory CO in PAR1 in male meiosis and the possible existence of distinct genetic control of recombination in this region (Burgoyne, 1982; Kauppi *et al.*, 2011).

Identification of a second active PAR1 sperm CO hot spot from analysis of patterns of haplotype diversity stresses the general effectiveness of such studies in hot spot identification in PAR1. It is worth noting however, that on account of the rapid breakdown of LD and high ‘background’ recombination activity in PAR1, detection of putative hot spots from population diversity data in PAR1 is more challenging than on autosomes (May *et al.*, 2002). Usefulness of PAR1 population data and their relationship to sperm data is discussed further in Chapter 4.

This study sheds light on additional phenomena that have previously not been identified at a pseudoautosomal sperm CO hot spot, including polymorphism in CO frequencies amongst men assayed at the hot spot. These features and their implications are discussed further in the following paragraphs. Finally, identification of this hot spot, in close proximity to the *SHOX* hot spot, has important implications for the overall fine-scale recombination landscape in PAR1 and is addressed in greater detail in section 3.3.3.

Figure 3.12 Genomic contexts and relative locations of PAR1 hot spots characterised to date



a. Relative positioning of the three known transcripts for the *SHOX* gene, namely, *SHOX*a, *SHOX*b and *PHOG* (Rao *et al*., 1997; Ellison *et al*., 1997). Data for all three transcripts were obtained from the Ensembl genome browser (http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000185960;r=X:585079-620146). Exons are indicated using filled bars and untranslated regions are in white.
b. The normal distribution of COs at the *SHOX* hot spot (black) and at the hot spot characterised presently (red). Dotted lines indicate the centre-points of the two hot spots. All co-ordinates are as per NCBI36/hg18.

3.3.1. Variation in CO activity between d10 and d62

Interestingly, CO frequency at the hot spot differs ~7-fold between the two men studied, at 0.041% (d62) and 0.29% (d10). Polymorphism in CO activity, to the extent of complete on/off polymorphism, has been well documented on autosomes (Jeffreys & Neumann, 2002; 2005; 2009; Jeffreys *et al.*, 2005; Neumann & Jeffreys, 2006; Webb, Berg & Jeffreys, 2007) but has not, until now, been reported within a pseudoautosomal region.

Data from autosomes show that in some hot spots, including *DNA2* in the MHC and *NID1* on chr1, polymorphisms in CO activity have been associated with specific DNA sequence variants close to the hot spot centre (Jeffreys & Neumann, 2002, 2005; discussed further in section 3.3.2). In other instances however, polymorphisms in CO frequencies appear to be independent of DNA variants in and around the hot spot (Neumann & Jeffreys, 2006). The best example of this is variation at MSTM1a and MSTM1b hot spots, located ~2 kb-apart on chr1 (Neumann & Jeffreys, 2006).

The MSTM1b hot spot was active in all 26 men assayed at the hot spot, but showed ~75-fold variation in sperm CO frequencies, ranging between $1.2x10^{-5}$ % and $90x10^{-5}$ %. Interestingly, analysis of haplotypes in and around the hot spot, showed instances of men showing very high and low CO frequencies, yet sharing identical haplotypes over 15 kb of sequence spanning the hot spot. This implied that variation in activity at MSTM1b was independent of local DNA sequence variation. MSTM1a, on the other hand, was only active in 3 of 26 men assayed at the hot spot and was thus an example of 'on/off' hot spot polymorphism. As at MSTM1b, analysis of active and suppressed MSTM1a haplotypes over a 100 kb-long interval, suggested that presence/ absence of the hot spot was unlikely to be determined by local DNA variants.

The unavailability of a large number of additional, more informative men that could be assayed at the SHOX-5 hot spot meant that the true extent of variation in CO activity could not be analysed further. Also, the two men assayed carried substantially different haplotypes, making it impossible to associate any local DNA sequence variants with reduced/enhanced CO activity at the hot spot or indeed discount the possible existence of such factors.

Furthermore, recombination activity for these men was not studied at either the *SHOX* hot spot or any of the flanking intervals also analysed using sperm assays, as they were not heterozygous for chosen selector sites. It is therefore, also not certain if reduced activity at this hot spot in d62 is compensated for by higher activity within nearby intervals (as indicated by Lien *et al.*, 2000). However, in light of identification of factors that influence hot spot usage and activity at human CO hot spots subsequent to this study, the issue of variation in activity will be re-visited in the overall discussion of this thesis (Chapter 8).

3.3.2. TD accompanying COs in d10

TD accompanying COs has been observed at several autosomal hot spots (Jeffreys & Neumann, 2002; 2005; 2009; Neumann & Jeffreys, 2006, Webb, Berg & Jeffreys, 2007). At hot spots *DNA2* and *NID1*, TD can be traced back to heterozygosity at a single SNP close to the hot spot centre. Thus, at hot spot *DNA2*, significant TD was observed in men heterozygous for an A/G polymorphism 70 bp away from the hot spot centre, with over-transmission of the 'A' allele. A/A and G/G homozygous men on the other hand, showed either very weak or no distortion. Furthermore, two G/G homozygous men displayed the lowest CO frequencies observed at the hot spot, a significant observation that supports a role for the SNP in directly regulating hot spot activity.

Jeffreys & Neumann (2002) suggested that this phenomenon could be explained using a model wherein DSBs are initiated more frequently on one haplotype compared to the other (in the case of *DNA2* this would refer to the haplotype carrying the 'A' allele). These DSBs would then be repaired (following 5′ to 3′ resection, strand-invasion, and dHJ formation) using information from the unbroken homologous chromosome, harbouring the suppressed haplotype, thus resulting in over-transmission of the latter into recombinant progeny. This biased gene conversion in favour of recombination-suppressing alleles constitutes a form of meiotic drive at the population level, that can nonetheless influence fixation behaviour of hot spot alleles (discussed further in Chapter 6).

It is important to note that such hot spot phenomena shed valuable light on the likely time-scales of hot spot evolution in human populations. This is perhaps best demonstrated at autosomal hot spot S2, located near the end of the short arm of chr3 (Jeffreys & Neumann, 2009). As with *DNA2*, hot spot S2 shows strong reciprocal CO asymmetry and TD that is exclusively associated with heterozygosity at a G/A polymorphism located ~60 bp away from the hot spot centre. Interestingly, the hot spot was only active in men carrying the 'G' allele at this SNP site and recombination was suppressed ~100-fold in A/A homozygotes. Analysis of hot spot haplotypes in 94 semen donors indicated that the hot spot-activating allele arose in human populations through a recent mutation event, ~70,000 years ago. The hot spot activating-allele was nonetheless under-transmitted to only 26% of recombinant progeny, suggesting that ~0.04% of 'G' alleles were lost from the population per generation. This recombination-based meiotic drive is sufficient to block the spread of this hot spot-activating allele in the population.

Thus, data from the S2 hot spot indicates that this active sperm CO hot spot was doomed to extinction right from the outset. In fact, data from extensive simulations are consistent with a lifespan of a mere 120,000 years for this hot spot. These time-scales of hot spot evolution further support the idea that hot spots represent rapidly evolving, transient features of the human genome (see Chapter 1).

Significant TD of hot spot alleles has so far not been observed in PAR1. TD observed at SNP 35/2637 in d10 was only on borderline significance (P=0.05) and may be a sampling effect. This is also supported by the fact that a more central marker 35/2287 did not show any evidence of TD (P=0.306). Further, even if genuine, the degree of TD of hot spot alleles was only modest (55:45) compared to other autosomal hot spots, including *NID1* (74:26) and S2 (62:38) (Jeffreys & Neumann, 2005; 2009). Owing to the presence of time constraints, transmission of 35/2637 alleles in other men heterozygous for the marker could not be further studied and the phenomenon verified.

It should also be mentioned that only three additional North European men heterozygous at 35/2637 carried suitable selector sites for sperm CO assays and none of these men were particularly informative over the hot spot centre. One possible way of further addressing this issue would thus be to investigate sperm CO activity at the hot spot in men from other, more diverse populations, who may carry additional polymorphisms closer to the hot spot centre. A panel of African semen donors is in fact available locally and has been studied as part of several other projects. Unfortunately, owing to time constraints, this panel could not be studied to further investigate the phenomenon of TD at this PAR1 hot spot. If validated, this phenomenon might shed valuable light on likely mechanisms and time-scales of hot spot evolution in PAR1 and facilitate meaningful comparisons with autosomes.

3.3.3. Implications of identification of this hot spot on the overall recombination landscape in PAR1

Regardless of variation amongst men, the average recombination activity at this hot spot (40 cM/Mb) is 45.5 times higher than the genome average for male meiosis (0.89 cM/Mb (Gyapay *et al.,* 1994). Furthermore, overall recombination activity over the 29 kb region analysed as part of this survey is 3.79 cM/Mb, which is ~4-fold higher than estimates from similar surveys in the class II region of the MHC and on chr1 (C.A. May, unpublished data; Jeffreys, Kauppi and Neumann, 2001; Jeffreys *et al*., 2005). It is worth mentioning that the autosomal survey extended over much larger intervals, but this alone is unlikely to cause the disparity in overall recombination activity between autosomes and PAR1. Rather, it is possible that this is the consequence of substantially higher CO frequencies at hot spots and elevated inter-hot spot recombination rates in PAR1, which may have evolved as a result of the requirement for at least one CO in this genomic region for faithful chromosome segregation during meiosis I.

Finally, the two PAR1 hot spots identified through this survey are ~18 kb apart. Based on this, hot spot spacing in PAR1 appears much denser than estimates from autosomes, which suggest that hot spots are located, on average, 50 kb apart (Myers *et al.,* 2005; Jeffreys, Kauppi & Neumann, 2001). It should be noted that autosomal ‘double hot spots’ and ‘hot spot clusters' do exist. Indeed, several double hot spots located within a couple of kb from one another have been identified on autosomes, including hot spots MSTM1a and MSTM1b mentioned above (Neuman & Jeffreys, 2006; Webb, Berg & Jeffreys, 2007; Jeffreys & Neumann, 2009). Likewise, hot spot clusters extending over ~7 kb have also been well identified in the MHC class II and chr1 surveys described in section 3.1.3. It is however unlikely that the two PAR1 hot spots constitute a double hot spot or form part of a cluster of hot spots, since these features are rarely separated by distances as great as 18 kb.

In summary, data emerging from the extended *SHOX* survey suggests that whilst similar underlying recombination processes likely operate throughout the human genome, the extraordinary evolutionary history of PAR1 and the need for an obligatory CO event in male meiosis modifies the overall recombination profile in the region, as exemplified by the presence of closely-situated active sperm CO hotspots and substantially elevated recombination activity outside of these hot spots (C.A. May, unpublished data).

Chapter 4. Search for the most highly active Xp/Yp pseudoautosomal hot spots

4.1. Introduction

High-resolution analysis of recombination in PAR1 is largely limited to data from the *SHOX* survey and as such little is known about the fine-scale distribution of recombination in the rest of this genomic region. Data from May *et al.* (2002) and from the work described in Chapter 3 both indicate that patterns of haplotype diversity are generally good indicators of contemporary PAR1 recombination rates, with both hot spots characterised so far present within intervals of LD breakdown and high-inferred historical activity. With the advent of projects like HapMap (The International HapMap Consortium, 2004, 2007), it has become possible to investigate population diversity data across PAR1 in more detail, in order to identify the most active PAR1 recombination hot spots.

Indeed, an extensive survey of patterns of allelic association on autosomes has led to the identification of a set of highly active autosomal hot spots (Webb, Berg & Jeffreys, 2008). In addition to providing important insights into the characteristics of these intense hot spots, this study has also shed some valuable light on the usefulness and limitations of utilising patterns of allelic association as means of identifying intense human recombination hot spots.

4.1.1. Identification of highly active autosomal sperm CO hot spots using patterns of marker association

As discussed previously in Chapter 1, recombination is one of the main forces that causes breakdown of allelic association. Very active hot spots should thus, in theory, exist within intervals that show the most extreme breakdown of LD. In order to identify such intervals and ultimately the most active human hot spots, Webb and colleagues (2008) created LDU maps for all autosomes, using Phase II HapMap genotype data on ~2.4 million SNPs for 60 individuals from a Utah population with northern and western European ancestry (CEU dataset). LDU maps thus generated showed good correlations with linkage maps and with estimates of historical recombination rates inferred from coalescent approaches, at least on a Mb-scale (Kong *et al.,* 2002; McVean *et al.,* 2004).

From this analysis, Webb and colleagues (2008) identified fifteen, 5 kb-long intervals containing a very large LDU step (>3 units in size) and showing high levels of inferred historical activity. All of these steps were verified at higher resolution in a panel of 92 North European semen donors, although step sizes were significantly reduced in several cases. Despite this, sperm CO assays on three men per LDU step showed that all LDU steps represented genuine sperm CO hot spots, whose activities were significantly higher than previously identified autosomal counterparts (P=0.0005). Interestingly, the hot spots identified varied widely in activity, ranging between 0.015% and 1% (median 0.13%), indicating that whilst surveys of LD are an effective way of identifying CO hot spots, they are poor at predicting true hot spot activities.

This autosomal survey not only more than doubled the repertoire of known sperm CO hot spots, but also highlighted several interesting features of these actively recombining regions. For example, although several hot spots showed polymorphism in activity, this variation was mostly restricted to weaker hot spots (frequency <0.1%). This was curious since higher recombination activity should facilitate faster spread of hot spot suppressing mutations through over-transmission to the recombinant progeny (as described for the *DNA2* hot spot in Chapter 3). Thus, more active hot spots, rather than less active ones, should show greater variation in activity.

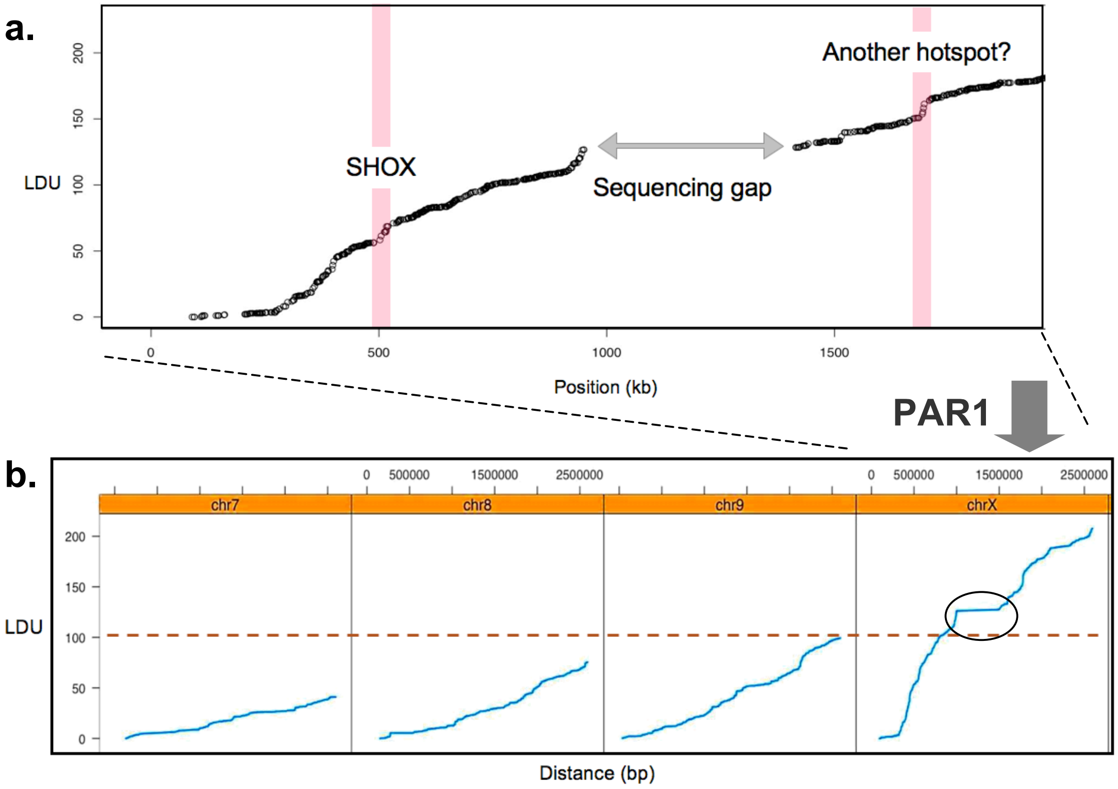
In addition, active sperm hot spots such as ones identified by Webb and colleagues (2008) allow various aspects of the dynamics of recombination to be studied in unprecedented detail via providing access to hundreds or even thousands of recombinants. Very active hot spots, with a good density of informative markers, are a resource for studying GCs and for investigating the shifting balance between frequencies of COs and NCOs. They can enable better definition of hotspot morphology and allow targeting of rare events, which may shed valuable light on recombination processes. Also, if two such hot spots were in close proximity, they could be used to analyse the phenomenon of CO interference at the molecular level. Finally and perhaps most importantly, a collection of such hot spots would be essential for the direct isolation of recombination intermediates from testes, which in turn will enable direct analysis of recombination initiation and processing of recombination intermediates in human meiotic cells.

4.1.2. This work

The need for an obligatory CO in PAR1 during male meiosis, the significantly higher than genome-average recombination activity obtained from linkage analysis, the extreme breakdown of LD observed in and around the *SHOX* gene and the identification of two fairly active recombination hot spots just from this limited survey, all suggest that this region of the genome is very active recombinationally and that most of this activity can probably be attributed to intense recombination hot spots. Given the presence of highly active hot spots (CO frequencies up to 1%) in rather unremarkable parts of the genome (Webb, Berg and Jeffreys, 2008), it seemed likely that equally or even more active hot spots exist in PAR1. If identified, these hot spots can serve as a valuable resource for investigating the dynamics and processes of recombination in this unusual genomic region and allow further comparisons between autosomal and PAR1 CO hot spots.

Data from the autosomal survey discussed in section 4.1.1 showed that analyses of patterns of population diversity are an effective means of locating intense recombination hot spots. Availability of HapMap population diversity data across PAR1 made it possible to use the same strategy for the identification and characterisation of the most intense pseudoautosomal hot spots. LDU mapping over PAR1 using HapMap phase II genotyping data for 60 CEU individuals (The International HapMap Consortium, 2007), revealed extreme breakdown of marker association spanning the entire region (A.J. Webb and C.A. May, unpublished data). This extent of LD breakdown was significantly greater than that seen on equivalent genomic regions on autosomes (Figure 4.1) and was thus a likely consequence of the evolutionary history of PAR1 and its role in male meiosis. Identifying specific LDU steps that could be representative of putative hot spots against the backdrop of general breakdown of association posed a considerable challenge. Nonetheless, one candidate interval, containing the strongest identifiable LDU step, was selected for further analysis. It is worth noting that at the time of this analysis, gaps in PAR1 primary sequence data still existed and additional targets may well exist within these regions.

Figure 4.1 LDU profile across PAR1 and comparisons with autosomes



a. LDU profile across PAR1 was obtained from Phase II HapMap genotypes (A.J. Webb and C.A. May, unpublished data). Extensive LD breakdown in PAR1 is evident from the steady increase in LDU across this region. Indeed, even the interval encompassing the active *SHOX* hot spot (highlighted in pink) appeared as a minor increase in metric LDU. Only one other step (also highlighted in pink), comparable to the *SHOX* hot spot, could be identified. This was chosen for further analysis. The grey arrowed bar represents gaps in genotype data (as of September, 2007).

b. LDU profile across PAR1 was compared with equivalent genomic regions on other metacentric chromosomes of similar size and morphology as the X chromosome. This comparison highlighted the substantially greater breakdown of marker association in PAR1 compared to the autosomes. The circled region represents gaps in PAR1 primary sequence data.

This chapter describes analysis of the interval thus identified, first by generating higher density SNP haplotypes for a panel of 92 North European semen donors and then through using high-resolution sperm analysis for detailed hot spot characterisation.

4.2. Results

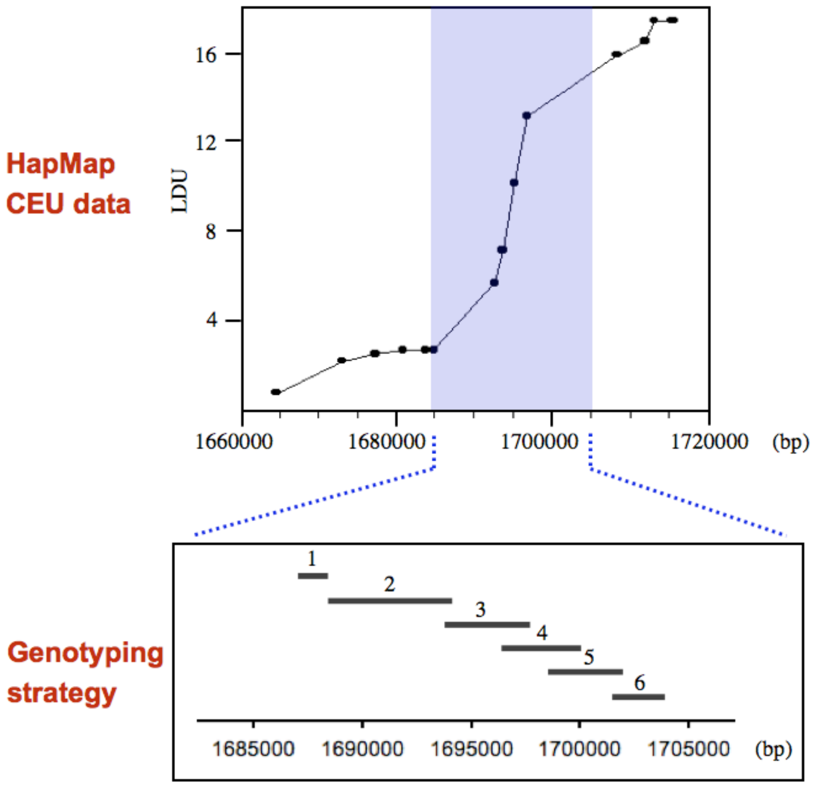
LDU analysis across PAR1 showed evidence for a large LDU step (of ~ 15 units) in the interval between 1685 and 1712 kb (NCBI36/hg18). Unlike autosomal LDU steps identified using the same strategy, this region spanned an unusually long interval of ~27 kb. Further higher-resolution analysis was thus limited to an ~17 kb-wide interval, flanking the centre of the LDU step, assuming that this represented the centre of the putative historical hot spot (Figure 4.2).

Prior to further studies, the genomic context of the LDU step was investigated. Thus, the DNA sequence was downloaded from the UCSC Genome Browser, SNPs were identified using dbSNP and grouped according to validation status and repeat DNA was identified using the 'Repeat Masker' function in the UCSC Genome Browser (see Appendix 4.1 for the annotated sequence information). This analysis showed a very high density of repetitive DNA (mostly Alus) in this region. This was not ideal, since it highly limited regions wherein primers could be positioned (primers in repeats, particularly Alus, are often non-specific) and substantially influenced the overall experimental design. Nonetheless, this interval was next subjected to LD analysis in a panel of 92 North European semen donors, in order to investigate if the large LDU step could be replicated in this panel and if so, to characterise it at a higher resolution.

4.2.1. High-resolution LD analysis in a panel of North European semen donors

For purposes of LD analysis, the ~17 kb interval selected for study was split into six overlapping amplicons (see Figure 4.2), varying in length between 1.4 kb bp and 5.8 kb. Positions of the amplicons largely depended on presence of single-copy DNA where primers could be designed. Overall sizes also had to be limited to less than 6 kb, in order to facilitate amplification from whole-genome amplified (MDA) DNA, which is routinely used for population screens. All primers were optimised using annealing and extension temperature titrations, prior to use. Nested PCRs were then used to amplify all six intervals in a panel of 92 North European semen donors. Primer combinations and PCR conditions used for amplification are listed in Appendix 4.2.

Figure 4.2 Strategy for investigating an ~18 kb interval of LD breakdown in a panel of North European semen donors



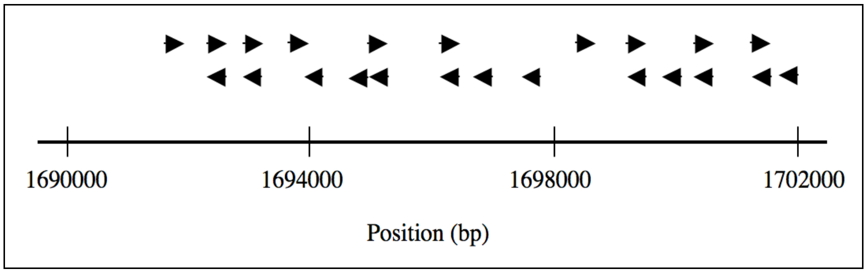
The top graph shows the largest step in metric LDU identified in PAR1 from analysis of HapMap CEU phase II genotype data. The interval chosen for further analysis is highlighted. Below, relative positions of genotyping amplicons are indicated using horizontal bars. Numbers above each bar relates to the local name given to the amplicon. All positions are as per NCBI36/ hg18.

PCR products thus generated were analysed using agarose gel electrophoresis to ensure that the product obtained was specific, of the required size, and had amplified to the same extent in all samples. Following this, the PCR products were dot blotted and SNPs within each amplicon genotyped using ASO hybridisation. SNP status was determined by sequentially hybridising dot blot filters with ASOs specific for one or the other SNP allele.

It is worth noting that a large number of reported SNPs in this region were present within Alus and other repeat structures. If the sequence context of a SNP is repeated elsewhere within the same amplicon, then ASOs designed to type the SNP bind to several positions within that amplicon, giving ambiguous results. In order to avoid such ectopic ASO hybridisation, smaller amplicons (of >1 kb) were used to type SNPs located within repetitive DNA. Details of these smaller amplicons, with primers and PCR conditions used, are listed in Appendix 4.2.

In order to identify additional previously unreported SNPs, preferably within single-copy DNA, ~10 kb of the region of interest (prioritised on the basis of relatively poor SNP density) was re-sequenced in eight North European semen donors, as illustrated in Figure 4.3. This led to the identification of several previously unreported polymorphisms, six of which had a minor allele frequency of greater than 0.1.

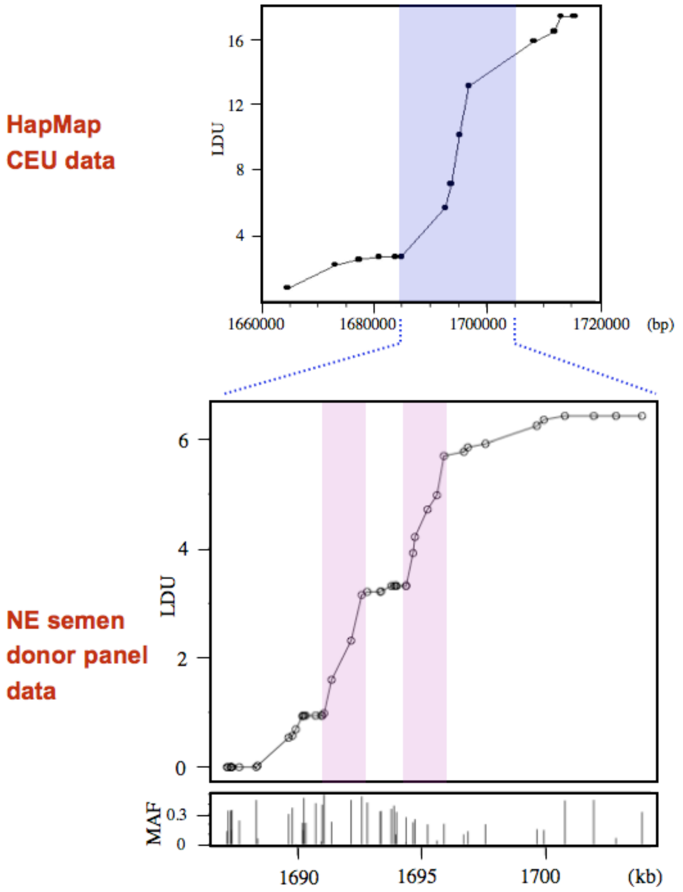
Figure 4.3 Re-sequencing strategy used to identify additional polymorphisms across the study interval



Re-sequencing efforts were concentrated over regions with lower density of polymorphisms. The chosen intervals were amplified from genomic DNA from eight semen donors, using genotyping amplicons 2-5 (see Figure 4.2). The protocol used for re-sequencing is as described in Chapter 2. Owing to the high concentration of poly A/Ts, sequence data had to be compiled from a number of short reads, shown by arrows, with direction of the arrow indicating whether forward or reverse sequencing primers were used to generate each read. Details of primers used in sequencing reactions are listed in Appendix 4.2. SNPs identified through re-sequencing are indicated in purple in the annotated sequence (see Appendix 4.1) and can be distinguished from others by their names, which include the word ‘seq’ at the end.

In total, 46 SNPs were genotyped across approximately ~17 kb and resulting patterns of marker association analysed using LDU mapping. An online program based on that described by Collins and Morton (1998) was used to calculate LD units (available at http://143.210.151.197/cgi-bin/LDMAP/ldumapping.pl). Details of types of polymorphisms typed, sequences of ASOs used and genotypes of the 92 semen donors are listed in Appendix 4.3. Figure 4.4 compares LDU profiles obtained from genotyping in the North European semen donor panel, with that previously obtained from analysis of HapMap CEU genotypes.

Figure 4.4 A higher-resolution LDU profile across the study interval



The top graph shows the LDU profile obtained from HapMap phase II (CEU) genotype data, whilst the one below shows the LDU profile (across the region highlighted in blue) obtained from genotyping 92 North European (NE) semen donors. Higher resolution data from the North European panel shows that this interval contains two smaller LDU steps (highlighted in pink), located ~ 2 kb apart. Minor allele frequencies of SNPs genotyped in the North European semen donor panel are also shown below the LDU profile for the dataset.

Whilst both maps are consistent with presence of an LDU step in this region, there are two significant differences between them. First, the size of the LDU step in North European semen donors is substantially smaller than the one seen in the HapMap dataset. This has been previously noted on autosomes and may in part, be due to the higher marker density of the North European semen donor dataset (Webb, Berg & Jeffreys, 2008). Large distances between polymorphic markers often create 'holes' in the LD map, wherein markers appear to be in free association; increasing marker density can reduce such inflations and provides a more accurate estimate of LD.

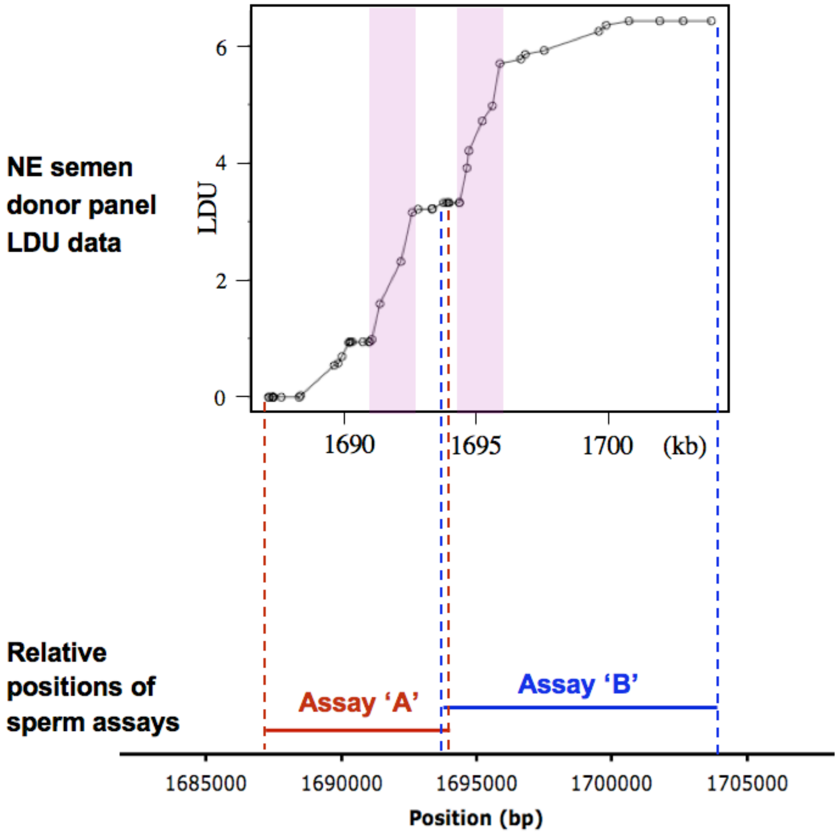
Second, at higher resolution the wider LDU step reported in HapMap, appeared to resolve into two distinct steps separated by a small region (of ~ 2kb) wherein markers were in association. Each of these steps were about 2 kb wide, reminiscent of the width of classical recombination hot spots. Furthermore, both of these LDU steps were captured in full, with markers on either side in association with each other (represented by plateaus on the map, see Figure 4.4).

Patterns of association seen in the North European semen donor panel, with two LDU steps within 2 kb of each other, are more consistent with presence of a putative double hot spot in this region. Double hot spots in such close proximity have previously been identified on autosomes (eg: Neumann & Jeffreys, 2006; Jeffreys & Neumann, 2009), but not yet in a pseudoautosomal region. In order to investigate contemporary recombination activities and distribution of COs in this region, sperm CO assays were next designed encompassing the interval studied.

For purposes of these high-resolution assays, the 17 kb region was split into two overlapping intervals, each harbouring one LDU step (see Figure 4.5). This was largely done in order to maximise the ease and specificity of amplification from sperm DNA in this repeat-rich region of the genome. It is worth mentioning that a few attempts were made to optimise ~15 kb-long PCRs that would encompass the entire interval, but these PCRs consistently yielded multiple non-specific products.

Although additives such as betaine and DMSO (dimethyl sulphoxide) can improve PCR specificity in GC-rich, repeat-dense genomic regions through minimising secondary structures, attempts at optimising PCRs using these reagents were not made. This was primarily owing to time constraints and in view of the fact that dividing the region into two smaller intervals would still address questions pertaining to the presence of a double sperm CO hot spot and allow their characterisation just as well as one larger assay would.

Figure 4.5 Strategy for sperm CO assays encompassing PAR1 LDU steps



The figure shows relative positions of sperm CO assays designed across each LDU step. It is worth noting that strategy for assay 'B' had to be subsequently modified such that it extended to ~1700 kb, rather than 1703 kb, as discussed in section 4.2.2.2.

4.2.2. High-resolution sperm analysis of the proximal LDU step (assay 'B')

The region encompassing the more proximal LDU step was studied first using sperm CO assays. Various stages of this assay are described in detail below.

4.2.2.1. Identifying selector sites and optimising allele-specific primers (ASPs)

First, SNPs rs4446909A/G, rs5989681C/G, 3920seqG/A and rs6644635C/T were chosen as potential 3' selector sites and 18 nucleotide-long ASPs designed, with the SNP positioned at the last 3′ base. All primers were tested for optimal efficiency and specificity as described in Chapter 3 (section 3.2.2). ASPs designed towards SNPs rs4446909A/G and rs5989681C/G provided the best yields and were most specific, hence these were chosen as primary and secondary 3′ selector sites respectively.

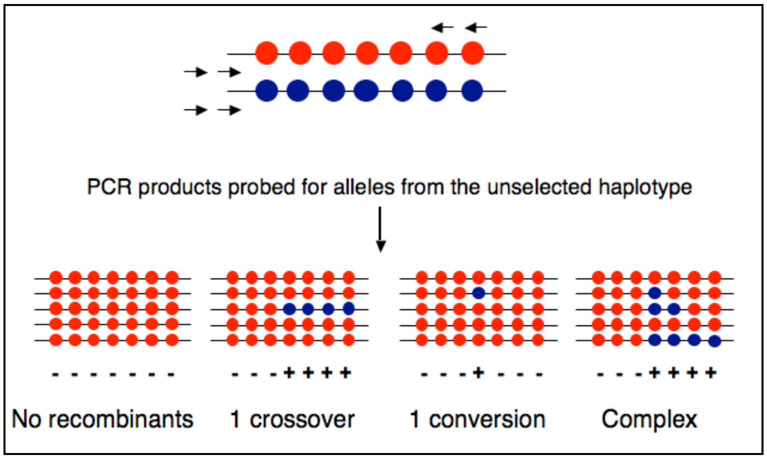
Choice of 5' selector sites for this assay was very limited owing to several factors. First, the region of marker association separating the two putative hot spots was rather small, highly restricting the choice of selector 5' selector sites. Second, several SNPs within this small region were present within repetitive DNA sequences, which adversely affected specificity of ASPs. This further limited the choice of selector sites for sperm CO assays.

Only two SNPs, rs6588802C/T and rs28541584A/C**,** 5′ of the LDU step met the criteria for potential selector sites, that is, they were present in single-copy DNA, within a suitably GC-rich sequence context and had high minor allele frequencies. However, attempts at optimising ASPs targeted towards these sites showed that the primer designed towards rs28541584A/C was not specific enough for use in repulsion-phase allele-specific PCR. Furthermore, optimal annealing temperatures of ~66°C for the reverse ASP RAS8802C/T, designed using rs6588802C/T as a selector site, were much higher than those for either forward ASP (54˚C and 64˚C for primary and secondary forward ASPs respectively). In view of this, a different type of sperm CO assay, compared with the one described in Chapter 3, was used to study this region, as described below.

4.2.2.2. Modified strategy for sperm CO analysis over the proximal LDU step

In this assay, parental haplotypes are amplified separately (typically from pools of 25 sperm molecules) using two rounds of nested PCRs, with ASPs in combination with universal primers. Haplotypes thus obtained are re-amplified using universal primers and the products dot blotted. Dot blot filters for any one haplotype are then probed for alleles belonging to the opposite haplotype in order to detect COs as well as NCOs, as shown below.

Figure 4.6 Detecting COs and NCOs using a modified sperm CO assay



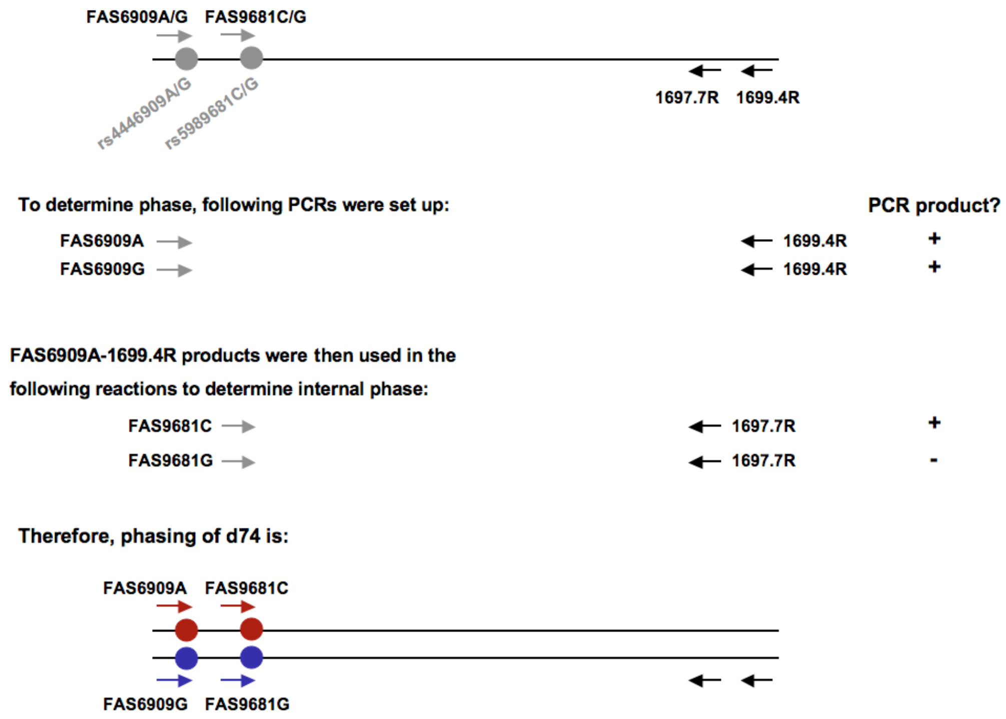
Detecting recombinants from sperm assays using forward ASPs in combination with reverse universal primers. Red and blue circles represent informative SNPs across the assay interval. Probing dot blot filters containing PCR products of any one haplotype, in this case the red haplotype, with alleles belonging to the unselected (blue) haplotype can yield a variety of results. A lack of any positive signals across the assay interval implies a lack of detectable recombinants. A single marker from the unselected haplotype can be interpreted as a NCO, whereas multiple consecutive markers from the unselected haplotype indicate a CO. It is worth noting that in rare cases pools may be complex and contain both types of events. In the example shown in the figure, a single-site and a double-site NCO are both masked by a CO and thus cannot be detected.

Since a pair of 3' selector sites were already identified and ASPs optimised, semen donors heterozygous for these sites were identified next. Amongst these men, ones with the highest density of informative markers across the assay interval were short-listed. Finally, two of these men, donors 74 and 30 (abbreviated as d74 and d30), for whom sperm DNA was available, were chosen for analysis.

4.2.2.3. Determining linkage phase of semen donors chosen for study

Linkage phase of primary and secondary selector sites was determined for both semen donors. Phasing of d30 is illustrated in Figure 4.7; d74 was phased using the same strategy, but using different universal reverse primers. PCR conditions for all phasing reactions are listed in Appendix 4.4.

Figure 4.7 Phasing d30 for high-resolution sperm assays


FAS6909A/G
FAS9681C/G
rs4446909A/G
rs5989681C/G
1697.7R
1699.4R
To determine phase, following PCRs were set up:
PCR product?
FAS6909A
1699.4R
+
FAS6909G
1699.4R
+
FAS6909A-1699.4R products were then used in the following reactions to determine internal phase:
FAS9681C
1697.7R
+
FAS9681G
1697.7R
-
Therefore, phasing of d74 is:
FAS6909A
FAS9681C
FAS6909G
FAS9681G

Grey circles indicate selector sites and grey arrows show positions of corresponding ASPs. Universal reverse primers are indicated using black arrows. Linkage phase was determined using two rounds of nested PCRs and analysing products using agarose gel electrophoresis. FAS6909A-1700.01R products were diluted five-fold before seeding secondary PCRs. Note: Figure is not to scale.

Thus, gel electrophoretic analysis of primary and secondary PCR products for d30 showed that rs4446909A and rs5989681C alleles lay on the same haplotype, whilst ‘G’ alleles of both SNPs were present on the opposite haplotype. d74 had the same phasing at selector sites.

Since recombination events using this type of sperm assay are detected by probing PCR products of one haplotype with alleles specific to the opposite haplotype, linkage phase of internal SNPs needed to be determined as well. Thus, positive secondary PCRs generated whilst phasing selector sites, which represented separated parental haplotypes, were re-amplified using various universal primer combinations to increase yields and the products dot blotted.

Since multiple SNPs within the assay interval were present within non-unique DNA sequences (mostly Alus), four small tertiary PCRs (ranging in size between 264 bp and 1.8 kb) were generated for each semen donor. This was done to ensure that only one match to the ASO sequences for these SNPs were present within each amplicon.

The tertiary PCRs used to re-amplify separated haplotypes varied between the two donors, since they carried different informative SNPs. Primer combinations and PCR conditions used to generate these smaller amplicons are listed in Appendix 4.4. Dot blot filters containing these products were then sequentially probed first with one allele of an informative internal SNP followed the other to determine linkage phase, as shown below.

Figure 4.8 Determining linkage phase of internal SNP alleles in d30

A.

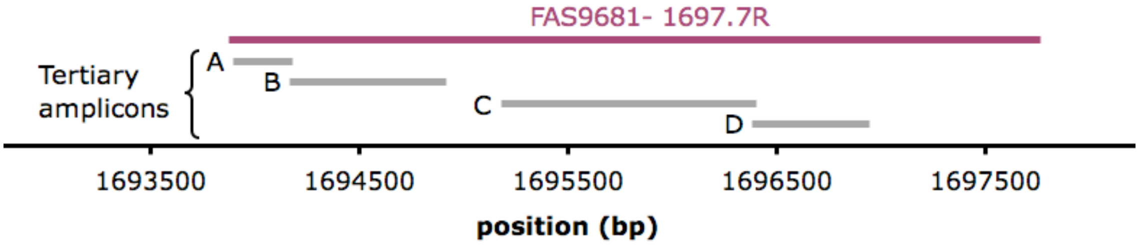
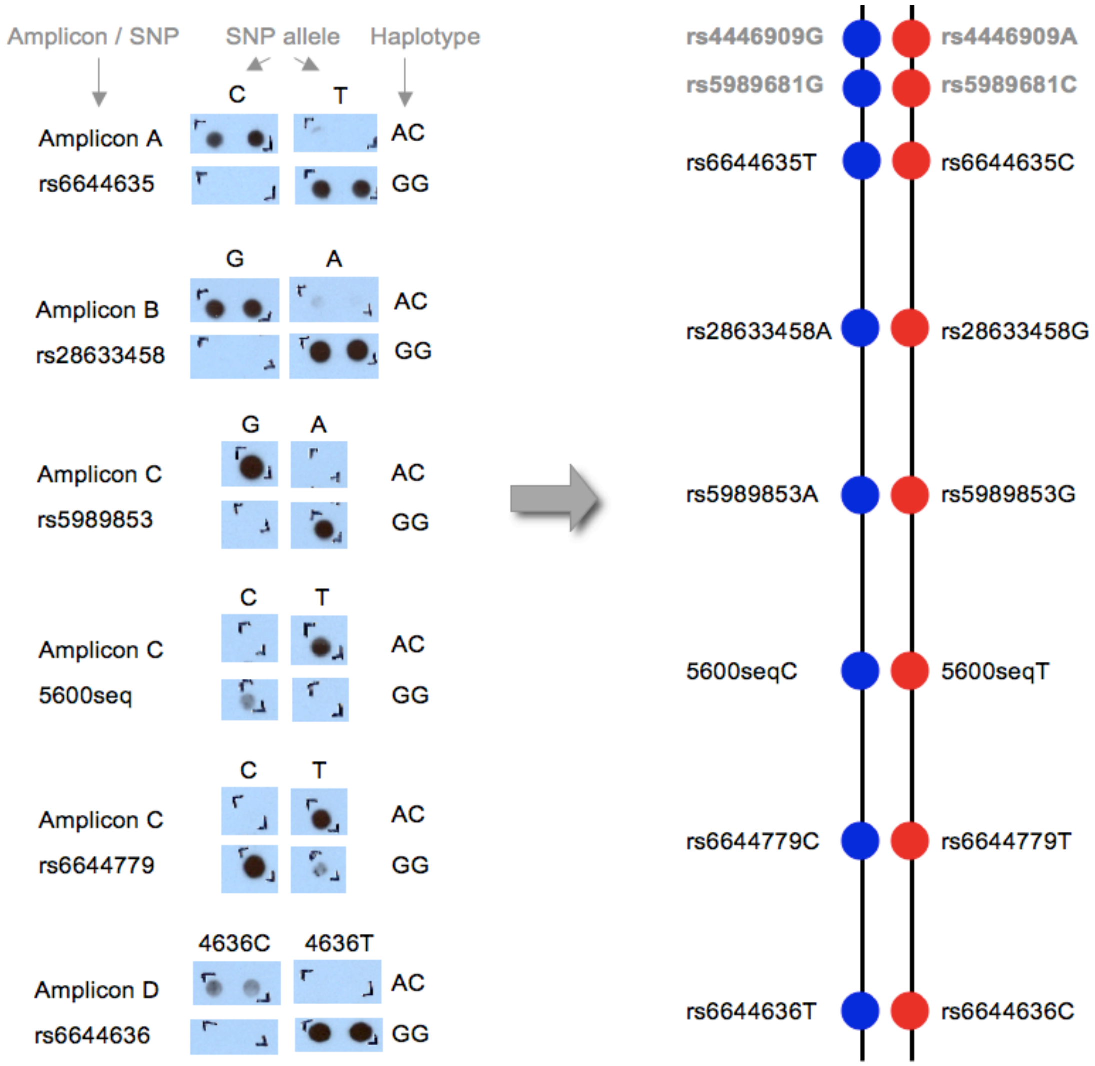


Figure 4.8 Determining linkage phase of internal SNP alleles in d30

B.


Amplicon / SNP
SNP allele
Haplotype
C
T
Amplicon A
rs6644635
AC
GG
G
A
Amplicon B
rs28633458
AC
GG
G
A
Amplicon C
rs5989853
AC
GG
C
T
Amplicon C
5600seq
AC
GG
C
T
Amplicon C
rs6644779
AC
GG
4636C
4636T
Amplicon D
rs6644636
AC
GG
rs4446909G
rs4446909A
rs5989681G
rs5989681C
rs6644635T
rs6644635C
rs28633458A
rs28633458G
rs5989853A
rs5989853G
5600seqC
5600seqT
rs6644779C
rs6644779T
rs6644636T
rs6644636C

A. Relative positions of tertiary amplicons used to amplify secondary phasing PCRs are shown in grey. The secondary amplicon used to seed these tertiary PCRs is also shown in pink.

B. Complete phasing data for d30. Parental haplotypes for d30 are referred to as ‘AC’ and ‘GG’ to denote that ‘A’ and ‘C’ alleles of selector sites rs4446909 and rs5989681 are present on the same haplotype, whilst ‘G’ alleles for both SNPs are present on the opposite haplotype. Both ‘AC’ and ‘GG’ haplotypes were re-amplified using smaller amplicons (see part A) to generate smaller tertiary products, which were subsequently dot blotted. In order to determine linkage phase of internal SNPs, dot blot filters containing both haplotypes were sequentially probed with each SNP allele. Any one SNP allele should ideally hybridise to one of the two parental haplotypes, thus allowing linkage phase to be determined. For example, rs664635 C allele only hybridises to haplotype ‘AC’ products, whilst rs6644635 T allele hybridises to haplotype ‘GG’ products. Thus, the ‘C’ allele of the SNP lies on the same haplotype as rs4446909A and rs5989681C and the ‘G’ allele on the haplotype containing rs4446909G and rs5989681G. Positions of SNPs are not to scale, but in order of position from 5' to 3'.

4.2.2.4. Bulk recovery of COs and NCOs from d30 and d74

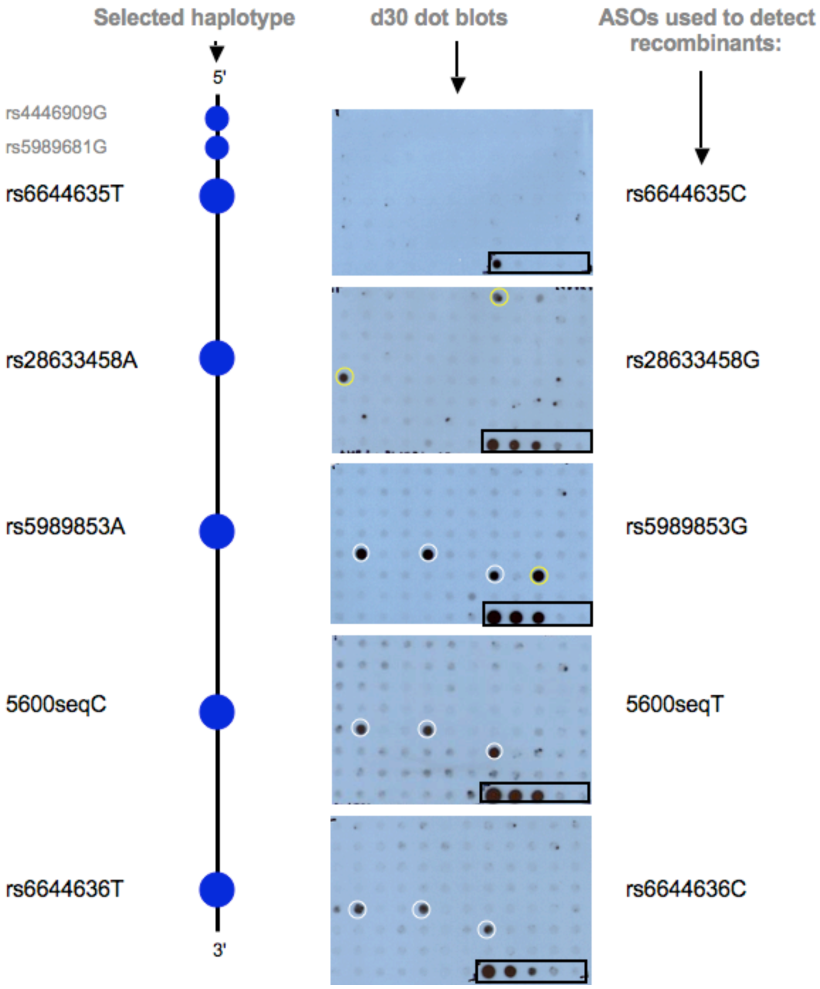
Sperm assays for the bulk recovery of recombination events were set up after both semen donors were fully phased across the assay interval. DNA inputs for this type of sperm assay are typically limited to ~50 molecules per pool (Kauppi, May & Jeffreys, 2009), in order to maximise non-selected allele signals during ASO hybridisation to dot blots. For d30, recombinants were recovered from both parental haplotypes using reciprocal assays, whereas only one orientation, that is haplotype, was screened for recombinants in d74. In total, 23,000 and 18,000 sperm molecules were screened for d30 and d74 respectively.

For both donors, the selected haplotype was amplified in 96-well PCR plates using two rounds of PCR, wherein forward ASPs were used in combination with universal reverse primers. The same primers and PCR conditions were used in primary and secondary PCRs as for phasing selector sites. Particular care was taken to minimise any risks of contamination and dedicated single molecule-clean pipettes and plastic-ware was used for sperm assays.

Secondary PCR products were re-amplified using universal primer combinations, in order to increase DNA yields for more efficient detection of recombinants. Tertiary PCR products obtained were electrophoresed on an agarose gel to verify product sizes, to ensure that ample PCR products were available for dot blotting and that amplification was uniform across all samples. Prior to dot blotting, the non-selected haplotype (amplified separately) was added to the last five wells of the 96-well PCR plate, such that the ratio of non-selected to selected haplotype products was 1:10, 1:30, 1:100, 1:300 and 1:1000. Using such a range of positive controls provides an effective means of estimating the proportion of DNA in a pool that accounts for the CO/NCO signal. Further, these controls provide an indication of the degree of allele specificity of various ASOs.

Dot blot filters containing PCR products of the selected haplotype were then probed with ASO alleles specific to the unselected haplotypes. Recombinants were detected and scored as described in Figure 4.9.

Figure 4.9 Detection of COs and NCOs in d30


Selected haplotype
d30 dot blots
ASOs used to detect recombinants:
5'
rs4446909G
rs5989681G
rs6644635T
rs28633458A
rs5989853A
5600seqC
rs6644636T
3'
rs6644635C
rs28633458G
rs5989853G
5600seqT
rs6644636C

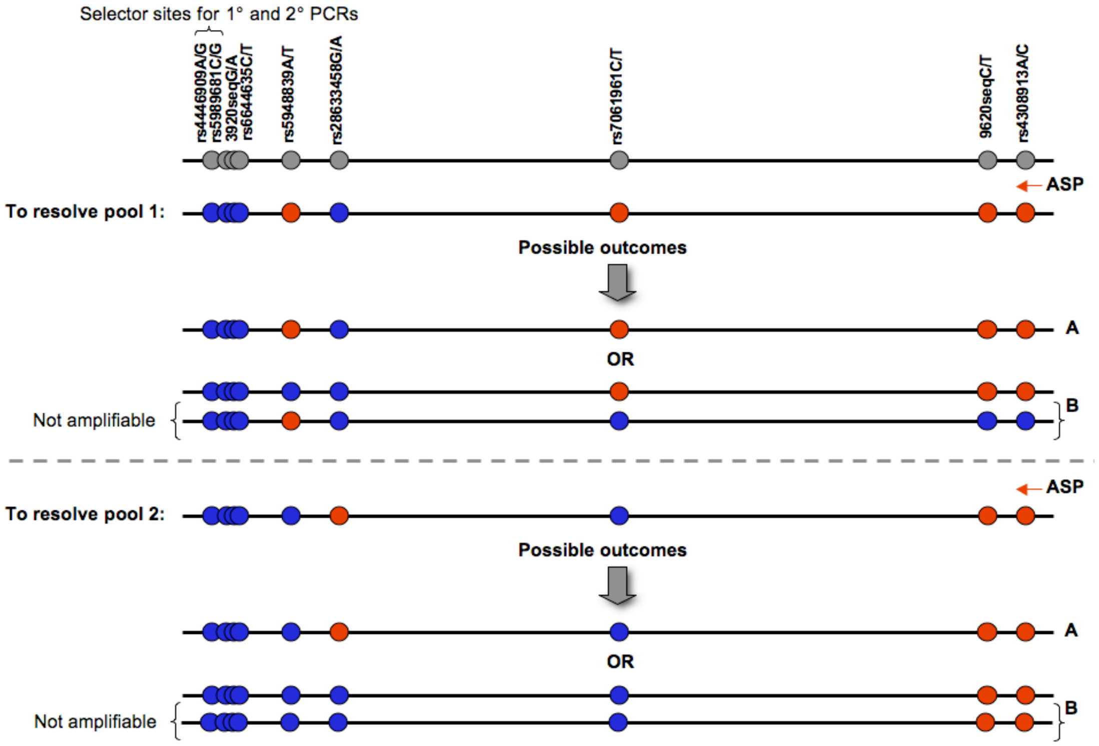
In the example shown, haplotype 'AC' was amplified in a 96-well PCR plate and the products were dot blotted. Last five reactions in the plate (bottom right) represent positive controls, containing a mixture of the non-selected and selected haplotypes in ratios of 1:10, 1:30, 1:100, 1:300 and 1:1000. The dot blot filter was sequentially probed for various SNP alleles specific to the unselected 'GG' haplotype. COs are indicated by presence of consecutive positive signals towards the distal end of the assay interval (affecting markers rs5989853, 5600seq and rs6644636) and are highlighted using white circles. NCOs are inferred where a single marker is positive for an allele specific to the unselected haplotype, with both flanking markers showing no positive signals. These are indicated using yellow circles.

Very faint positive signals, similar to positive controls containing the non-selected and selected haplotypes in ratios of 1:300 and 1:1000, which affected single markers were usually discounted from analysis, as these were likely to represent PCR mis-incorporations, which arose during early cycles of PCR. COs, which affected multiple markers and were thus genuine, acted as valuable internal controls for distinguishing recombinant signal intensities from those likely to represent PCR mis-incorporations. In total, 34 and 64 COs (uncorrected) were recovered for d30 and d74 respectively, whilst 14 and 29 NCOs were detected.

It is worth noting that two pools in d74 showed multiple switching of markers between haplotypes. These events could be interpreted in two ways, one invoking a genuine 'complex' event generated through patchy repair of heteroduplex DNA and the other assuming that this pattern was the result of presence of two independent events (one CO and one NCO) in the same pool. Ideally, to resolve these events, allele-specific PCRs directed towards the most distal site should be used (as shown in Figure 4.10).

Unfortunately SNP rs4308913A/C, which was critical for resolving these events, lies within an Alu repeat and ASPs directed towards this site were largely non-specific. As an alternative, the probability of seeing multiple independent events in the same pool was estimated via data simulation, using a program written by Alec Jeffreys. The expected number of such events were determined and compared to the observed data. This analysis showed that the observed number of such 'mixed' pools was not significantly different to numbers expected, if two independent events occurred in the same pool by chance (*P*=0.965). These pools were therefore resolved as containing two simple events, namely, one CO and NCO.

Figure 4.10 Resolving pools showing multiple switches of marker alleles between two haplotypes


Selector sites for 1° and 2° PCRs
rs4446909A/G
rs5989681C/G
3920seqG/A
rs6644635C/T
rs5948839A/T
rs28633458G/A
rs7061961C/T
9620seqC/T
rs4308913A/C
ASP
To resolve pool 1:
Possible outcomes
A
OR
Not amplifiable
B
ASP
To resolve pool 2:
Possible outcomes
A
OR
Not amplifiable
B

Two pools showing multiple switches between haplotypes in d74. Circles represent informative SNPs. The amplified haplotype is the blue haplotype. In order to resolve these pools, ASPs directed towards the rs4308913 ‘red’ allele could be used, as shown below. These allele-specific PCRs may have several outcomes; the two most likely ones are shown below. Outcome ‘A’ (for both pools) would suggest that the observed pattern represents recombination events in the same molecule and thus invokes a ‘complex’ event. Outcome ‘B’ however, would imply that the observed pattern is representative of events in two separate molecules in the same pool as only one exchange event is recovered through allele-specific PCR. The other recombinant molecule, containing a NCO at rs5948839 and rs26833458 for 1 and 2 respectively, would not amplify in allele-specific PCRs specific to the rs4308913 ‘red’ allele.

4.2.2.5. Frequency and distribution of COs in d30 and d74

COs were scored as described in Figure 4.9. Using this strategy, only the most 5' COs could be detected from each CO-positive pool and additional COs that might be located closer to the distal end were masked. On account of this, numbers of COs scored were Poisson corrected (as described in Chapter 3). CO frequency (%) was thus calculated as follows:

CO frequency (%) = (Poisson-corrected COs/total number of molecules screened) x 100

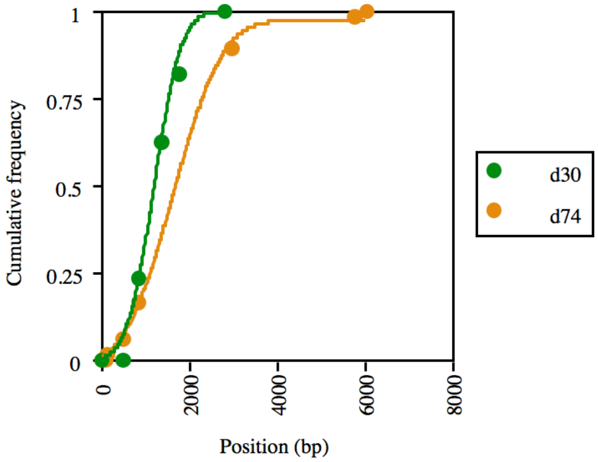
CO data from d30 and d74 is summarised in Table 4.1. d30 was assayed in both orientations, that is, both haplotypes were screened for recombinants. Since no significant differences either in rate or distribution was observed between the two orientations, the data shown below represent the average. d74 was assayed in one orientation only.

Table 4.1 Summary of CO data for d30 and d74

Donor name	Poisson-corrected COs	Molecules screened	CO frequency (%)	Lower 95% C.I. (%)	Upper 95% C.I. (%)
d30	34	22750	0.150	0.108	0.208
d74	67	18350	0.367	0.283	0.462

Both men showed evidence for clustering of COs into a recombination hot spot. However, CO frequencies at the hot spot (0.2-0.5%) were much lower than those observed at the most active autosomal hot spots identified by Webb and colleagues (2008). This hot spot will subsequently be referred to as hot spot PAR1-P (or PAR1 proximal). Least-squares best-fit values for hot spot centres and hot spot width, assuming a normal distribution of COs, were calculated using programs written by Alec Jeffreys and are summarised in Table 4.2.

Figure 4.11 Distribution of COs in d30 and d74



Cumulative frequency of COs for each man is plotted against distance (in bp). Positions in the graph are relative to the assay interval, with 0 bp corresponding to 1693891 bp (NCBI36/hg18). Underlying least-square best-fit cumulative frequency distributions, assuming normal distribution of COs, is also shown for each man (using green and orange lines for d30 and 74 respectively). Cumulative frequency data for d30 represents the average of two orientations whereas that for d74 represents data from the single analysed orientation.

Table 4.2 Summary of hot spot centres, width and peak activity for d30 and d74

Donor name	Hot spot centre (bp)	Width (bp)	Peak activity (cM/Mb)
d30	1695100	1220	124
d74	1695600	3490	168

Hot spot centres for d30 and d74 differ by ~500 bp. However, the hot spot in d74 maps to a much wider region (1693800 - 1697300 bp), which fully encompasses the interval harbouring the d30 hot spot (1694500 - 1695730 bp). Distributions of COs are not significantly different between the two donors either (P=0.09). Hence, there is no evidence suggesting that the two donors use different hot spots located in close proximity to one another.

4.2.2.6. Frequency and distribution of NCOs in d30 and d74

NCOs were scored as described in Figure 4.9. Numbers of NCOs calculated from these estimates will be conservative, since two or more identical NCOs in the same pool will be masked. Also, multiple-site NCOs will mask single-site events at these markers in the same PCR pool. Thus, for analysis of frequencies of NCOs, Poisson-corrected numbers of NCOs were used. Poisson-corrected numbers of NCOs = $-\ln[n/(p+n)]$, where, n= NCO-negative PCRs and p= NCO-positive PCRs. Frequency of NCOs for a particular marker = (Poisson-corrected NCOs / no. of molecules that can be screened for NCOs) x100. Calculation of frequencies of NCOs is shown for both donors in Table 4.3.

Table 4.3 Frequencies of NCOs in d30 and d74

- d30 (orientations combined)

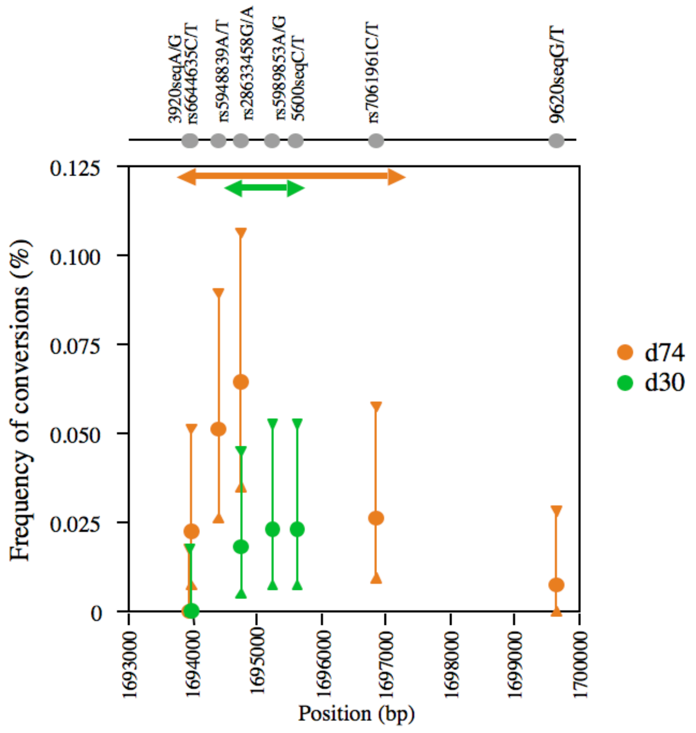
Marker	rs6644635	rs28633458	rs5989853	5600seq
Uncorrected NCOs	0	4	5	5
Poisson-corrected NCOs	0	4.01	5.01	5.02
Molecules screened	22800	22600	22200	22000
Frequency of NCOs	0	0.018	0.023	0.023

- d74 (one orientation only)

Marker	3920seq	rs6644635	rs5948839	rs28633458	rs7061961	5600seq
Uncorrected NCOs	0	4	9	11	4	1
Poisson-corrected NCOs	0	4.01	9.06	11.09	4.01	1.00
Molecules screened	18350	18300	18250	18100	16900	16800
Frequency of NCOs	0	0.022	0.050	0.061	0.024	0.006

In general, d74 shows higher frequencies of NCOs than d30, despite the presence of more central informative markers in the latter. Thus, the closest informative marker to the centre of the hot spot in d30, namely rs5989853, lies only ~135 bp away from the centre. In d74, however, the closest marker to the centre, namely rs28633458, is ~816 bp away. Relative distributions of NCOs in both men are shown in Figure 4.12.

Figure 4.12 Frequency and distribution of NCOs in d30 and d74



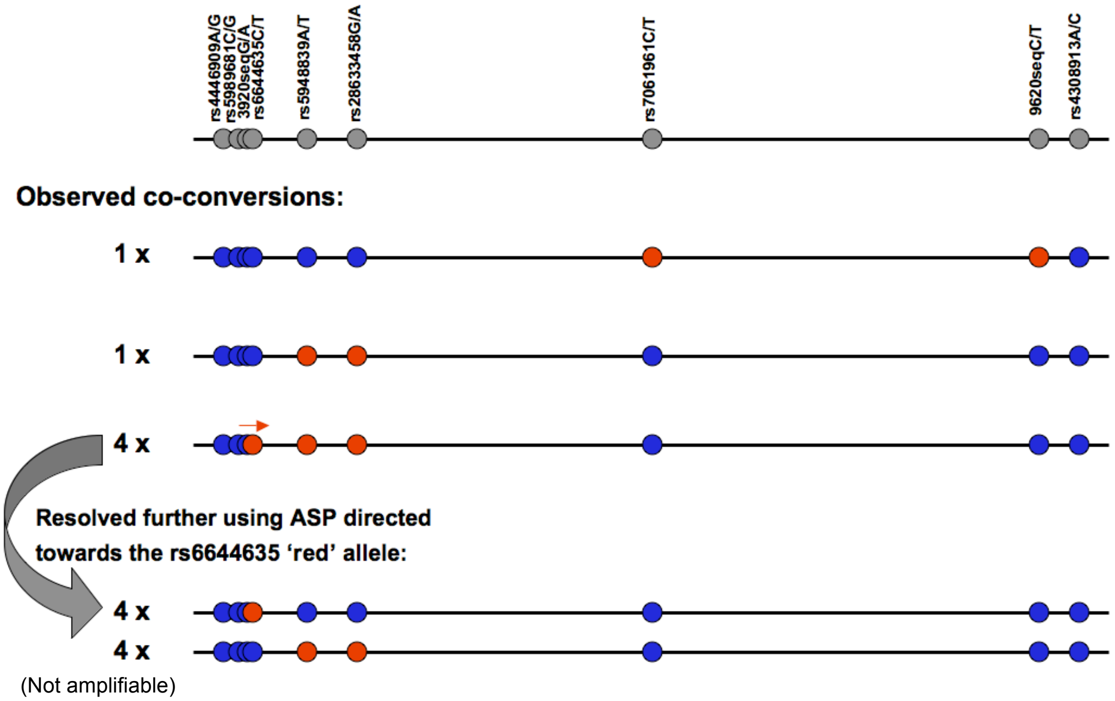
Frequency of NCOs in d30 and d74. No significant differences were detected in either numbers or distribution of conversions between the two orientations in d30, hence conversion data from both orientations are combined. The location of the CO hot spot for each donor is indicated using arrowed bars (orange for d74 and green for d30) to highlight co-localisation with COs. Vertical bars represent 95% C.I.s.

As shown in the above figure, NCOs in both donors largely co-localise with COs, with markers closest to hot spot centres showing the highest frequencies of NCOs. Thus rs26633458A/G, which is 816 bp away from the centre of the CO hot spot in d74, shows the highest frequency of NCOs. The same is true for d30, although a marker further away from the centre shows similarly high frequencies of NCOs as the central marker. It is also important to note that whilst NCOs largely co-localise with COs, they are not restricted to the CO hot spot. In fact, apparent NCOs can be detected as far as 4090 bp from the centre of the CO hot spot in d74.

Since all NCOs detected in d30 were single-site events, minimum tract lengths can in theory be just 1 bp, whereas maximum tract lengths ranged between 848 and 1451 bp (median = 880 bp). In d74 however, several pools contained ‘multiple-site’ NCOs involving several markers (shown in Figure 4.13). In order to estimate tract lengths of NCOs, it is necessary to determine whether these represent multiple independent events in the same pool or a single conversion tract involving all the aforementioned markers.

‘Triple-site’ NCOs observed could be partially resolved experimentally, using ASPs directed towards rs6644635C/T, into single-site NCOs at rs6644635C/T and ‘double-site’ NCOs involving rs5948839A/T and rs28633458A/G (see Figure 4.13). The latter however, could not be resolved further as ASPs targeted towards either of these sites were not specific enough.

Figure 4.13 'Multiple-site' NCOs observed in d74


rs4446909A/G
rs5989681C/G
3920seqG/A
rs6644635C/T
rs5948839A/T
rs28633458G/A
rs7061961C/T
9620seqC/T
rs4308913A/C
Observed co-conversions:
1 x
1 x
4 x
Resolved further using ASP directed towards the rs6644635 'red' allele:
4 x
4 x
(Not amplifiable)

In total, four 'triple-site' NCOs involving rs6644635C/T, rs5948839A/T and rs28633458A/G, one 'double-site' NCO involving rs5948839A/T and rs28633458A/G and another one involving rs7061961C/T and 9620seqC/G were detected. Of these, only the 'triple-site' NCOs could be resolved further using allele-specific PCRs directed towards internal sites. Thus, an ASP specific to the rs6644635 'red' allele was designed and used to selectively amplify molecules containing a NCO at this site. The products obtained were dot blotted and probed with rs5948839 and rs28633458 ASOs, in order to determine the allelic status of these markers in rs6644635 NCO-positive molecules. This analysis showed that in all four cases, rs6644635 NCOs and rs5948839 and rs28633458 NCOs were on different molecules and hence, the observed patterns represented at least two independent events.

In order to determine whether remaining 'double-site' NCOs, involving rs5948839A/T-rs28633458A/G and rs7061961C/T-9620seqC/G, represented two independent events in the same pool or genuine co-conversions, the probability of seeing two such events in the same pool by chance was calculated and the expected numbers of these events determined, as shown in Table 4.4.

Table 4.4 Estimating the probability of genuine co-conversions in d74

Type of multiple-site event	Observed numbers	Expected number of such events*	Probability of both events in same pool by chance
rs5948839- rs28633458	4	0.33	0.0011
rs7061961-9620seq	1	0.01	0.01

* Assuming these represent two independent events that occur in the same pool by chance.

Thus, for both types of 'multiple-site' NCOs, the probability of these representing two independent events is very small (*P*=0.004 and 0.025, for rs5948839- rs28633458 and rs7061961-9620seq co-conversions respectively). It is thus, likely that these pools contain genuine co-conversion events. Based on this assumption, maximum and minimum tract lengths for single-site NCOs and co-conversions are listed in Table 4.5.

Table 4.5 Estimation of NCO tract lengths for d74

Type of event	Minimum length (bp)	Maximum length (bp)
Single-site NCOs	1	102 - 4904
rs5948839- rs28633458 co-conversions	392	2870
rs7061961- 9620seq co-conversions	2870	5188

4.2.3. Sperm analysis of the distal LDU step (assay 'A')

The interval encompassing the more distal LDU step was analysed using a conventional sperm CO assay, wherein two rounds of repulsion-phase allele-specific PCRs are used to selectively amplify recombinant molecules from pools of sperm DNA, as described in Chapter 3.

Markers rs28612605A/G and rs28640198A/G were chosen as primary and secondary 5' selector sites respectively and corresponding ASPs designed and optimised. The small region of marker association separating the two LDU steps and the presence of very few SNPs within single-copy DNA in this region, largely limited the choice of 3' selector sites. Positions of SNPs rs6644635C/T and 3920seqA/G were ideal for primary and secondary selector sites respectively, as these allowed for sufficient overlap between the two assay intervals but did not extend into the recombination hot spot corresponding to the proximal LDU step. However, the reverse ASP targeted towards the 3920seq ‘A’ allele was not specific, whilst that for the ‘G’ allele produced very low yields and only at low temperatures. At these temperatures, 5' ASPs were no longer specific. Hence, 3920seqA/G could not be used as a selector site. Instead, rs5989681C/G, the closest marker 5' to 3920seqA/G, was chosen as the secondary selector site.

Whilst a number of men in the semen donor panel were heterozygous for the chosen selector sites, d74 was chosen for analysis even though he was not particularly informative over the interval assayed. This was done in order to get a complete recombination profile for the entire region in a single man and to determine whether both hot spots were active in this man and if so, to compare their relative activities.

The linkage phase of selector sites in d74 was established and DNA inputs for the bulk recovery of COs was determined, as described in Chapter 3, sections 3.2.3 and 3.2.5 respectively. CO assays were carried out in one orientation only, using the most specific and efficient combination of ASPs, as shown in Figure 4.14.

Figure 4.14 Strategy for sperm CO assay 'A'

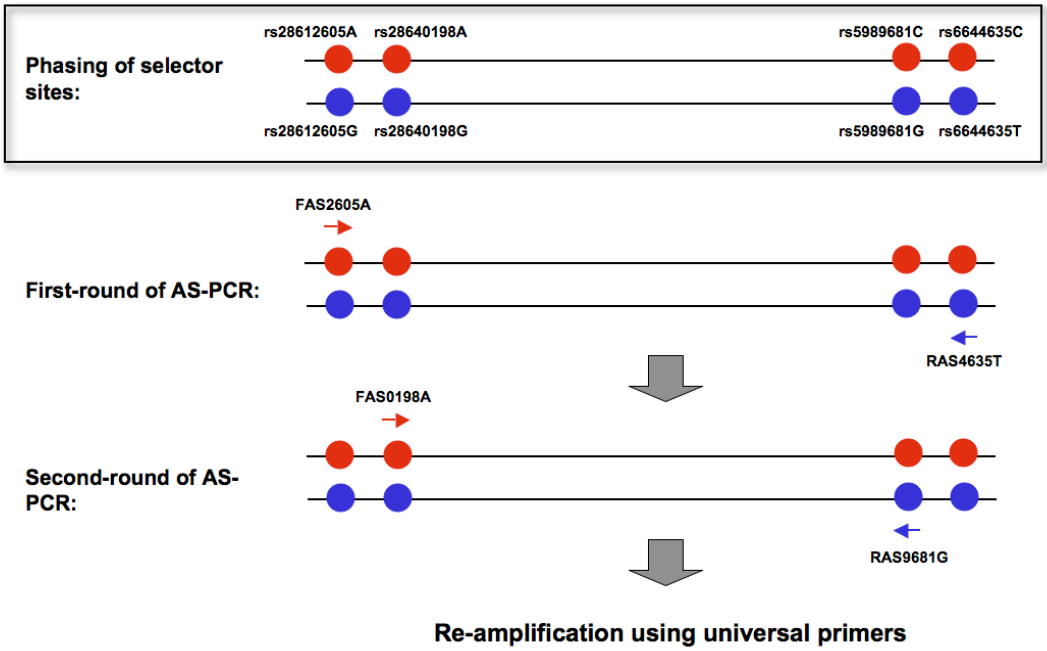

Phasing of selector sites:
rs28612605A
rs28640198A
rs5989681C
rs6644635C
rs28612605G
rs28640198G
rs5989681G
rs6644635T
FAS2605A
First-round of AS-PCR:
RAS4635T
FAS0198A
Second-round of AS-PCR:
RAS9681G
Re-amplification using universal primers

Figure shows phasing of d74 selector sites and combination of ASPs used to selectively amplify CO-containing molecules from pools of d74 sperm DNA. FAS2605A and RAS4635T were used in the first-round repulsion-phase PCR and FAS0198A and RAS9681G in the second round. Selector sites are indicated using circles. Haplotypes have been arbitrarily assigned red and blue colours for purposes of clarity. Details of PCR conditions used in this assay are listed in Appendix 4.5.

Secondary PCR products obtained from two rounds of repulsion-phase PCRs were re-amplified using internal primers. Tertiary products thus generated were dot blotted. The same dot blots also contained parental haplotypes, in order to determine haplotypes of internal SNPs.

CO breakpoints were mapped by typing internal SNPs, using ASO hybridisation to dot blots (see Chapter 3, section 3.2.6 for more details). It is worth noting that smaller tertiary amplicons could not be generated to type markers within repetitive elements owing to time constraints. Thus, all informative markers within the interval assayed were not typed. Recombination events were scored and CO frequencies determined as described in Chapter 3, sections 3.2.7/8. COs mapping to terminal intervals were excluded from analysis since these events cannot be verified and may well represent bleed-through of parental haplotypes (which occur since ASPs are not 100% specific and may mis-prime from the unselected haplotype, as explained in Chapter 3). Thus, both 5′ and 3′ end interval events, which comprised 32% and 14% of all COs respectively, were rejected in order to avoid artificially inflating estimates of CO frequencies. Data from this sperm CO assay is summarised in Table 4.6.

Table 4.6 Summary of d74 CO assay 'A'

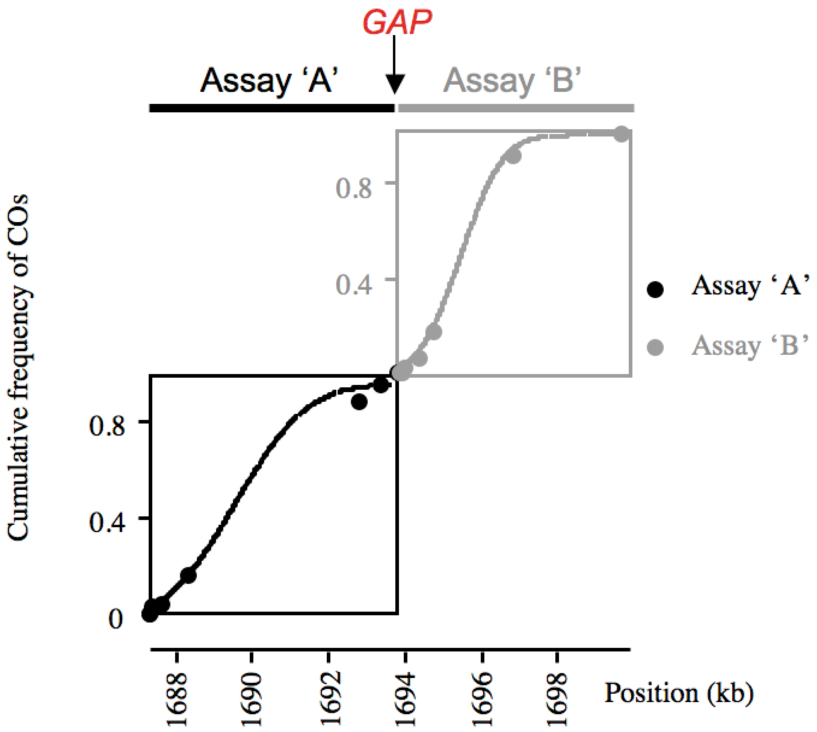
Molecules screened	Corrected COs	CO frequency (%)	Hot spot centre (bp)	Width (bp)	Peak activity (cM/Mb)
25091	114	0.455	1689600	5950	121

The exclusion of end interval COs and the lack of informative markers over a nearly 4.5 kb region within the assay interval made it very difficult to determine whether a classical CO hot spot was present here. However, high CO frequencies approximating ~0.5% argue in favour of the presence of a CO hot spot in this region. Values of hot spot centre, width and peak activity were thus determined assuming that the region contained a hot spot wherein CO-breakpoints were normally distributed. The distribution of COs over the hot spot is discussed further in the following section. This hot spot will subsequently be referred to as PAR1-D (PAR1- distal).

4.2.4. Overall recombination profile in d74

Data from sperm assays confirmed the presence of two active CO hot spots in d74, with the centres of the two hot spots ~6 kb away from each other. Both hot spots were similar in activity, with the distal hot spot recombining at 0.455% (95% C.I. = 0.375-0.544 %) and the proximal one at 0.367% (95% C.I. = 0.283-0.462 %). This represented the first example of a pseudoautosomal 'double hot spot' in humans. The distribution of COs in d74 across the entire interval is shown in Figure 4.15.

Figure 4.15 Distribution of COs in d74



The cumulative frequencies of COs across both assay intervals in d74. Black circles represent data from assay 'A', whilst grey circles show assay 'B' data. Underlying least-square best-fit cumulative frequency distributions are indicated using black and grey lines respectively. It is worth noting that recombination data could not be obtained over an interval of 109 bp between the two assay intervals, since COs mapping to terminal intervals were excluded from analysis in assay A.

4.3. Discussion

Work in this Chapter describes the identification of two closely-spaced, actively-recombining sperm CO hot spots, from analysis of population-diversity data. This work has highlighted that population studies in PAR1 suffer from some significant difficulties and limitations (discussed below). Furthermore, hot spots identified through this work provide the first example of a pseudoautosomal 'double- hot spot'. These hot spots provide an important resource for investigating local variations in PAR1 recombination rates and can potentially be used to investigate CO interference in PAR1 at very high resolution. Finally, this work provides additional insights into the distributions and frequencies of NCOs in PAR1 and facilitates valuable comparisons with previous data from autosomes and from the *SHOX* hot spot, as discussed in the following paragraphs.

4.3.1. Difficulties and limitations of using LD profiles for hot spot identification in PAR1

In this work, the largest step in metric LDU within PAR1 was investigated and not one, but two sperm CO hot spots were identified. The first limitation of HapMap population data lay in the fact that it was unable to detect both hot spots. Instead, it was consistent with the presence of one, unusually broad, recombination hot spot. More examples like this have been reported on autosomes. For instance, distribution of sperm COs in 3 out of 16-intervals studied by Webb and colleagues (2008; discussed in section 4.1.1) were consistent with the presence of double hot spots. However, population data for at least two of these intervals failed to identify both hot spots, instead showing a single broad interval of elevated activity. Further, the LDU step was centred at the region between the two sperm hot spots, which is in fact recombinationally suppressed. Collectively, these examples highlight the imperfections of recombination profiles obtained using HapMap population data and stress the importance of hot spot characterisation using sperm approaches.

Secondly, detecting hot spots using LD profiles is considerably more difficult in PAR1 compared to autosomes, on account of the high background recombination rates in this region and extensive LD breakdown (see Figure 4.1). Curiously however, although both hot spots identified presently are reasonably active (~0.3%), they are not nearly as active as some hot spots identified on autosomes from similar surveys (CO frequencies up to 1%). Since this interval presents the largest step in metric LDU in PAR1, it raises important questions as to whether more active hot spots exist in this region and if so, how does one detect them.

A complementary approach of detecting active CO hot spots in humans is from high-resolution pedigree analysis data, as described in Chapter 1, section 1.5.2.6 (Coop *et al.*, 2008; Kong *et al.*, 2010). Unfortunately, these studies have to-date only focussed on autosomes and do not provide information on putative PAR1 hot spots. Moreover, in many cases, very few COs can be detected from pedigree analysis in a particular region, limiting inferences that can be made from them. Sufficiently high-resolution pedigree data encompassing PAR1, in combination with population diversity data, will however no doubt prove valuable in terms of shedding additional light on the likely distributions of CO hot spots across PAR1 and provide additional targets for sperm assays.

4.3.2. Utilising a pair of pseudoautosomal hot spots for investigating the processes of recombination

This work describes the identification of two very closely located sperm CO hot spots. It is clear that further work is required to fully characterise both hot spots and to address the following issues. First, d74 was not very informative across hot spot PAR1-D and this made it very difficult to ascertain whether a classical recombination hot spot was present in this region. It also adversely affected the accuracy of estimates of hot spot centre and width. Typing additional, more informative donors will no doubt allow the hot spot to be characterised with greater accuracy and detail. Second, typing additional men over hot spot PAR1-P will further address the issue of largely differing hot spot widths observed between d30 and d74 and will hopefully provide a more robust estimate for hot spot width. Finally, a small gap exists between the two intervals assayed. It would be ideal to be able to close this gap and better still, to have a small over-lap between the two intervals, for purposes of obtaining a complete dataset.

Once fully characterised, this pair of hot spots could provide a valuable resource for investigating recombination processes. First, these hot spots would allow us to address if and how, CO activities at nearby hot spots vary relative to each other. Data from d74 indicates similar activities at both hot spots. It will be interesting to see if this is maintained in other men or if one or the other hot spot emerges more active relative to the other in different men. A more extensive survey of activity will also allow comparisons with autosomal double hot spots, including hot spots MSTM1a/MSTM1b (discussed in Chapter 3, section 3.3). Finally, if polymorphic, these hot spots could shed further light into the dynamics and processes of hot spot evolution, as discussed previously in Chapter 3. A ~2.5 fold variation in activity between d30 and d74 at hot spot PAR1-P, indeed hints that such polymorphisms are likely.

Two closely situated hot spots also provide a valuable resource for investigating double COs using batch sperm typing. This in turn, could provide important insights into the phenomenon of interference in PAR1. Double COs in PAR1 have in fact, previously been reported by several studies, although these have been separated by much larger distances (>100 kb) (Henke *et al*., 1993; Rappold *et al*., 1994; Schmitt *et al*., 1994; Lien *et al*., 2000; Flaquer *et al*., 2009). The presence of more than one CO within such a small genomic region has raised several questions regarding the mechanisms of interference in this region. Indeed, Schmitt *et al.* (1994) speculated that interference in PAR1 could be a function of genetic, rather than physical distance.

Earlier models of interference had suggested that interference depends on and diminishes with physical distance (Fox 1973; King & Mortimer, 1990). However, observations indicating longer interference distances in organisms showing low recombination rates per kb (eg. *Drosophila melanogaster*) and shorter distances in organisms with higher rates of recombination per kb (eg. *Saccharomyces cerevisiae*), suggested a relationship between genetic distance and levels of interference. A model proposed by Foss *et al*. (1993), further elaborated on this. This model suggests that interference results from constraints imposed on the resolution of randomly-spaced recombination intermediates. By constraints, Foss *et al*. (1993) refer to the requirement for a certain number of NCOs in between two COs. In other words, it is the number of intervening NCOs, rather than physical distance, that determines the interference distance between adjacent COs. Additional support for this model comes from earlier observations by Stadler (1959), suggesting that NCOs in *Neurospora* promote the occurrence of COs in their vicinity and from the work of Mortimer & Fogel (1969), who suggested that in *Saccharomyces*, COs and NCOs alternate along the length of a tetrad.

A standard CO assay, similar to that used to characterise hot spot PAR1-D, but extending across both hot spots cannot be used to detect double recombinants. Indeed, a modified assay similar to the one used to study recombination at the proximal hot spot, will need to be designed. Preliminary attempts have been made at carrying out such assays, but these PCRs have either failed or provided multiple non-specific products. It is likely that this was the consequence of the highly repetitive genomic context of the interval assayed, coupled with the large size of the amplicon used (>15 kb). Using additives such as DMSO or betaine however, might improve amplification efficiency and specificity, thereby facilitating an effective sperm assay. Modulation of PCR buffer concentrations and that of other reagents might also help in assay optimisation. The aforementioned optimisations could not be carried out owing to time constraints.

4.3.3. NCOs at hot spot PAR1-P and comparisons with autosomal hot spots

This work provides valuable information on the distributions and frequencies of NCOs across hot spot PAR1-P in two different North European men. Meiotic NCOs in humans have previously been investigated at hot spots *DNA3*, *DMB2*, *SHOX*, as well as at the β-globin hot spot (Jeffreys & May, 2004; Holloway, Lawson & Jeffreys, 2006). In the following paragraphs, data from these studies will be compared with NCO data from hot spot PAR1-P and similarities and differences between them highlighted.

4.3.3.1. Distribution of NCOs at PAR1-P

First, previous data mainly from two autosomal hot spots *DNA3* and *DMB2*, are consistent with presence of highly-localised NCOs, with peaks in NCO activity coinciding with the centre of the CO hot spot (Jeffreys & May, 2004). The same is true for NCOs at hot spot PAR1-P, wherein markers closest to the hot spot centre show the highest NCO frequencies in both d30 and d74.

It is important to mention however that although most NCOs at PAR1-P largely co-localise with COs, apparent NCOs can be detected up to 4 kb away from the centre of the CO hot spot (see Figure 4.12). This is in sharp contrast to data from the *SHOX* hot spot, wherein a marker merely 500 bp away from the centre did not show any evidence of NCOs (Jeffreys & May, 2004). Difference in gradient of NCOs at PAR1-P, relative to previously studied hot spots, is also evident from d30 sperm data. Thus, markers rs5989853 and 5600seq, which lie 135 and 515 bp away from the hot spot centre respectively, show identical frequencies of NCOs. It is worth bearing in mind here that sample sizes of NCOs screened, that is 14 and 29 for d30 and d74 respectively, are rather small. Hence, conclusions regarding gradients of NCOs should be treated with caution and verified using larger datasets.

4.3.3.2. Tract lengths of NCOs

The majority of NCOs (~90%) recovered involved transfer of a single marker allele from one haplotype to another, consistent with previous data from Jeffreys and May (2004). Tract-lengths associated with these NCOs ranged between 1-880 bp for d30 and between 1-4904 bp in d74, with median maximum lengths being 880 and 1613 bp respectively. On account of low marker densities in both men, it is very likely that true tract-lengths are much shorter and similar to existing estimates of 55-290 bp (Jeffreys & May, 2004). Interestingly however, data from d74 shows some evidence of apparent co-conversion of markers located 2870 bp apart from each other. This is unprecedented, since previously co-conversions have only been reported in markers lying in very close proximity to one another. Thus, the longest NCO tract at hot spot *DNA3* encompassed four different markers distributed over only 300 bp (Jeffreys & May, 2004). It will be interesting to see if a more extensive survey of NCOs reveals more instances of co-conversions involving markers located >1 kb apart. If genuine, such events could provide important insights into the processes of meiotic recombination in the human male germ-line, including extents of DNA resection and branch migration.

4.3.3.3. Relative frequencies of COs and NCOs

CO: NCO ratios for both d30 and d74 was ~6:1. Although, larger sample sizes of NCOs are ideally required for robust estimates of CO:NCO ratios, those currently observed are well within the range reported at other autosomal hot spots, which varies between 2.7:1 at *DNA3* and 1:12 at β-globin (Holloway, Lawson & Jeffreys, 2006). It has been suggested that preference for one or the other outcome of recombination at different hot spots is consistent with COs and NCOs being generated by different pathways, although mechanisms influencing this choice remain completely unknown (Holloway, Lawson & Jeffreys, 2006).

Chapter 5. Identification and characterisation of the first PAR2 recombination hot spot

5.1. Introduction

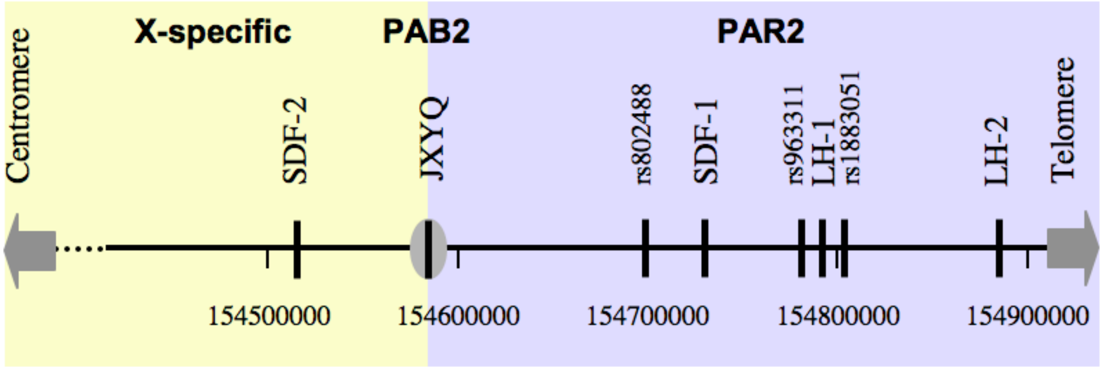
Cytogenetic data suggest that meiotic associations between distal ends of X and Y-chromosomes rarely persist into mid- to late pachytene, during which stage COs are formed (Chandley *et al.* 1984, 1987; Speed & Chandley 1990, Armstrong *et al.* 1994). These observations raised important questions regarding the nature of genetic exchange within PAR2 and led to speculation that recombination in PAR2 mainly occurred in the form of NCOs, rather than COs (Kvaløy *et al.* 1994). Nonetheless, several lines of evidence suggest that reciprocal exchanges do indeed occur in this region.

First, low-resolution linkage data from Freije *et al*. (1992) showed evidence of exchange within PAR2 in three-generation CEPH pedigrees. In this study, Freije *et al*. (1992) studied inheritance patterns of X-specific marker SDF-2 and Xq/Yq pseudoautosomal marker SDF-1 in 32 informative CEPH pedigrees. This revealed four PAR2 exchange events in 195 informative male meioses and another event in 238 informative female meioses. Thus, PAR2 appeared recombinationally active in both sexes, but like PAR1, male rates were substantially higher than female rates.

A subsequent study also based on CEPH pedigrees and using three additional polymorphic markers, LH-1, LH-2 and JXYQ, identified six COs from 288 informative male meioses (Li & Hamer, 1995). All COs mapped between markers JXYQ and SDF-1. The female recombination event, previously reported by Freije *et al.* (1992), was not replicated in this study. Li & Hamer (1995) further showed that whilst there was significant breakdown of association between PAR2 and sex-specific regions of X and Y, the three PAR2 markers, SDF-1, LH-1 and LH-2, remained linked with one another. This suggested a non-uniform distribution of COs in PAR2, with most events occurring between JXYQ, which lies at the PAR2 boundary and SDF-1, which lies ~147 kb distal to it.

More recently, Flaquer *et al.* (2009) studied the inheritance patterns of six markers, including SDF-1, LH-1 and LH-2, in 28 CEPH pedigrees. However, they only reported one exchange event in male meiosis and no events in female meiosis. Curiously, the study also suggested that recombination rates in PAR2 were similar in the two sexes. This is in contrast to previous published data and is probably a consequence of the analysis being based on a single CO. Relative positions of the markers analysed in these studies are shown in Figure 5.1.

Figure 5.1 Relative positions of markers used in the study of recombination within PAR2 in CEPH pedigrees



Co-ordinates given are for chrX, NCBI36/hg18. Freije *et al.* (1992) analysed inheritance of markers SDF-1 (also referred to as DXYS154) and SDF-2; Li & Hamer (1995) used SDF-1, SDF-2, LH-1, LH-2 and JXYQ; Flaquer *et al.* (2009) used SDF-1, LH-1, LH-2, rs802488, rs963311 and rs188305. Curiously, rs188305 maps to chromosome 7, as reported in the UCSC Genome Browser and dbSNP. However, rs1883051 does map to PAR2. We suspect that the 'rs' number provided in the paper is incorrect and have used the position of rs1883051 in the figure. Physical positions of these markers were not stated in respective papers that used them for analysis of recombination; locations used in the figure were calculated as described in Appendix 5.1.

In addition to pedigree studies, evidence of crossing-over in this region was also obtained through single-sperm typing. Thus, Schmitt *et al.* (1994) studied 505 sperm from two men and reported exchange in PAR2 in 0.3% of male meioses. Subsequently, Lien *et al.* (2000)

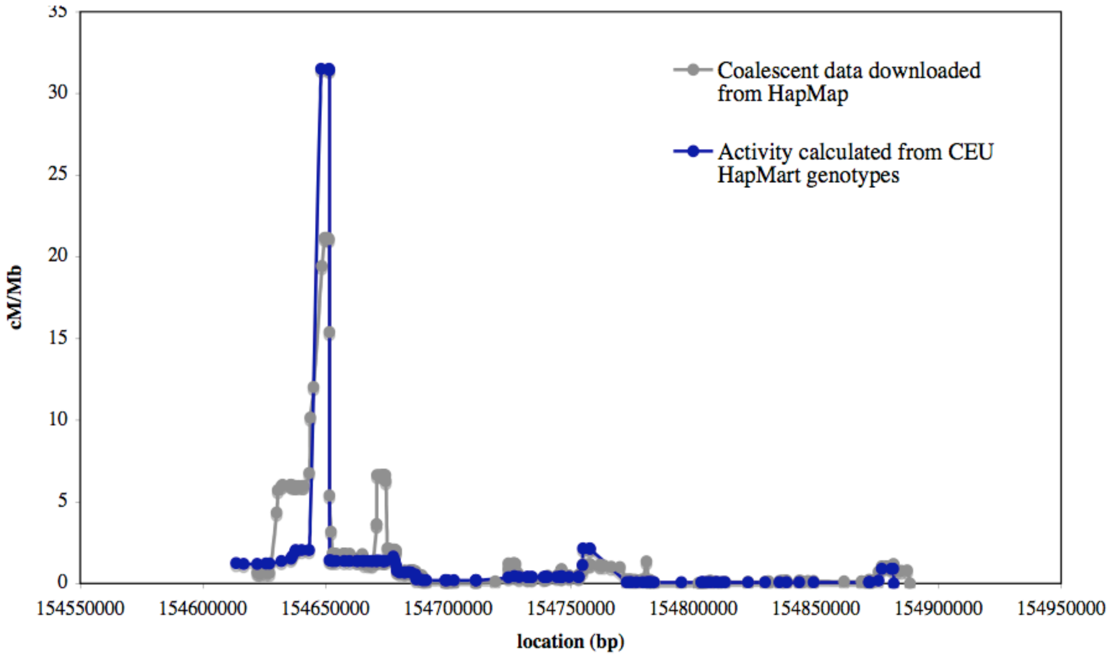
analysed 1912 sperm from 4 men and reported PAR2 recombination frequencies of 0.7%. The most distal marker in both studies was DXYS154, otherwise known as SDF-1. Estimates of recombination from these studies were thus, only based on data from the proximal 147 kb of this 330 kb-long region.

5.1.1. This work

This work is aimed at further exploring meiotic recombination in PAR2, at determining whether COs in this genomic region cluster into hot spots and if so, how active are these compared to ones on autosomes and in the Xp/Yp pseudoautosomal region. Inferred historical recombination profiles obtained from analysis of population diversity data are generally good indicators of contemporary recombination rates, as discussed previously in Chapters 3 and 4. Hence, coalescent data for PAR2 published by HapMap (2007) were first examined in order to gain an idea of the recombination profile across PAR2. This showed that most historical recombination events in PAR2 occurred in two clusters, located within the proximal 95 kb of PAR2 (see Figure 5.2).

Coalescent analysis was next repeated on 115 markers, using an online program (available at http://143.210.151.197/cgi-bin/LDhat/ldhat.pl), but using a more stringent set of controls. Thus, 64 HapMap markers for which less than 70% of individuals were unambiguously typed were excluded from analysis. Also, only unrelated individuals were included, as using closely related individuals may underestimate breakdown of marker association. Curiously, the smaller peak in historical activity could not be replicated using this dataset (Figure 5.2), indicating that this may well have been an artefact in the original dataset. On account of this, this former interval was not subjected to further analysis.

Figure 5.2 Historical recombination profiles across PAR2



The grey line on the graph represents historical recombination activity across PAR2, obtained from coalescent analysis of Phase II HapMap data (CEU). Values for recombination activity (cM/Mb), calculated from HapMap genotype data, were downloaded directly from the HapMap website (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2006-10_rel21_phaseI+II/rates/). For the second dataset (blue line on the graph), HapMap (release 27) genotypes for the CEU population were downloaded from HapMart. Markers for which less than 70% of individuals were unambiguously typed, as well as related individuals, were excluded from analysis. Values for recombination activity (cM/Mb) were obtained using an online coalescent analysis program. All positions are given on ChrX, NCBI36/hg18.

The larger peak in recombination activity, flanked on either side by recombinationally suppressed regions, is typical for a historical recombination hot spot. Interestingly, exchange intervals reported in PAR2-recombinant CEPH pedigrees, completely encompassed this putative hot spot interval (Freije *et al.*, 1992; Li & Hamer, 1995). In order to investigate whether these *de novo* paternal exchanges occurred within the putative hot spot interval, these recombination events were investigated further via typing additional SNPs across this region of high historical activity. Subsequently, sperm assays were also carried out in order to confirm the existence of an active recombination hot spot and to characterise it at micro-Morgan resolution.

5.2. Results

5.2.1. High-resolution mapping of exchanges in PAR2-recombinant CEPH pedigrees

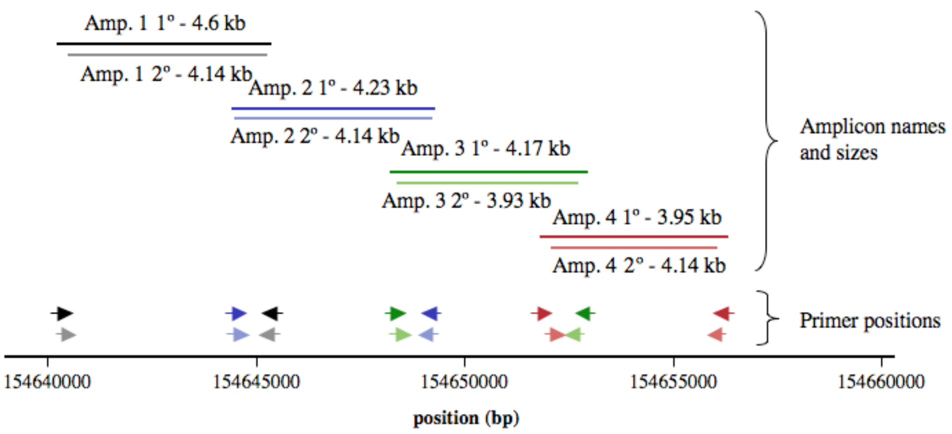
5.2.1.1. Choice of CEPH DNAs for study

In order to determine whether exchange events reported in CEPH pedigrees co-localised with the putative Phase II HapMap historical hot spot (Figure 5.2) and to verify and characterise these events at higher resolutions, the five reported PAR2-recombinant CEPH pedigrees 1362, 1423, 1377, 1332 and 1416 were studied first. For all families, except 1423, all four grandparents and both parents were included in analysis. For family 1423, sufficient DNA for the maternal grandmother was not locally available at the time of this study. All recombinant children, as reported by Freije *et al.* (1992) and Li & Hamer (1995) were also analysed. Where possible, at least one non-recombinant son and daughter from all families were included, in order to confirm the transfer of marker alleles between males and females. Since all children in family 1377 are male, only one recombinant and one non-recombinant son were studied for this pedigree. The choice of non-recombinant children was based mainly on DNA availability.

5.2.1.2. PCR amplification of the interval encompassing the putative PAR2 historical hot spot and mapping recombinants in CEPH pedigrees

The DNA sequence encompassing the peak in historical activity was first retrieved and annotated using the UCSC genome browser (see Appendix 5.2 for the annotated sequence). Next, four overlapping amplicons were designed over the interval. Sizes of amplicons were limited to typically < 5kb, in order to increase ease of amplification using CEPH DNAs. Some of these samples had been frozen/thawed several times, which adversely affected DNA quality. Relative positions of amplicons used are shown in Figure 5.3. Primer combinations and PCR conditions used for amplification are listed in Appendix 5.3. It is worth noting that PCRs for a non-recombinant daughter (142305) in family 1423 consistently failed. This sample could therefore, not be included in this study.

Figure 5.3 Strategy for amplifying the interval of high historical recombination activity


Amp. 1 1° - 4.6 kb
Amp. 1 2° - 4.14 kb
Amp. 2 1° - 4.23 kb
Amp. 2 2° - 4.14 kb
Amp. 3 1° - 4.17 kb
Amp. 3 2° - 3.93 kb
Amp. 4 1° - 3.95 kb
Amp. 4 2° - 4.14 kb
Amplicon names and sizes
Primer positions
154640000
154645000
154650000
154655000
154660000
position (bp)

Key: Primary PCR primers

Symbol on figure	Primer name	Amplicon name
→	1.1F	Amplicon 1
←	2.2R	Amplicon 1
→	64.46F	Amplicon 2
←	S24	Amplicon 2
→	64.85F	Amplicon 3
←	7.1R	Amplicon 3
→	65.21F	Amplicon 4
←	6.1R	Amplicon 4

Key: Secondary PCR primers

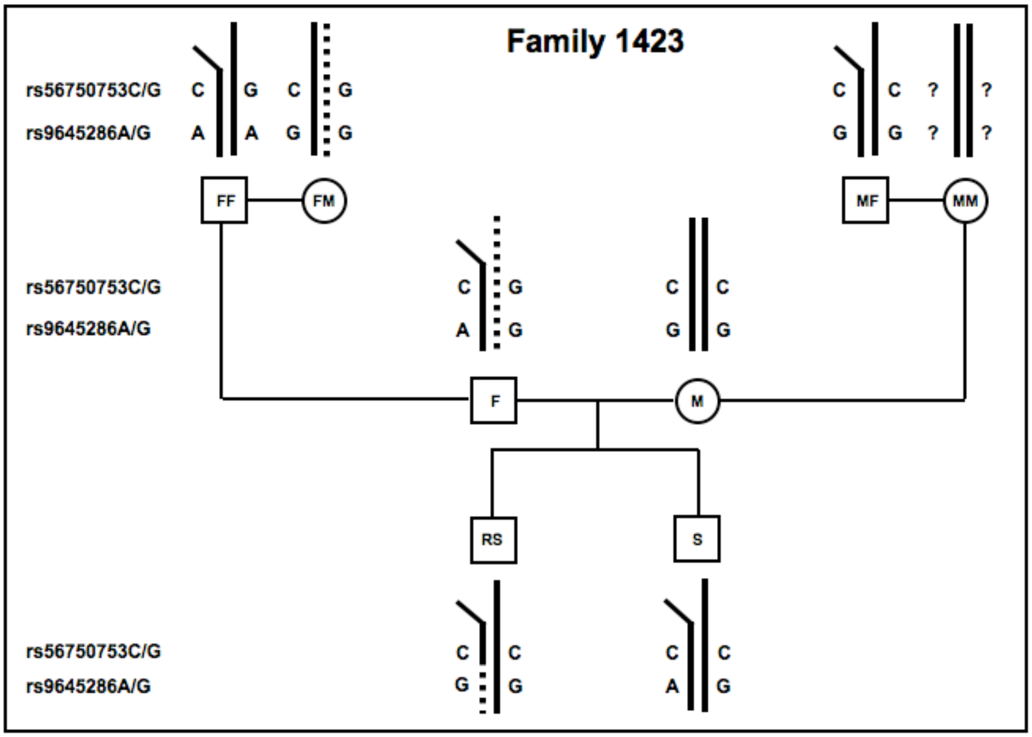
Symbol on figure	Primer name	Amplicon name
→	1.2F	Amplicon 1
←	2.1R	Amplicon 1
→	64.47F	Amplicon 2
←	S8	Amplicon 2
→	64.86F	Amplicon 3
←	S4	Amplicon 3
→	S3	Amplicon 4
←	6.2R	Amplicon 4

Four overlapping amplicons were designed across the interval. Amplicon sizes were limited to less than 5kb and nested PCRs were used for each amplicon. All positions are as per ChrX, NCBI36/hg18.

PCR products thus obtained were electrophoresed on an agarose gel in order to verify sizes of respective products and to ensure all samples had amplified evenly. PCR products were subsequently dot blotted. SNPs recorded in dbSNP, present within these amplicons, were then genotyped using ASO hybridisation to dot blots (examples shown in Appendix 5.5).

An exchange event was inferred when genotype data suggested that a recombinant son or daughter inherited a SNP allele present on his/her father's X- or Y- chromosome respectively, as shown in Figure 5.4.

Figure 5.4 Mapping recombination events in three-generation CEPH pedigrees

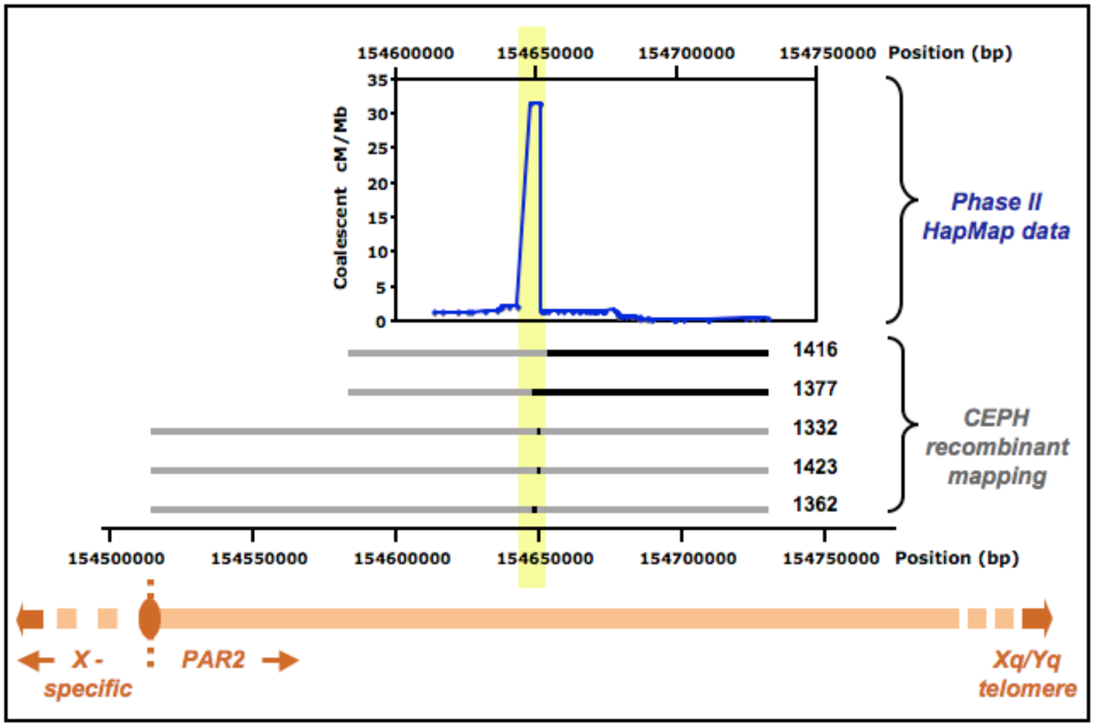

Family 1423
rs56750753C/G
rs9645286A/G
C
A
G
A
C
G
G
G
FF
FM
C
G
C
G
?
?
?
?
MF
MM
rs56750753C/G
rs9645286A/G
C
A
G
G
C
G
C
G
F
M
RS
S
rs56750753C/G
rs9645286A/G
C
G
C
G
C
A
C
G

Inheritance patterns of two informative markers rs56750753 and rs9645286 are shown. The X chromosome inherited from the mother in the pedigree is represented using a solid line, whilst that inherited from the father is indicated by a dotted line. The Y chromosome is bent. The recombinant son (RS) inherited the rs5670753C allele on his Y chromosome from his father's Y, but the rs9645286G allele from his father's X, thus indicating the presence of a CO event between X and Y chromsomes in the interval between the two SNPs.

In the example shown in Figure 5.4, genotype data for SNPs rs56750753 and rs9645286 showed that both alleles on the Y chromosome of the non-recombinant son were inherited from his father's Y chromosome. The Y chromosome of the recombinant son, on the other hand, carried rs56750753C from his father's Y chromosome, as well as rs9645286G from his father's X chromosome. This indicated the presence of an exchange event between X and Y- chromosomes, which mapped to the interval between rs56750753 and rs9645286.

Using genotype data at 17 SNPs, exchange events in families 1362, 1423 and 1332 were mapped to intervals of 2044bp, 1649bp and 867bp respectively. This presented a major improvement in terms of resolution, compared to the original studies that reported these events (data summarised in Figure 5.5). Moreover, this study also showed that exchange events reported in these families overlapped with the peak in PAR2 historical recombination activity.

Figure 5.5 Data summary from CEPH recombinant mapping



Grey bars represent exchange intervals determined by Freije *et al.* (1992) and Li & Hamer (1995). For families 1362, 1423 and 1332, these represent the interval between SDF-1 and SDF-2. For families 1377 and 1416, these indicate the distance between JXYQ and SDF-1. Black bars represent exchange intervals determined from the present study. For families 1362, 1423 and 1332, these represent intervals between rs68008496-rs700443, rs56750753-rs3870343 and rs700442-rs71190340 respectively. The exchange events in family 1377 could not be mapped at a similarly high resolution on account of paucity of informative markers. Current data suggest the event must map distal to rs68008496, placing it in the interval between this marker and SDF-1. In family 1416, the interval studied presently does not appear to be involved in the recombination event previously reported. Rather, the exchange must map to the interval between rs9645286 and SDF-1. The graph represents data from coalescent analysis of phase II HapMap genotype data (CEU). The yellow bar highlights that in at least three CEPH families, the exchange event maps to the interval of high historical recombination activity identified from HapMap Phase II data (2007).

In family 1377, all but two markers were uninformative and these two markers were towards the 5′ end of the interval studied. Thus, it was not certain if the reported event fell within the region studied (3′ of the informative markers), although the possibility cannot be discounted either. In pedigree 1416 however, informative markers towards the distal end of the 15kb interval confirmed that this region was not involved in the exchange event reported by Li & Hamer (1995). Full details of how exchange intervals were calculated for each pedigree are shown in Appendix 5.5.

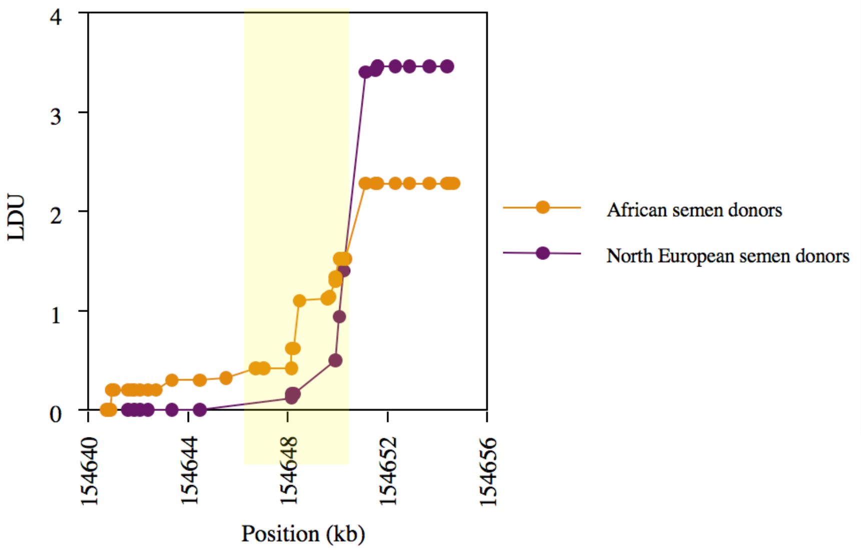
5.2.2. Analysis of patterns of linkage disequilibrium over the target interval

Since two separate lines of evidence pointed towards recombination in the interval between 154,640,762 and 154,655,521 (chrX, NCBI36/hg18), this region was next subjected to LD analysis in 97 North European and 74 African semen donors. This analysis was done in order to investigate historical patterns of recombination at higher resolutions than that of the HapMap dataset and to determine if these patterns were consistent with the presence of a recombination hot spot in this region in either/both populations. LD analysis in semen donor panels was also imperative for subsequent high-resolution sperm assays, which are required to fully characterise recombination hot spots.

For this analysis, additional SNPs in the target interval were first identified through dbSNP. In order to determine SNP genotypes in the African semen donor panel, sample DNAs were amplified using the same primers and PCR conditions as those used to study recombinants in CEPH pedigrees (see Appendix 5.3). PCR products thus obtained were dot blotted. ASO hybridisation to dot blots was used to determine SNP genotypes, as shown previously in Chapter 4. The North European semen donors were genotyped by M. Denniff.

In total, 36 markers were typed for the North European panel and 51 markers for the African panel. Complete genotype data from both panels is shown in Appendix 5.6. Genotype data from both panels were used to determine LD maps across the region. Values of LDU for each marker, as well as their minor allele frequencies are shown in Appendix 5.6. LDU profiles for both donor panels are shown in Figure 5.6.

Figure 5.6 LDU profiles across the target interval



Values for LDU are plotted along the Y-axis and distance is plotted along the X-axis. Both populations show breakdown of LD over the highlighted region, spanning approximately 4 kb. There appears to be greater LD breakdown in Europeans, as reflected by a larger step size. However, this may be a consequence of fewer informative markers being present in the Europeans compared with Africans.

Thus, LDU profiles in both sets of semen donors show localised LD breakdown over the same interval, flanked on either side by regions wherein markers are in association. This profile is typical for a 'classical' recombination hot spot, as discussed in Chapter 4. The target interval was next subjected to high-resolution sperm CO analysis, to determine whether the region contained an active hot spot and if so, to define the morphology of the hot spot at sub-kilobase level.

5.2.3. High-resolution sperm analysis

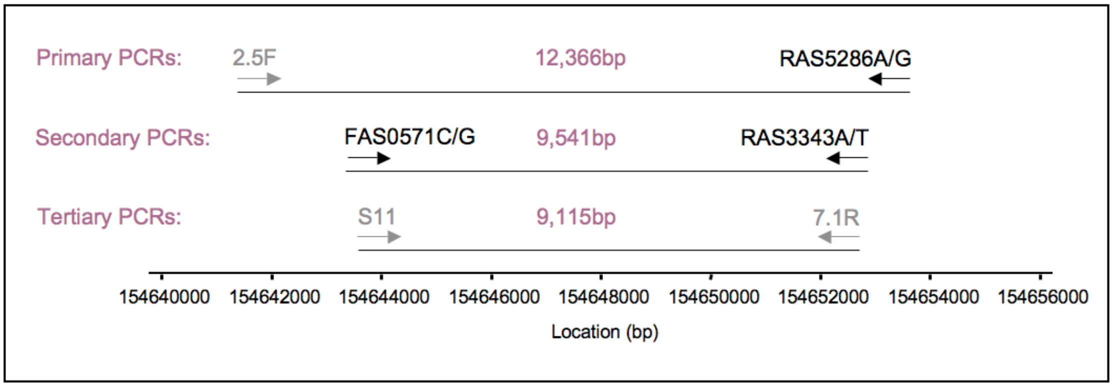
First, semen donors amenable to CO assays were short-listed. These donors carried at least one informative marker distal and proximal to the interval showing LD breakdown, as well as a good density of informative markers across the potential assay interval itself. Amounts of single-molecule clean sperm DNA available for a given donor was also taken into account, prior to making final selections.

Second, feasibility of using various selector sites was tested. As in chapter 4, the LDU profile across the region was used to guide the initial choice of selector sites. Three candidate SNPs 5′ to the LDU step, namely, rs6655071G/C, rs57915757C/T and rs5940571G/C, and three SNPs 3′ to the step, namely, rs17653343A/T, rs9645286A/G and rs306887C/T, were short-listed as potential selector sites. These SNPs were specifically chosen owing to the relatively high frequency of men heterozygous for these markers.

Third, ASPs directed towards these SNPs were designed. On average ASPs were 18bp long, with the variant allele positioned at the last 3′ base. All ASPs were tested over a range of temperatures, typically, 56°C, 58°C, 60°C, 62°C and 64°C, for optimal efficiency and specificity, as described in detail in chapter 3. Whilst all three ASPs 3′ of the LDU step could be optimised to the required extent, only one of the three ASPs 5′ of the step produced similar results. Due to this, the strategy for CO assays was modified as follows.

In primary PCRs, 3′ reverse-ASP RAS5286A/G, designed against rs9645286, was used in combination with universal primer 2.5F. This effectively separated the two parental haptotypes. Primary products were then used to seed secondary PCRs. In this round, ASPs FAS0571C/G and RAS3343A/T, designed against rs6655071 and rs17653343 respectively, were used in repulsion-phase in order to selectively amplify recombinant molecules from separated haplotypes. Following this, PCR products were re-amplified using universal primers S11 and 7.1R. Relative positions of primers are shown in the figure below.

Figure 5.7 Strategy for PAR2 sperm CO assays

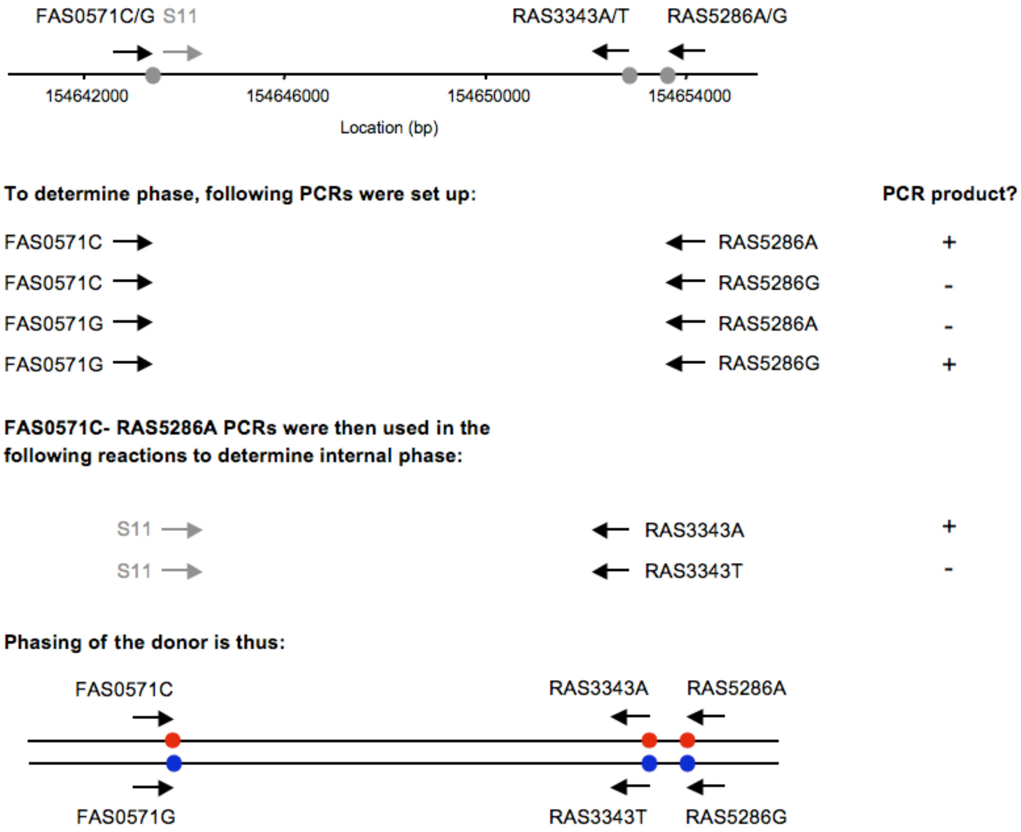

Primary PCRs:
2.5F
12,366bp
RAS5286A/G
Secondary PCRs:
FAS0571C/G
9,541bp
RAS3343A/T
Tertiary PCRs:
S11
9,115bp
7.1R
154640000
154642000
154644000
154646000
154648000
154650000
154652000
154654000
154656000
Location (bp)

Primary, secondary and tertiary PCR amplicons used in the sperm assay. ASPs are shown in black and universal primers are in grey. Sizes of PCR products are indicated in bp above each amplicon. All coordinates are as per chrX, NCBI36/hg18.

Fourth, selected sperm donors were phased for the sperm assay described above. To determine linkage phase, the primary reverse ASP was used in combination with the secondary forward ASP. This was done in order to avoid generating phasing amplicons which encompass both forward and reverse primary selector sites and can thus, potentially contaminate subsequent assays. Positive PCRs were used to phase the secondary reverse ASP, as shown in Figure 5.8.

Primer sequences and PCR conditions used in these experiments are listed in Appendix 5.3. All donors tested had the same phasing for both reverse ASPs, that is RAS5286A-RAS3343A and RAS5286G-RAS3343T. However, linkage phase of FAS0571C/G in relation to these, varied from donor to donor.

Figure 5.8 Linkage phasing for a sperm CO assay


FAS0571C/G S11
RAS3343A/T
RAS5286A/G
154642000
154646000
154650000
154654000
Location (bp)
To determine phase, following PCRs were set up:
PCR product?
FAS0571C
RAS5286A
+
FAS0571C
RAS5286G
-
FAS0571G
RAS5286A
-
FAS0571G
RAS5286G
+
FAS0571C- RAS5286A PCRs were then used in the following reactions to determine internal phase:
S11
RAS3343A
+
S11
RAS3343T
-
Phasing of the donor is thus:
FAS0571C
RAS3343A
RAS5286A
FAS0571G
RAS3343T
RAS5286G

ASPs are indicated in black, whilst universal primers are in grey. RAS5268A/G is the external (primary) ASP used in the assay, while FAS0571C/G and RAS3343A/T are internal (secondary) ASPs. Red and blue colours were assigned arbitrarily in order to distinguish the two haplotypes.

Finally sperm CO assays were carried out in seven men, as described in Figure 5.7. Primer combinations and PCR conditions used for amplification are listed in Appendix 5.3. Tertiary PCR products from the assay were dot blotted and allelic status of internal SNPs determined using ASO hybridisation to dot blots. CO breakpoints were mapped and events scored as described in Chapter 3, sections 3.2.6 and 3.2.7.

5.2.4. Frequency and distribution of COs across the assay interval

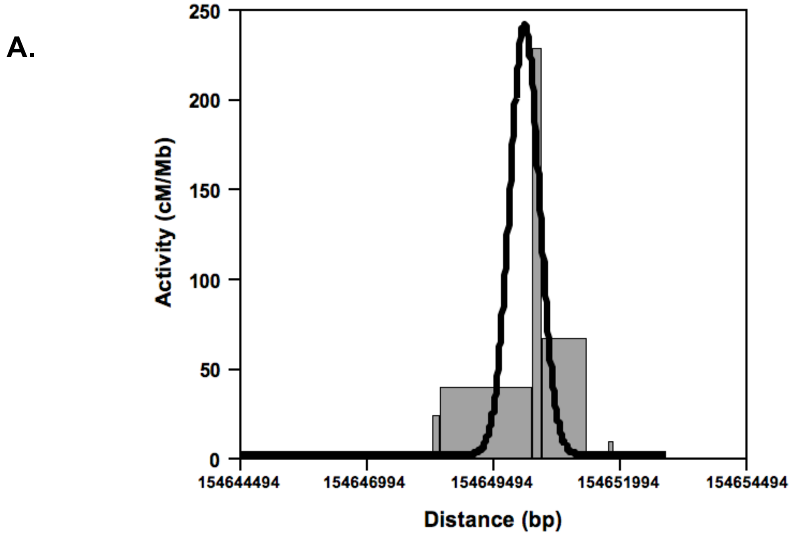
Sperm assays in seven men revealed that COs in this interval cluster to form a hot spot, wherein CO-breakpoints appear to be normally distributed. Five of these men were analysed using reciprocal CO assays and the remaining two men were assayed in one orientation only. For all men, at least a hundred COs were mapped per orientation. In total 2,357 COs were scored from 428,189 molecules. Figure 5.9 shows the distribution of events across the hot spot in all seven semen donors. Interestingly, CO frequencies varied amongst the men analysed, ranging from 0.176% for donor 27 to 0.914% in donor 66, as shown below. This is explored in more detail in Chapter 6.

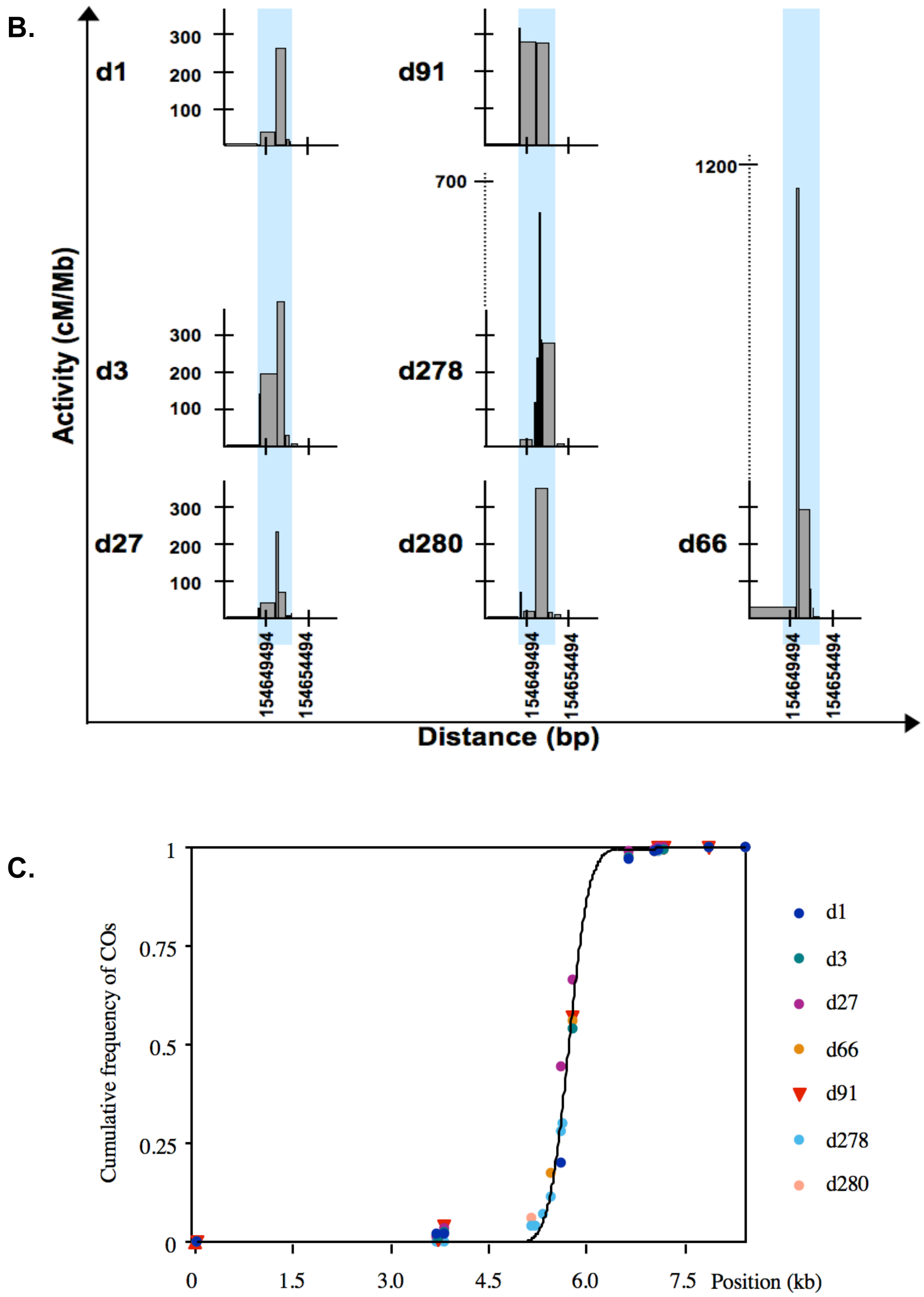
Table 5.1 Data summary from PAR2 CO assays

Table shows the number of molecules screened, corrected COs, recombination frequency, 95% width of the hot spot and peak activity for the seven men assayed. Donors 1, 3, 27, 66 and 91 were assayed in both orientations and values listed below represent the average of both orientations. Donors 278 and 280 were only assayed for one orientation. For details on how these values were calculated, see Chapter 3, section 3.2.7.

Donor	No. of mols.	Corrected COs.	RF (%)	Hot spot centre	95% width, bp	Peak, cM/Mb
d1	69500	587	0.845	154,650,280	997	1343
d3	41600	320	0.756	154,650,190	1830	659
d27	165570	292	0.176	154,650,100	1170	240
d66	62100	568	0.914	154,650,190	1150	1252
d91	49260	362	0.734	154,650,130	2120	553
d278	22340	127	0.567	154,650,270	1290	700
d280	17840	101	0.567	154,650,260	1660	544

Figure 5.9 Distribution of COs across the assay interval





A. Distribution of COs across the assay interval in d27. The bell–shaped graph shows the underlying distribution of COs, assuming a normal distribution of CO-breakpoints. This is consistent with the region harbouring a conventional recombination hot spot. Figure only shows data for d27, but the same is true for the rest of the men (not shown).

B. CO activity for each interval (in cM/Mb) is plotted against distance (in bp). Values shown for d1, d3, d27, d66 and d91 represent the average of two orientations. There is good consistency amongst men in terms of overall distribution of events, as indicated by blue highlights.

C. Cumulative frequency distribution of COs is shown for all seven men. The black line represents the best-fit cumulative normal distribution of COs across the hot spot.

In addition to variation in CO frequency and peak activity, hotspot centres and 95% width also showed slight differences amongst donors. Nonetheless, mean hotspot centre (using data from seven men) was calculated at 154,650,203 (Chr X, NCBI36/hg18), whilst average hotspot width was 1,460 bp (extending from 154,649,473 to 154,650,933). The latter is very similar of the width of most autosomal hotspots and suggests similarities in underlying recombination processes in the two genomic regions.

5.3. Discussion

This work represents the first in-depth analysis of recombination within the Xq/Yq PAR or PAR2. Examination of HapMap Phase II historical recombination rates across PAR2 initially suggested that most recombination in this region occurred within the proximal 95 kb. Independent coalescent analysis of HapMap Phase II genotypes using a more stringent set of controls further narrowed this estimate and suggested that most recombination events occurred within an ~16 kb wide interval. This peak in historical recombination, showing activities up to ~32 cM/Mb, extended from ~ 154643-154660 kb (ChrX, NCBI36/hg18; see Figure 5.2).

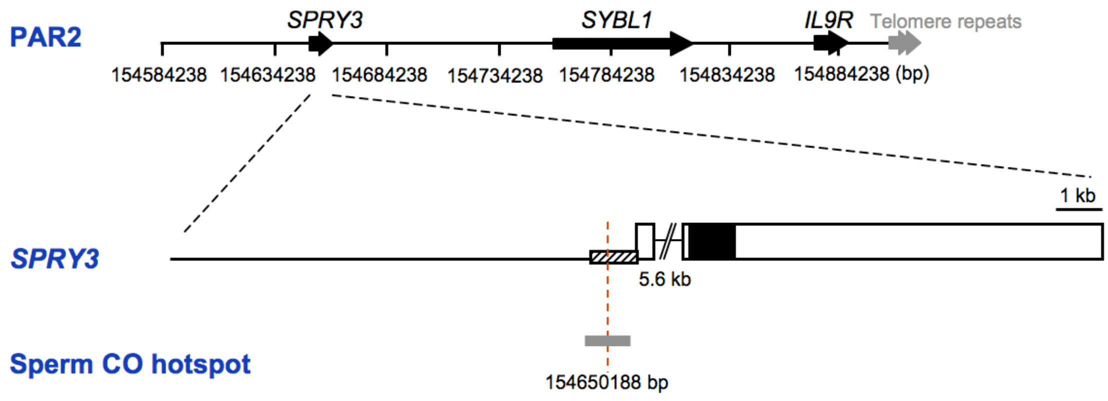
High resolution mapping of exchanges previously reported in CEPH pedigrees further confirmed highly localised COs in PAR2, with at least 3 out of 5 exchanges reported by Freije *et al.* (1992) and Li & Hamer (1995) overlapping the aforementioned peak in historical recombination activity. The exchange event reported in family 1377 was subsequently mapped to an interval of 31 kb, extending from 154648-154679 kb by (C.A. May and A. Veselis). However, no exchanges were detected in family 1416 from analysis across an ~190 kb interval (154648-154838 kb), extending up to SDF-1.

High-resolution LD analysis over a 15 kb interval encompassing this peak in historical activity revealed strong LDU steps in both North European and African semen donors and sperm assays confirmed the presence of an active CO hot spot in this region. This hot spot shows many similarities to its autosomal and PAR1 counterparts, as discussed below.

5.3.1. Genomic context of the hot spot identified

The hot spot identified maps within a putative promoter sequence for the *SPRY3* (sprouty homolog 3) gene (De Bonis *et al.,* 2006) and is henceforth referred to as the *SPRY3* hot spot (see Figure 5.10). Sperm CO hot spots located within promoter sequences have previously been identified on autosomes. Examples of such hot spots include *DNA1* within the MHC class II region and the β-globin hot spot on chr11 (Jeffreys, Kauppi & Neumann, 2001; Holloway, Lawson & Jeffreys, 2006). Collectively, these hot spots are reminiscent of α-hot spots in yeast, which preferentially occur within promoter sequences (see Chapter 1, section 1.3.2.1). Generally however, hot spots occurring within promoter sequences are uncommon in humans. In fact, as discussed in Chapter 1, data from genome-wide surveys suggest that most hot spots tend to avoid genes and recombination rates usually tend to increase with increasing distances from nearby genes (Myers *et al.,* 2005; Coop *et al.*, 2008).

Figure 5.10 PAR2 sperm CO hot spot lies within the *SPRY3* gene


PAR2
SPRY3
SYBL1
IL9R
Telomere repeats
154584238
154634238
154684238
154734238
154784238
154834238
154884238 (bp)
1 kb
SPRY3
5.6 kb
Sperm CO hotspot
154650188 bp

Relative position of *SPRY3* within PAR2 and a magnified schematic of the gene is shown. It contains two exons, indicated by rectangular bars. The striped and black bars represent the promoter and coding region respectively. The position of the sperm CO hot spot is shown (and is to scale with respect to the gene schematic). The hot spot centre is indicated by a red dotted line. Note: The illustration is based on a figure from De Bonis *et al,* 2006).

5.3.2. Morphology of the *SPRY3* hot spot

Data from seven men assayed were consistent with presence of an ~1 kb- wide CO hot spot, with the centre positioned at ~154650 kb (ChrX, NCBI36/hg18). Overall, properties of the *SPRY3* hot spot were very similar to its autosomal and PAR1 counterparts, including a hot spot width of 1-2 kb and quasi-normal distribution of CO breakpoints. These commonalities amongst hot spots imply that the same underlying recombination processes most likely operate across the human genome.

CO frequencies as high as 0.914% recorded at this hot spot, make *SPRY3* not only the most active pseudoautosomal hot spot characterised to date, but also one of the most active hot spots identified in the human genome in general. Interestingly, men assayed at the hot spot showed a fair degree of variation in CO frequencies, ranging between 0.176 and 0.914% (median = 0.734%; mean = 0.652%).

Polymorphism in CO frequencies amongst men is a relatively common phenomenon and has been well documented at several autosomal hot spots (see Chapter 3, section 3.3). Work described in preceding Chapters indicates that such variations are likely to be a feature of PAR1 hot spots as well. Underlying causes of variation in CO activity however, tend to vary from one hot spot to another (see Chapter 3, section 3.3). Thus, for hot spots *NID1* and *DNA2*, variation in activity is determined by status at a single central SNP (Jeffreys & Neumann, 2002; 2005). On the other hand, the ~75-fold variation in activity observed at hot spot MSTM1b appears to be independent of local sequence features (Neumann & Jeffreys, 2006). The present survey of recombination at the *SPRY3* hot spot is too limited to ascertain the true extent of variation in activity and to investigate factors that influence the same. The work in the following Chapter aimed to explore this in detail.

Chapter 6. Factors influencing CO activity at the PAR2 *SPRY3* hot spot

6.1. Introduction

Recent years have witnessed a steady progress in our understanding of the features of human recombination hot spots; however, factors influencing the location and activity of these hot spots have remained somewhat elusive. To date, studies indicate two different levels of hot spot regulation, namely, *cis*- and *trans*-regulation. The former relates to local DNA sequences that influence recombination at the level of individual chromatids, whilst the latter refers to genetic loci located away from hot spots that influence their activity. In the following paragraphs, identification of these factors and their known and predicted effects are explored.

6.1.1. *Cis*-regulation at hot spots

Genome-wide linkage analysis in human pedigrees and large-scale surveys of patterns of LD have revealed many genomic features that show some correlation with location of hot spots. These include the GC-content of a genomic region and the presence of certain repetitive elements, and have been discussed in detail in Chapter 1, sections 1.5.2 and 1.5.3. Here, data on certain sequence motifs that have been associated with a substantial number of LD hot spots are briefly recapitulated and described in further detail.

The first evidence supporting association of a particular sequence motif with substantial numbers of LD hot spots came from Myers *et al*. (2005), who compared recombinationally hot and cold regions in order to identify sequence features enriched exclusively in the former. This led them to a sequence CCTCCCT, which was associated with 11% of the LD hot spots included in the study, as previously discussed in Chapter 1.

Further evidence for influence of this motif on recombination activity came from sperm analysis at the *DNA2* hot spot (Jeffreys & Neumann, 2002). Polymorphism in CO activity and TD accompanying COs at the *DNA2* hot spot were linked to status at a single SNP located ~70 bp away from the centre (Jeffreys & Neumann, 2002; see Chapter 3.3). Examination of the sequence context of this SNP revealed that it was present at the third position of the CCTCCCT motif (Myers *et al*., 2005). Intriguingly, it was the motif-disrupting allele at this SNP that suppressed CO activity in *cis* at the *DNA2* hot spot. A similar phenomenon was identified at the *NID1* hot spot, wherein a SNP allele disrupting the CCCCACCCC motif, associated with 3% of LD hot spots studied by Myers *et al*. (2005), also suppressed recombination activity in *cis* (Jeffreys & Neumann, 2005).

In addition to CO hot spots, these sequences motifs were enriched within many unstable minisatellite sequences (Myers *et al*., 2005). For example, each repeat unit of the highly unstable CEB1 minisatellite contained between one and three copies of the CCTCCCT motif (Buard & Vergnaud, 1994). Unstable minisatellites such as CEB1 expand largely through inter-allelic transfer of repeat units, mostly during the recombinational repair of meiotic DSBs. Enrichment of these motifs within both minisatellites and CO hot spots hinted towards a common role for them in these recombinational processes.

More recently, Myers *et al*. (2008) identified a more degenerate 13-mer motif, CCNCCNTNNCCNC from analysis of LD hot spots inferred from Phase II HapMap data (2007). A relative of the previously-described 7-mer motif, this motif was associated with as many as 41% of LD hot spots studied. It also existed in the vicinity of the estimated centres of hot spots *DNA3* and *DMB2* in the class II region of the MHC, as well as the MS32 hot spot on Chr1 (Jeffreys *et al.,* 2001; Jeffreys *et al.,* 1998). Furthermore, a C/G polymorphism that was linked to CO activity at the MS32 hot spot, was located just 1 bp-away from this 13-mer motif. Myers *et al*. (2008) suggested that reduced CO frequencies associated with the C allele of the SNP were conferred by virtue of its position in relation to the motif.

Furthermore, this 13-mer motif was present within hot spots of non-allelic homologous recombination, such as those associated with Charcot-Marie-Tooth disease type 1A and hereditary neuropathy with liability to pressure palsies. It was also enriched within hyper-variable minisatellites MS1, CEB1, B6.7 and D7S22 (Myers *et al*., 2008). In summary, the findings of Myers *et al*. (2008) suggested a role for the 13-mer motif in various forms of recombination-associated processes in humans.

The 13-mer degenerate motif was however, present at less than half of the LD hot spots surveyed and therefore, did not account for the activity of remaining hot spots. Furthermore, the motif was present within regions that did not constitute hot spots or show elevated recombination activity, thus suggesting that it was neither necessary nor sufficient for hot spot activity. Considering this, it seemed likely that additional yet-unidentified factors existed that influenced global distribution of recombination events in the human genome.

6.1.2. *Trans*-regulation of hot spots- Evidence from mice

Evidence of *trans*-regulation of mammalian hot spots mainly comes from studies on mouse hot spots, in particular *Psmb9* (see Chapter 1, sections 1.4.2.5 and 1.4.2.6). Data from mouse hot spots, including *Psmb9* and *Hlx1*, highlighted the presence of two overlapping loci, *Dsbc1* and *Rcr1*, on chr17, which seemed to control activity of specific mouse hot spots on chromosomes 1, 15 and 18 (Grey *et al*., 2009; Parvanov *et al*., 2009). Several commonalities between the two loci, including regulation of the same hot spots, indicated that they most probably represented effects of the same gene. In order to identify this gene, Parvanov *et al.* (2010) studied 1580 progeny from crosses between B6 and CAST mouse strains, which differed in activity at *Rcr1/Dsbc1*-regulated hot spots *Psmb9, Hlx1* and *Esrrg-1*. From this study, they were able to map the *trans*-acting regulators to a 181-kb interval containing four different genes, *Psmb1, Tbp, Pdcd2* and *Prdm9.*

Proteins encoded by three of these genes, *Psmb1, Tbp* and *Pdcd2*, showed no differences in amino acid sequences between B6 and CAST mice and thus most likely did not account for the observed differences in hot spot activity between the two strains. However, protein variants encoded by the *Prdm9* gene did differ between B6 and CAST strains. Thus, the PRDM9 variant in B6 mice contained 12 zinc-fingers (ZnFs) in the terminal C2H2 ZnF-domain, whilst the variant in CAST mice carried only 11. Both strains also showed several amino acid substitutions at positions determining DNA-binding specificity of the protein, relative to one another. On account of this, PRDM9 was chosen as the primary candidate gene.

Several other properties of PRDM9 further supported a role for the protein in hot spot specification. First, PRDM9 contains a PR/SET domain responsible for promoting tri-methylation of lysine 4 of histone 3 (H3K4). This chromatin modification has been shown to precede recombination at the *Psmb9* and *Hlx1* hot spots (Buard *et al.,* 2009). H3K4 trimethylation has also been associated with recombination initiation sites in *S. cerevisiae*, consistent with a role for PRDM9 in hot spot specification (Borde *et al*., 2009). Second, studies in *Prdm9* knockout mice have shown reductions in the number of meiotic *Dmc1* foci and lack of co-localisation *Dmc1* with *γ-H2AX* (Hayashi, Yoshida & Matsui, 2005). Furthermore, meiosis was blocked by pachytene in these mice, resulting in sterility in both sexes and azoospermia in males (Hayashi, Yoshida & Matsui, 2005).

Based on the properties of PRDM9, Parvanov *et al*. (2010) concluded that *Dsbc1/ Rcr1* most likely represented the effects of the *Prdm9* gene. At the same time as the Parvanov *et al*. (2010) study, work by Myers *et al*. (2010) and Baudat *et al*. (2010) provided additional evidence suggesting a role for PRDM9 in regulating recombination hot spots in humans (as discussed below).

6.1.3. Regulation of human hot spots by PRDM9

In humans, data from bioinformatic analyses suggested that PRDM9 was most likely the only C2H2 ZnF protein that bound the 13-bp hot spot motif and matched the patterns of degeneracy associated with the same (Myers *et al*., 2010). Also, the predicted binding sequence for the chimpanzee PRDM9 was very different to its human counterpart. This was consistent with the lack of enrichment of the 13-mer motif amongst LD hot spots in chimpanzees and provided a basis for the absence of hot spot sharing between the two species (Myers *et al*., 2010).

Additional evidence suggesting a role for PRDM9 in influencing LD hot spot usage came from analysis of sequences encoding the ZnF array of PRDM9 in a panel of CEPHs and Hutterites (Baudat *et al*., 2010). First, the analysis revealed variation in sequences encoding the PRDM9 ZnF array, with six and three different *PRDM9* alleles identified amongst these two sample sets respectively. The most common allele in both populations, occurring at frequencies of 90% and 94% respectively amongst CEPHs and Hutterites individuals, was the A allele, which encoded the protein variant predicted to bind the 13-mer hot spot motif. Alleles B, C, D, E and F were also identified amongst CEPHs. The protein encoded by allele B was identical to the A variant, except for one amino acid change in the 6th ZnF. Alleles C, D, E and F encoded proteins showing several changes with respect to the A variant in ZnFs 8-11. Most changes affected positions -1, 2, 3 and 6 of the ZnF α-helices, which determined DNA-binding specificities of the protein. Alleles A and B were also identified amongst Hutterites, in addition to a third allele I, which occurred at a frequency of 2%.

The protein encoded by allele I showed several changes in its ZnF array, such that it was not predicted to bind the 13-mer hot spot motif. Baudat *et al*. (2010) showed that A/I heterozygous males and females showed significant reductions in usage of LD hot spots, relative to individuals carrying two A alleles. This observation supported a relationship between *PRDM9* status and hot spot usage in humans. Baudat *et al*. (2010) suggested that the I-allele might trigger a different set of hot spots, which have yet to leave their mark on population diversity data.

Finally, in order to demonstrate that the aforementioned influence of *PRDM9* on hot spot usage was mediated via binding of the protein to hot spot sequences, Baudat *et al.* (2010) studied the interaction between PRDM9 A and I variants and their predicted binding motifs (determined using the Zinc Finger Database). They were able to demonstrate that the PRDM9 A variant showed high affinity for the 13-mer hot spot motif *in vitro*, but low affinity towards motifs carrying mutations in highly conserved bases, as well as to the predicted binding motif for the protein encoded by allele I. Conversely, the I variant was specific for the predicted I motif and showed poor affinity towards the 13-mer hot spot motif.

Collectively, these studies have defined an entirely new level of regulation of meiotic recombination in humans and mice. It is worth emphasising however, that the Baudat *et al.* (2010) study focussed entirely on genome-wide *LD* hot spot usage in humans. It was therefore interesting to see how variation in *PRDM9* might affect CO activity at individual sperm CO hot spots. A local effort was thus initiated to investigate the impact of *PRDM9* variation on recombination activity in a panel of 10 CO hot spots. A substantial part of the work described in this Chapter is part of this larger survey (Berg *et al*., 2010).

6.1.4. This work

Sperm analysis in seven men revealed considerable variation in CO frequencies at the PAR2 *SPRY3* hot spot. Work described in this Chapter was aimed at investigating whether any *cis*-acting factors, such as motif-disrupting SNPs, or *trans*-regulators of recombination, specifically *PRDM9,* accounted for the observed variation in hot spot activity. In order to study the potential influence of any such factor(s) and their relative contributions, the survey of CO activity was extended further and a total of 28 men were assayed, as described in the following sections.

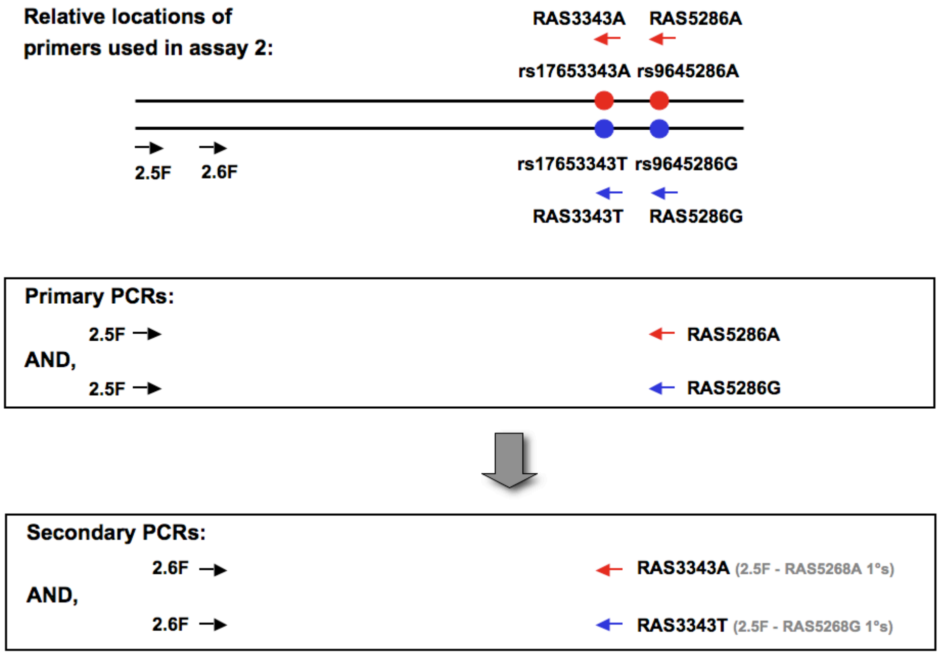
6.2. Results

6.2.1. Strategy for extended survey of recombination at the *SPRY3* hot spot

In order to investigate the true extent of variation in CO activity and to identify potential *cis*- and *trans*-acting factors influencing this, additional men were assayed at the *SPRY3* hot spot. Men carrying different *PRDM9* alleles were intentionally assayed, in order to effectively test for any effects of this putative *trans*-regulator of recombination (discussed further in later sections of this Chapter). Two different types of sperm analyses were used in this survey. The first has already been described in Chapter 5 and was used to assay donors heterozygous at markers rs5940571G/C, rs17653343A/T and rs9645286A/G, which were used as selector sites. This assay will henceforth be referred to as ‘assay 1’ for purposes of clarity. CO activities for a total of 14 men were determined using this assay. Another 12 men were studied using a different sperm assay, henceforth referred to as ‘assay 2’. This assay enabled high-resolution analysis of both COs and NCOs, as shown in Figure 6.1. This type of sperm assay was previously used to characterise hot spot PAR1-P described in Chapter 4.

In this instance, the assay was used to study men who were heterozygous at 3' selector sites, rs17653343A/T and rs9645286A/G, but homozygous at the 5' selector site, rs5940571G/C, used in assay 1. First, donors that had been short-listed for the assay on the basis of density of informative markers and availability of single-molecule clean DNA, were phased using the same strategy as described in Chapter 4, section 4.2.2.3. All donors had the same phasing at selector sites, which is predictable given their close proximity within a region of strong marker association. In the sperm assay itself, two rounds of allele-specific PCRs were used to selectively amplify individual haplotypes, as shown in Figure 6.1. The DNA inputs for these assays were limited to 55 molecules per pool, although inputs as low as 23 molecules per pool were used in some cases. PCR conditions used in this assay are listed in Appendix 6.1.

Figure 6.1 Strategy for 'assay 2'


Relative locations of primers used in assay 2:
RAS3343A
RAS5286A
rs17653343A
rs9645286A
2.5F
2.6F
rs17653343T
rs9645286G
RAS3343T
RAS5286G
Primary PCRs:
2.5F
RAS5286A
AND,
2.5F
RAS5286G
Secondary PCRs:
2.6F
RAS3343A (2.5F - RAS5268A 1°s)
AND,
2.6F
RAS3343T (2.5F - RAS5268G 1°s)

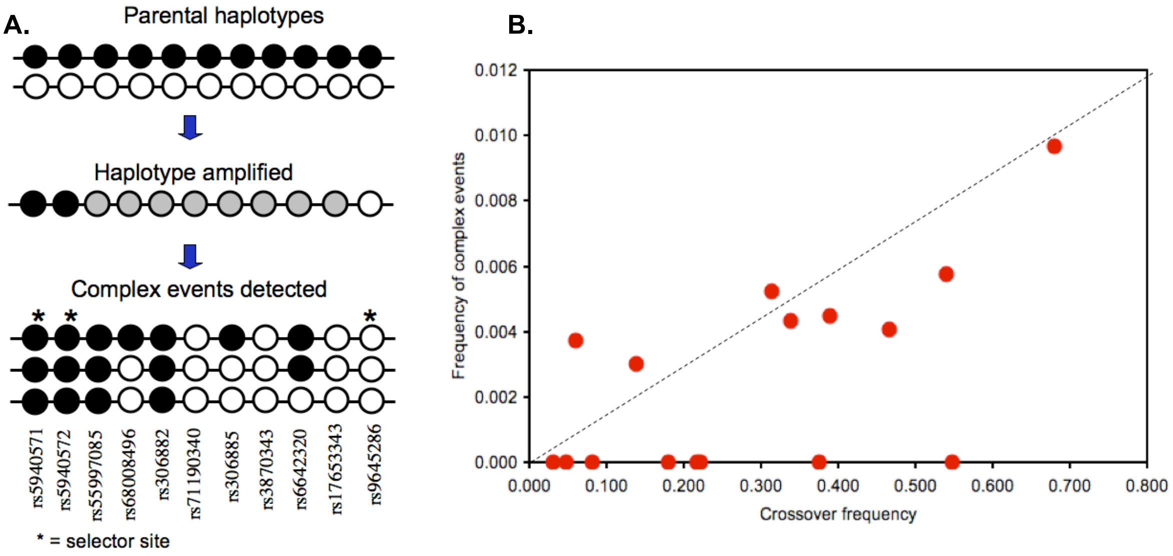
Primer combinations used to amplify constituent semen donor haplotypes in assay 2. Positions of selectors sites are indicated using circles. ASPs are indicated using red/ blue arrows positioned above corresponding selector sites. Universal primers are indicated using black arrows. Primary PCR products were 12,366 bp long and generated using 2.5F with either RAS5285A or RAS5286G. Secondary products were 11,574 bp long and generated using 2.6F and RAS3343A/T. Primaries used to seed different secondary PCRs are indicated in the figure. All donors studied have the same phasing at selector sites.

Secondary PCR products were re-amplified using the same primer combinations, as used in assay 1 (see Chapter 5, section 5.2.3) to increase yield of PCR products, which were subsequently dot-blotted. Dot blot filters containing products of one haplotype were then probed with oligos specific to the allele of the unselected haplotype, in order to detect recombination events. Two men, d27 and d66, were analysed using both types of assays in order to test for any significant differences between them. Neither man showed any significant differences in data obtained from the two assays, as determined using the Fisher's exact test (2-tailed *P*=0.272 and 0.677, for d27 and d66 respectively).

In total, thirteen North European and fifteen African men were assayed across the hot spot. Sixteen men were assayed in both orientations and none of them showed any significant differences in CO frequencies between the two orientations ($P>0.1$). Given this, remaining men were assayed in one orientation only. Between 10,010 and 191,170 molecules were screened per man and recombination events thus recovered were scored and analysed as described in previous chapters.

Data from all donors were consistent with the presence of a recombination hot spot, with quasinormal distribution of CO breakpoints. Greater than 99% of the COs analysed were simple, with exchanges mapping to a single interval between heterozygous SNPs. This was consistent with previous data from autosomal hot spots (Webb, Berg & Neumann, 2008). All complex events, wherein intervals of exchange could be unambiguously mapped, extended out of the CO hot spot (example shown in Figure 6.2A). Such complex events, with multiple exchange points mapping out side of the hot spot, have been previously documented in mismatch repair deficient *Mlh1*$^{-/-}$ mice (Guillon *et al.,* 2005). It has previously been suggested that the frequency of complex events increases as the frequency of COs increases (Webb *et al*., 2008). However, no significant relationships were detected between CO frequencies and the frequencies of complex events in the present study ($P>0.05$, Mann-Whitney U test), as shown in Figure 6.2.

Figure 6.2 Relationship between hot spot activity and the frequency of complex events



A. Complex events recovered from d1. Data from a single orientation are shown; distances between informative SNPs are not to scale. Selector sites are indicated by asterisks.

B. Relationship between CO frequency and the frequency of complex events is shown for all donors for whom complex events could be unambiguously interpreted. The dotted line indicates the expected relationship, as suggested by Webb, Berg & Jeffreys (2008).

6.2.2. Extreme variation in CO frequencies at the *SPRY3* hot spot

Data from the 28 men thus studied revealed large variations in sperm CO frequencies, which ranged from 0% to 0.914%. At the time of the study, this presented the greatest variation documented at a human sperm hot spot. Table 6.1 summarises CO data for all men, including numbers of molecules screened, total number of COs analysed, CO frequencies (%) and corresponding lower and upper confidence limits.

Table 6.1 CO frequencies of men assayed at the *SPRY3* hot spot

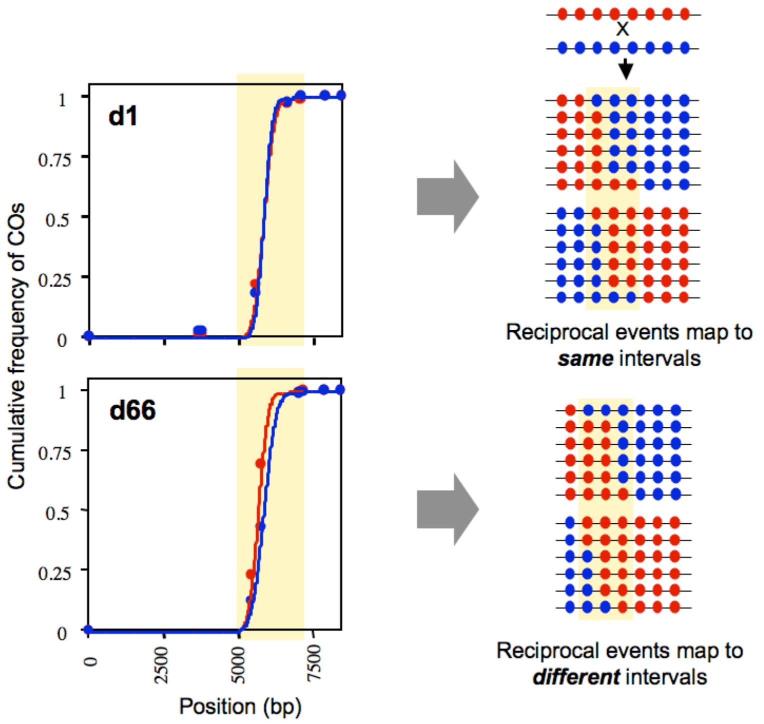
Note: d1, d3, d5, d23, d25, d27, d39, d66, d72, d87, d91, d178, d240, d244, d252, d253, d274 and d279 were analysed in both orientations and CO frequencies listed below represent data from both orientations. For donors 27 and 66, data from only assay 1 is shown since there are no significant differences in data from the two assays for these men and assay 1 constitutes the larger and more robust dataset of the two. Men are listed in ascending order of CO frequencies.

Donor name	No. of mols. screened	Corrected COs	CO Frequency (%)	Lower 95% C.I.	Upper 95% C.I.
d254	14181	0	0	0.000	0.027
d252	26399	2	0.008	0.001	0.027
d279	10010	2	0.02	0.002	0.072
d5	138741	45	0.032	0.024	0.043
d240	16518	6	0.036	0.013	0.079
d72	19149	9	0.047	0.021	0.089
d60	39359	21	0.054	0.033	0.082
d282	26916	17	0.064	0.037	0.101
d87	28927	18	0.067	0.037	0.098
d23	13277	9	0.068	0.031	0.129
d250	18557	16	0.087	0.050	0.140
d178	12758	14	0.109	0.061	0.184
d39	11906	19	0.162	0.098	0.251
d27	165568	292	0.176	0.157	0.198
d246	19458	40	0.207	0.148	0.281
d251	18918	47	0.251	0.184	0.331
d184	19818	54	0.271	0.203	0.352
d274	23001	67	0.283	0.226	0.368
d25	20361	72	0.349	0.279	0.446
d244	17677	69	0.378	0.305	0.494
d277	23061	113	0.488	0.402	0.585
d232	19073	94	0.495	0.400	0.603
d278	22341	127	0.567	0.473	0.673
d280	17837	101	0.567	0.462	0.687
d91	49260	362	0.734	0.660	0.812
d3	41633	320	0.756	0.687	0.857
d1	69445	587	0.845	0.778	0.915
d66	62106	568	0.914	0.841	0.991

6.2.3. Heterozygosity at SNP rs700442 causes reciprocal CO asymmetry and transmission distortion accompanying COs

Data from all men assayed in both orientations showed that COs were always reciprocal in rate. However, CO breakpoint mapping showed that they were not always reciprocal in distribution, with some men showing reciprocal CO asymmetry and significant transmission distortion (TD) of SNP alleles accompanying COs. The figure below illustrates this using CO distribution in d1 and d66 as examples.

Figure 6.3 Reciprocal CO asymmetry at the *SPRY3* hot spot



The cumulative frequency distribution of COs in each orientation in d1 and d66. Red and blue colours have been arbitrarily assigned to the two orientations. Illustrations are shown alongside graphs to depict the transmission of hot spot alleles in the two donors. Informative SNPs are indicated using red or blue circles. Note: positions on graph are relative to the assay interval for assay 1, wherein the first base corresponds to 154644494 bp (NCBI36/hg 18, chr X).

In d1, the cumulative frequency distributions of COs from both orientations mirror each other and reciprocal COs map to the same intervals (Figure 6.3). This is consistent with 50:50 transmission of hot spot alleles from a particular haplotype to the recombinant progeny. In d66 however, cumulative frequency distributions of COs from the two orientations are displaced by ~200 bp at central markers. This indicates that reciprocal COs map to different intervals, thereby leading to over-transmission of hot spot alleles from the red haplotype into the recombinant progeny. This deviation from the 50:50 transmission ratio, or TD, was strongest (66:34) at markers closest to the hot spot centre.

Analysis of haplotypes of all semen donors studied (listed in Appendix 6.2) showed that TD accompanying COs was only seen in men heterozygous for SNP rs700442C/T (abbreviated 0442), located ~ 45 bp away from the hot spot centre. Men homozygous at this marker, such as d1, showed no significant distortion in transmission of hot spot alleles. Figure 6.4 illustrates this further, using data from d1 and d3, who were T/T and C/C homozygous at 0442 respectively and also from d27, d66 and d91, who were C/T heterozygous for the SNP. Significant TD was only observed in 0442 heterozygotes, with the 'C' allele being consistently over-transmitted ($P \leq 0.0004$).

Alleles linked to 0442C also showed significant TD ($P \leq 0.002$). Thus, the C allele at rs306882, located 164 bp away from 0442, was significantly over-transmitted in d27. The same is true for the rs56750753 G allele in d66, located some 312 bp away from 0442. These SNPs however, did not show any significant TD in 0442 homozygous men, thereby indicating that they were most likely passively being swept into biased GC tracts associated with 0442.

Curiously, the G allele at rs67804731 (or 4731), which lies on the opposite haplotype relative to 0442C, showed significant over- transmission in d91 ($P=0.0005$). 4731 was located nearly 2 kb-away from 0442 and did not fall within the *SPRY3* hot spot, as defined in Chapter 6. Given that the two SNP alleles showing significant over-transmission lay on opposite haplotypes, this phenomenon could not be accounted for by kb-long conversion tracts accompanying COs. Whilst it remains formally possible that over-transmission of the G allele at 4731 was the result of biased heteroduplex repair at this site, this was unlikely given the fact that the phenomenon was not observed in any other man. In fact, d5 showed slight over-transmission of the A, rather than G allele at 4731, although this was not significant ($P=0.146$).

Figure 6.4 Heterozygosity at 0442 is associated with TD in COs

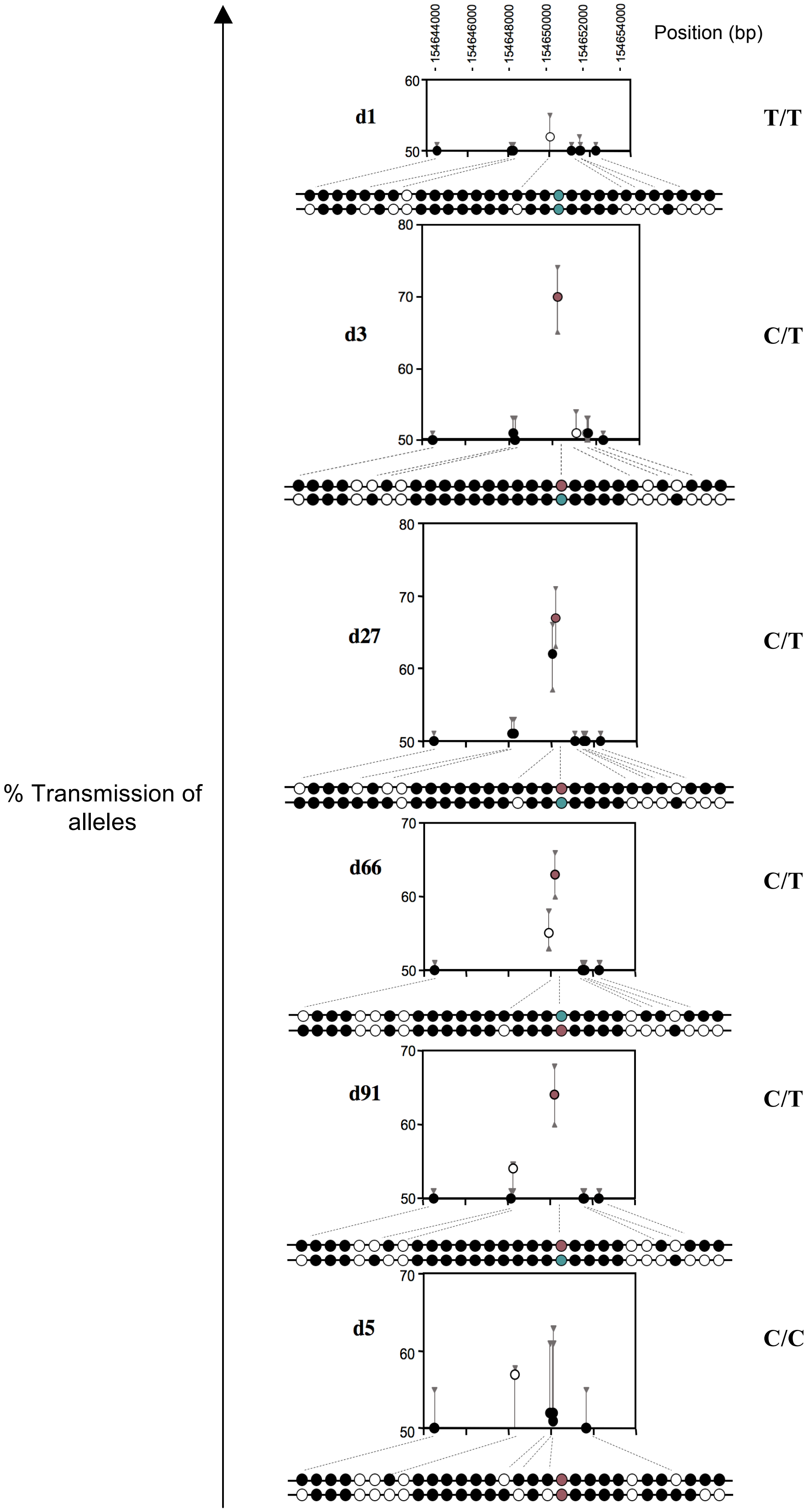


Figure 6.4. Transmission frequencies of hot spot alleles to CO progeny are shown for d3, d27, d66 and d91. 0442 status (C/C, C/T or T/T) of each donor is also shown alongside respective graphs. 95% confidence intervals on transmission frequencies are indicated using vertical bars. Donor haplotypes over the assay interval are indicated. Black and white colours have been arbitrarily assigned to SNP alleles. SNP positions indicated in the haplotypes are not to scale, but in order of position (from centromere to telomere). Last pair of informative SNPs for all men represent 3' selector sites. C allele of 0442 is indicated in pink, whilst the T allele is in cyan. All positions are as per chrX, NCBI36/hg18.

It is important to note that significant TD at 0442, with over-transmission of the C allele, was only observed in men for whom large numbers of COs (typically >100) were scored. Thus, 5 of the 12 0442-heterozygous men showing significant TD were all men for whom large numbers of COs were screened. This association was significant ($P<0.01$) and highlighted the importance of large sample sizes in detection of often subtle, variations in transmission ratios of hot spot alleles. Amongst the remaining 7 men, over-transmission of the 0442 C allele was seen in 4. The remaining 3 of 12 men showed marginal over-transmission of the 0442 T allele. This was however not significant ($P\geq0.226$) and was most likely a consequence of low sample sizes of COs analysed for these men. Transmission frequencies of hot spot alleles for all men analysed in both orientations are listed in detail in Appendix 6.3.

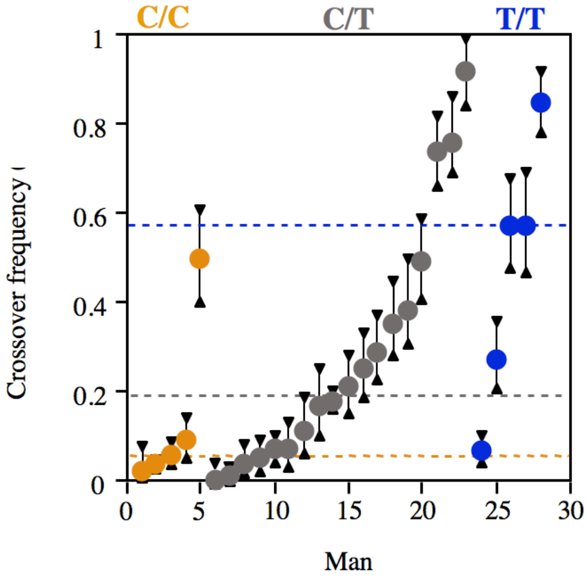
Average transmission ratio of 0442 C:T alleles in COs was calculated using data from 4 men who showed significant TD at the SNP and for whom greater than 100 COs were scored. These men were specifically chosen in order to attain the most robust estimates possible. Thus, data from d3, d27, d66 and d91 showed that average transmission ratio of C:T alleles in COs was 66:34, which corresponded to a conversion bias of 1.94:1 during CO, in favour of the C allele. Hot spot centres for these donors were displaced by an average of 329 bp between the two orientation of COs, an estimate comparable with previous data from *DNA2* and *NID1* hot spots, wherein centres are displaced by 430 bp and 440 bp respectively (Jeffreys & Neumann, 2002; Jeffreys & Neumann, 2005). It is worth noting however, that combined cumulative distributions in these donors correlated perfectly with that of 0442 homozygous men, implying that the overall location of the hot spot remained the same.

Data from autosomal hot spots showing the same phenomenon suggested that TD accompanying COs could be best explained by a difference in recombination initiation frequencies between haplotypes carrying the 0442- C and T alleles, with initiating DSBs occurring more frequently on the haplotype carrying the T allele. Thus, repair of these breaks using information from the haplotype carrying the less active C allele, would lead to over-transmission of this allele in the recombinant progeny. If this model was correct for the *SPRY3* hot spot, then men homozygous for the active T allele should show higher CO frequencies compared to men carrying two copies with the less-active C allele. To test for this, 28 men assayed at the *SPRY3* hot spot were classified according to their 0442 status, in order to test for any differences in mean and median CO frequencies between the three groups, namely, 0442 TT, CT and CC men

6.2.4. 0442 influences CO frequencies *in-cis* at the *SPRY3* hot spot

Of the 28 men assayed at the *SPRY3* hot spot, 5 were C/C homozygous, 5 were T/T homozygous and 18 were C/T heterozygous at 0442. A simple comparison of CO frequencies between C/C, C/T and T/T individuals revealed pronounced effects of 0442 SNP status, with median CO frequencies of C/C, C/T and T/T men being 0.054, 0.192 and 0.567% respectively, as shown in Figure 6.5. Notably, the difference in rankings of 0442 C/C and T/T men was significant (*P*=0.030; Mann-Whitney U test).

Figure 6.5 Variation in CO frequencies between 0442 C/C, C/T and T/T men



CO frequencies for 28 men assayed at the *SPRY3* hot spot. 0442 C/C men are shown in orange, C/T men in grey and T/T men in blue. Men in each group are ranked by CO frequencies. Vertical lines represent 95% confidence intervals. Median CO frequencies for each group are indicated by horizontal dotted lines.

Interestingly, 0442 C/T men showed a wide range of CO frequencies, varying between 0% and 0.914%. Given this degree of variation, it seemed unlikely that 0442 is the only factor influencing recombination activity at this hot spot. Indeed, multiple *cis*- or *trans*-acting factors have been shown to influence recombination activity at the mouse *Psmb9* hot spot, as discussed in Chapter 1, section 1.4.2.5 (Shiroishi *et al.*, 1991; Baudat & de Massy, 2007).

Intriguingly, analysis of the sequence context of 0442 showed that the C allele of the SNP disrupted the hot spot-motif CCNCCNTNNCCNC in the seventh position. Data from Myers *et al.* (2010) suggested that this hot spot-motif most likely served as a binding site for meiosis-specific zinc-finger protein PRDM9, which had been implicated as a regulator and specifier of recombination hot spots in mice, as described in Section 6.1.3 (Parvanov *et al.*, 2010; Myers *et al.*, 2010; Baudat *et al.*, 2010). Variation in the zinc-finger array of PRDM9 had also been associated with shifts in LD hot spot usage in humans (Baudat *et al.*, 2010). Therefore, potential effects of PRDM9 variation on CO frequencies at the *SPRY3* hot spot were investigated.

6.2.5. *Trans*-acting factor PRDM9 profoundly influences frequency of COs at the *SPRY3* hot spot

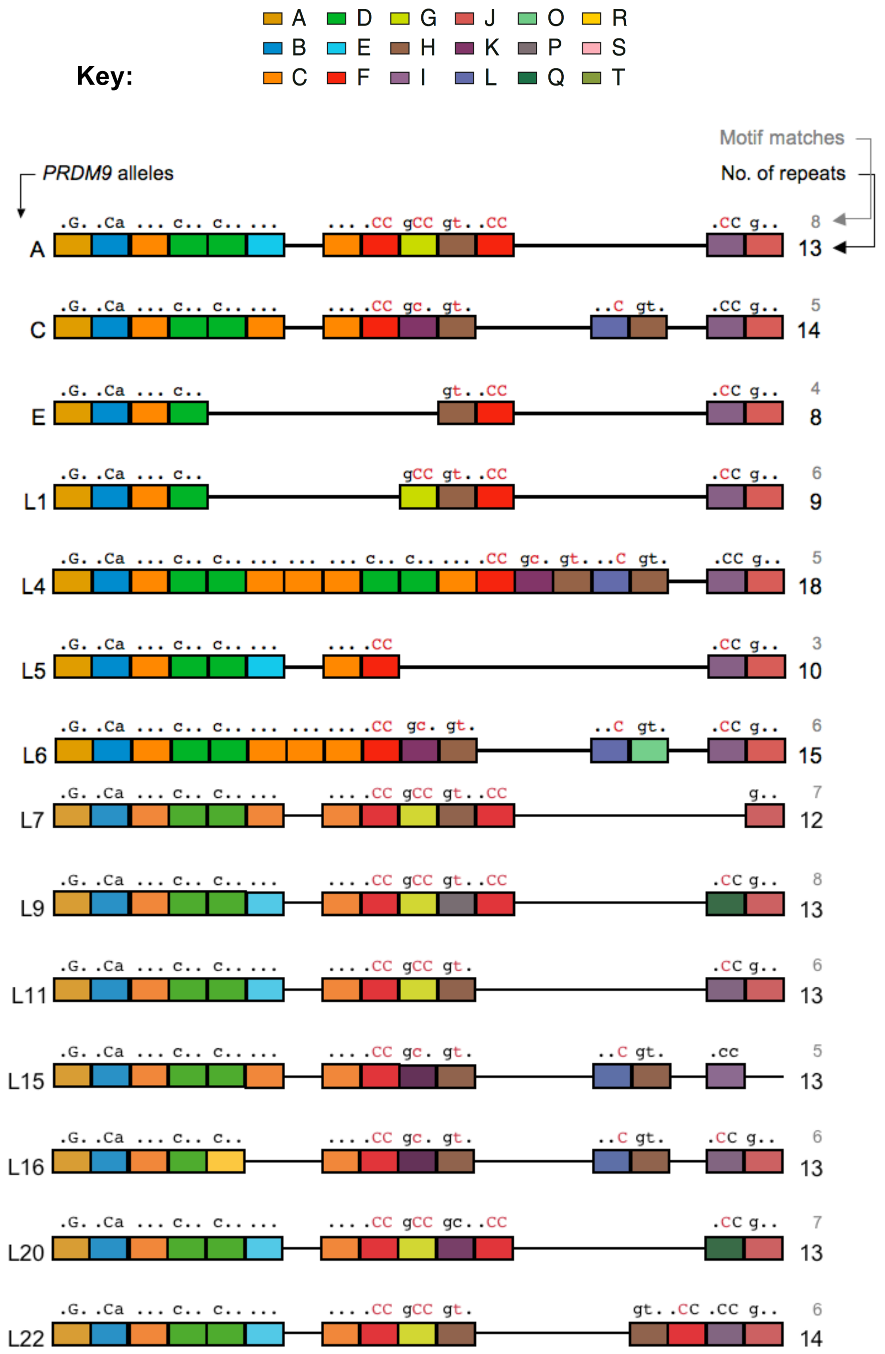
The ZnF domain of PRDM9 is encoded by a minisatellite, with one repeat unit coding for an individual ZnF. To study the influence of variation in the PRDM9 ZnF array on CO activity, minisatellite repeats encoding this array were characterised by R. Neumann, I.L. Berg and A.J. Jeffreys in 74 African and 156 European semen donors. This revealed extreme variation in sequences coding for the ZnF domain, particularly in Africans, with substantial diversity in both numbers and types of repeat units. These repeat units were named A to T; Appendix 6.4). Most of the variation was limited to sequences coding for the DNA-binding residues of PRDM9 ZnFs, that is, positions -1, 2, 3 and 6 of the alpha helix. Variation within sequences coding for other residues were also identified, but were relatively infrequent (see Appendix 6.4).

Overall, this survey revealed *PRDM9* alleles that code for zinc-finger arrays of variable lengths, ranging between 8 and 18 zinc-fingers. *PRDM9* allele A, previously identified by Baudat *et al.* (2010), still emerged as the most common allele. Alleles B-E reported by Baudat *et al.* (2010) were also identified, but alleles F and I were not seen. Previously unidentified *PRDM9* alleles were assigned names L1-L24 (where ‘L’ indicates Leicester). Predicted binding sequences for zinc-finger arrays encoded by various *PRDM9* alleles were determined using the Zinc Finger Database. This showed that most non-A alleles encoded ZnF arrays that had poor predicted ability to bind the 13-mer hot spot motif. Predicted binding sequences for PRDM9 alleles seen amongst donors assayed at the *SPRY3* hot spot are shown in Figure 6.6.

If PRDM9 indeed triggers hot spot activity by binding to predicted sequence motifs then, in theory, men carrying two copies of motif-recognising *PRDM9* alleles, such as the common A allele, should show greater hot spot activity compared to men who carried only one/ no copies of such alleles. Men assayed at the *SPRY3* hot spot were thus grouped according to their *PRDM9* status, as shown in

Table **6.2** and median CO frequencies amongst the three groups, namely, men carrying two A alleles (A/A), one A and one non-A allele (A/N), and finally two non-A alleles (N/N) were compared.

Figure 6.6 *PRDM9* alleles seen in sperm donors analysed at the *SPRY3* hot spot


Key:
A D G J O R
B E H K P S
C F I L Q T
Motif matches
No. of repeats
PRDM9 alleles
A .G. .Ca ... c.. c.. ... ... .CC gCC gt. .CC .CC g.. 8 13
C .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .CC g.. 5 14
E .G. .Ca ... c.. gt. .CC .CC g.. 4 8
L1 .G. .Ca ... c.. gCC gt. .CC .CC g.. 6 9
L4 .G. .Ca ... c.. c.. ... ... c.. c.. ... .CC gc. gt. ..C gt. .CC g.. 5 18
L5 .G. .Ca ... c.. c.. ... ... .CC .CC g.. 3 10
L6 .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .CC g.. 6 15
L7 .G. .Ca ... c.. c.. ... ... .CC gCC gt. .CC g.. 7 12
L9 .G. .Ca ... c.. c.. ... ... .CC gCC gt. .CC .CC g.. 8 13
L11 .G. .Ca ... c.. c.. ... ... .CC gCC gt. .CC g.. 6 13
L15 .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .CC 5 13
L16 .G. .Ca ... c.. c.. ... .CC gc. gt. ..C gt. .CC g.. 6 13
L20 .G. .Ca ... c.. c.. ... ... .CC gCC gc. .CC .CC g.. 7 13
L22 .G. .Ca ... c.. c.. ... ... .CC gCC gt. gt. .CC .CC g.. 6 14

Various *PRDM9* alleles are shown. Alleles A, C and E were identified by Baudat *et al.* (2010), whilst the remaining were characterised by R. Neumann, I.L. Berg and A.J.Jeffreys (Berg *et al.,* 2010). Coloured boxes represent various minisatellite repeat units, shown in the key and described in Appendix 6.4. Total number of repeat units for each allele are indicated alongside in black. Each repeat encodes one ZnF, predicted binding sequences for which are indicated above respective repeat arrays. Bases in red indicate best matches to the 13-mer hot spot motif CCNCCNTNNCCNC to which the PRDM9 A-variant is presumed to bind. Maximum numbers of such matches for all alleles are indicated in grey. Repeat units for all alleles are aligned with respect to the most common allele A.

Table 6.2 *PRDM9* allelic status of men assayed at the *SPRY3* hot spot

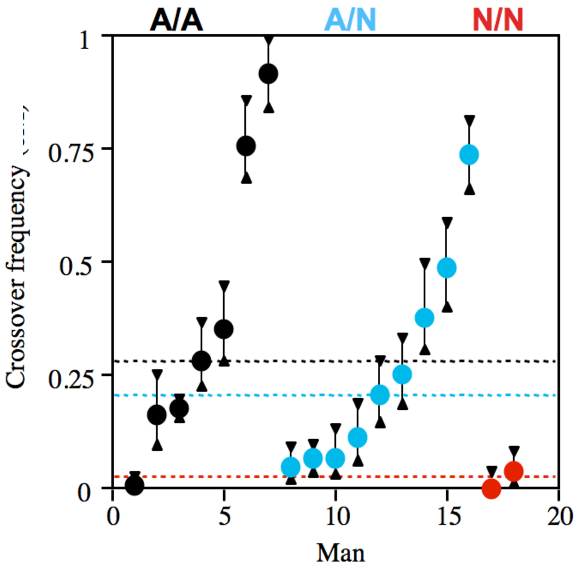
CO frequencies of 11 A/A, 14 A/N and 3 N/N men are shown. A/A men are shown in black, A/N men in blue and N/N men in red. Population origins of each man are also shown; 'Eu' indicates European and 'Af' indicates African. Motif matches represent the maximum number of bases of the 13- bp motif that can be recognised by the ZnF array of PRDM9 encoded by various alleles. Status at 0442 is also indicated as this too influences CO frequencies at the hot spot. Donors within each group, A/A, A/N and N/N, are arranged in ascending order of CO frequency.

Donor name	Population origin	*PRDM9* status	Motif matches	rs700442 status	CO frequency, %
d252	Af	A/A	8/8	C/T	0.008
d5	Eu	A/A	8/8	C/C	0.032
d250	Af	A/A	8/8	C/C	0.087
d39	Eu	A/A	8/8	C/T	0.162
d27	Eu	A/A	8/8	C/T	0.176
d274	Af	A/A	8/8	C/T	0.283
d25	Eu	A/A	8/8	C/T	0.349
d278	Af	A/A	8/8	T/T	0.567
d3	Eu	A/A	8/8	C/T	0.756
d1	Eu	A/A	8/8	T/T	0.845
d66	Eu	A/A	8/8	C/T	0.914
d72	Eu	A/L9	8/8	C/T	0.047
d60	Eu	C/A	5/8	C/C	0.054
d282	Af	L6/A	5/8	T/T	0.064
d87	Eu	A/E	8/4	C/T	0.067
d23	Eu	A/L1	8/6	C/T	0.068
d178	Af	C/A	5/8	C/T	0.109
d246	Af	C/A	5/8	C/T	0.207
d251	Af	A/L16	8/5	C/T	0.251
d184	Af	A/L7	8/7	T/T	0.271
d244	Af	L22/A	8/6	C/T	0.378
d277	Af	A/L5	8/4	C/T	0.488
d232	Af	A/E	8/4	C/C	0.495
d280	Af	L4/A	5/8	T/T	0.567
d91	Eu	A/L20	8/7	C/T	0.734
d254	Af	C/L16	5/5	C/T	0.000
d279	Af	L6/C	5/7	C/C	0.020
d240	Af	L15/L11	5/5	C/T	0.036

Comparison of CO frequencies amongst 0442 C/T men, carrying two, one and no A alleles showed remarkable effects of *PRDM9* status on CO frequencies, with median frequencies corresponding to 0.283%, 0.207% and 0.018% for A/A, A/N and N/N men respectively (see Figure 6.7a). It thus seemed that at least some of the variation in CO frequencies observed amongst 0442 C/T men was due to variation in their *PRDM9* status. Significant differences were also observed amongst *PRDM9* A/A, A/N and N/N men, when this analysis was extended to all men assayed at *SPRY3*, regardless of their 0442 status (see Figure 6.7b). Thus, median CO frequencies amongst N/N men was >14-fold lower than A/A men (*P*<0.01) and >11-fold lower than A/N men (*P*<0.05).

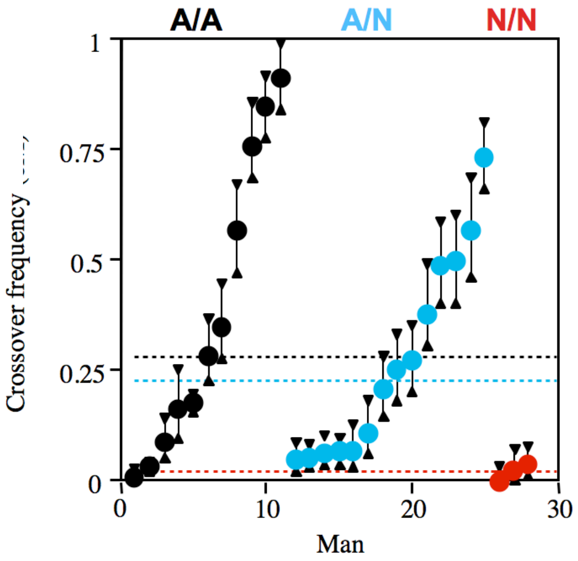
Figure 6.7 Comparison of CO frequencies between *PRDM9* A/A, A/N and N/N men assayed at the *SPRY3* hot spot

a. 0442 C/T men only



a. Classifying 0442 heterozygous men according to their *PRDM9* status. CO frequencies (%) for *PRDM9* A/A, A/N and N/N men are represented using black, blue and red circles respectively. Vertical bars represent 95% confidence intervals on CO frequencies. Median frequencies for each group are indicated using dashed lines in colours representing that group.

b. All men assayed at the hot spot



b. Classifying all men assayed at the *SPRY3* hot spot according to their *PRDM9* status shows profound effects of the *trans*-acting factor on CO frequencies. Again, *PRDM9* A/A, A/N and N/N men are represented using black, blue and red circles respectively. Median frequencies are indicated using dashed lines and 95% error bars are indicated using vertical lines.

It thus seemed that activity at the *SPRY3* hot spot was mainly triggered by the *PRDM9* A allele, very possibly via binding of the protein to the predicted A motif. Also, data from the three N/N men indicated that alleles C, L6, L11, L15 and L16 were incapable of activating the *SPRY3* hot spot, although this was probably true for other N alleles too, given the reduced median frequencies observed amongst A/N men.

Furthermore, it appeared that very subtle changes in the PRDM9 ZnF array created protein variants that were incapable of activating the *SPRY3* hot spot. For example, A/N d72 carried *PRDM9* alleles A and L9. The ZnF array encoded by the latter differs from the A variant by a single Lys→Glu substitution in the 9th ZnF, which is not predicted to alter the DNA-binding specificity of the protein. Nonetheless, CO frequencies for d72 were reduced six-fold relative to the median frequencies for A/A men with same 0442 status as d72. Interestingly, similar examples were also observed at other autosomal hot spots, as discussed in section 6.3 of this Chapter.

Sequence variations in *PRDM9*, however, did not always create hot spot inactivating alleles. Rather, some variations seemed to enhance hot spot activity. For example, A/N d91 carried PRDM9 alleles A and L20, which encoded a protein that was at best, predicted to recognise 7/8 bases of the A motif. Nonetheless, CO frequency for this man was three times higher than the median frequency for A/A men with the same 0442 status. Intriguingly, L20 had similar effects on hot spot MSTM1b, but not on two other autosomal hot spots. This seemingly hot spot-specific effect of allele L20 is discussed further in section 6.3 of this Chapter.

6.2.6. Combined effects of *PRDM9* and 0442 allelic status on CO frequencies

CO data from the *SPRY3* hot spot showed that *PRDM9* N/N men were always highly suppressed at the hot spot, regardless of their 0442 status. This observation was consistent with PRDM9 being the predominant factor influencing CO activity at this hot spot and suggested that it probably acted prior to influences exerted in-*cis* by 0442.

Generally, 0442 and *PRDM9* appeared to have consistent effects on CO frequencies. For instance, 0442 C/C homozygotes showed some of the lowest CO frequencies observed amongst men carrying two *PRDM9* A alleles (P=0.05). Conversely, 0442 T/T homozygotes showed some of the higher CO frequencies observed amongst *PRDM9* A/A men. However, this was not always the case and some inconsistencies were observed, as discussed in the following paragraphs.

Generally, amongst men carrying identical *PRDM9* alleles, 0442 T/T and/or 0442 C/T men showed higher CO frequencies, compared to C/C men. However, the reverse was true in two different cases and C/T men appeared to be recombinationally suppressed relative to C/C men. Thus d87, who carried *PRDM9* alleles A and E and was C/T heterozygous at 0442, showed significantly lower CO frequencies compared to d232, who carried the same PRDM9 alleles, but was C/C homozygous. Similarly, 0442 heterozygous semen donor d252 showed significantly lower CO frequencies, than both C/C men carrying the same

PRDM9 alleles. It is, in theory possible that COs in d232 and/or d252 are suppressed on account of additional *cis*- or *trans*-acting factors.

Furthermore, there remained a substantial amount of variation in CO frequencies that could not be accounted for by either 0442 or *PRDM9* status, an observation consistent with the presence of additional regulators of recombination at the *SPRY3* hot spot. Thus, an ~100-fold variation in CO activity was observed amongst seven *PRDM9* A/A, 0442 C/T men, as shown in Table 6.3. This issue has been addressed to some extent since this work, as described in section 6.3 of this Chapter.

Table 6.3 Variation in CO frequencies amongst PRDM9 A/A, 0442 C/T men

Donor name	CO frequency (%)
d66	0.914
d3	0.756
d25	0.349
d274	0.283
d27	0.176
d39	0.162
d252	0.008

CO frequencies for seven *PRDM9* A/A, 0442 C/T men are listed above in descending order of CO frequencies.

6.3. Discussion

This Chapter represents the first comprehensive effort at investigating factors influencing CO activity at a pseudoautosomal hot spot. Indeed, work described here showed that at least one *cis*-acting factor, namely SNP 0442 and one *trans*-acting factor, namely PRDM9, greatly affected CO frequencies at the hot spot. Nonetheless, ~70% of this variation could not be explained by either (Sarbajna *et al.*, submitted), suggesting the presence of other influences on CO activities at the hot spot. These findings and their implications are discussed further in the following paragraphs.

6.3.1. Central hot spot SNP 0442 influences CO frequencies in *cis* at the *SPRY3* hot spot

Sperm assays in 28 men at the *SPRY3* hot spot showed that CO frequencies at the hot spot were regulated in *cis* by SNP 0442, which lay ~ 45 bp away from the hot spot centre. Thus, men carrying two copies of the active T allele of the SNP showed significantly higher CO frequencies than men carrying two copies of the hot spot suppressing C allele (P=0.03). Further, men heterozygous for the SNP showed TD accompanying COs, with over-transmission of the hot spot suppressing C allele. This was consistent with a disparity in recombination initiation frequencies between haplotypes carrying C and T alleles. The resulting ~2:1 bias in GC accomanying COs, in favour of the less active C allele, resulted in the presence of meiotic drive at the *SPRY3* hot spot. Similar meiotic drives in favour of recombination-suppressing alleles have been documented at several autosomal hot spots, including *NID1*, wherein this drive was expected to significantly increase the likelihood of fixation of the hot spot-attenuating allele (Jeffreys & Neumann, 2005).

Greater extents of TD were observed at hot spot *NID1*, relative to *SPRY3*, with transmission ratios corresponding to 74:26 and 66:34 respectively, in favour of recombination-suppressing alleles (Jeffreys & Neumann, 2005). Nonetheless, on account of the higher average CO frequencies at *SPRY3*, the strength of meiotic drive at this hot spot was expected to be greater at the population level. Indeed, the estimated degree of distortion in population-level transmission ratios of 0442 C:T alleles (50.0496:49.9504), was ~5.5 times greater than that observed at *NID1* (50.009:49.991) (Jeffreys & Neumann, 2005). This in turn implied faster time-scales of fixation of the hot spot suppressing C allele at *SPRY3*, in comparison to the ~12,000 generations predicted in the case of *NID1* (Jeffreys & Neumann, 2005).

Interestingly, the C allele at 0442 disrupted a nearly perfect version of the 13-mer hot spot motif. It has been suggested that this 13-mer motif is the likely binding site for the most common PRDM9 variant A (Myers *et al*., 2010; see section 6.1.3). Indeed, work described in this Chapter showed that in addition to 0442, CO frequencies at the *SPRY3* hot spot were also influenced by variations in the ZnF array of PRDM9, as discussed below.

6.3.2. Hot spot activity at *SPRY3* is also regulated by in *trans* by meiosis-specific protein PRDM9

The >100-fold variation in CO frequency observed amongst 0442C/T men at the *SPRY3* hot spot suggested that this *cis*-acting SNP was not the only factor influencing CO frequencies at this hot spot. Indeed, classifying men assayed at the *SPRY3* hot spot according to their *PRDM9* status revealed profound effects of the latter on CO activity. Thus, *PRDM9* A/A men showed significantly higher CO frequencies compared to A/N and N/N men (P<0.05), consistent with hot spot activation by the PRDM9 A-variant, possibly via binding to the 13-mer hot spot motif.

PRDM9 N/N men carrying alleles C, L6, L11, L15 and L16, which encoded protein variants that could at best recognise 5/6 bases of the 13-mer motif, were recombinationally suppressed at the *SPRY3* hot spot (P<0.03). Further, significantly lower CO frequencies amongst *PRDM9* A/N men compared to A/A men (P<0.05), suggested that most, but not all, N alleles were probably incapable of activating the *SPRY3* hot spot. Indeed, a *PRDM9* A/L20 man showed CO frequencies three-times higher than the median for A/A men with the same 0442 status.

Curiously however, a man carrying *PRDM9* A and L9 alleles, both of which encode protein variants that recognise all 8 bases of the 13-mer hot spot motif, showed a six-fold suppression in CO activity relative to A/A men of the same 0442 status. Analysis of haplotypes of the men assayed revealed no significant additional *cis*-influences that could explain higher and lower than average CO frequencies in the *PRDM9* A/L20 and A/L9 individuals respectively (C.A. May, *pers. comm.*). However, once effects of 0442 and PRDM9 had been taken into account, the study lacks statistical power to detect more subtle effects. Moreover, it is also important to test CO frequencies in men carrying alleles L20 and L9 in combination with known non-activating N alleles, in order to further validate hot spot activating and non-activating effects of the two alleles respectively.

Nonetheless, these unexpected observations raised questions regarding the role of the 13-mer motif (if any) and the nature of its interaction with PRDM9, especially in light of similar observations recorded at autosomal hot spots that were also analysed as part of a much larger survey by Berg *et al*. (2010). Key findings from this larger survey and correlations with *SPRY3* data, are discussed in the following paragraphs.

6.3.3. Insights from a detailed survey investigating the impact of variation in *PRDM9* on recombination activity at CO hot spots

In order to study in detail the influence of variation in the ZnF array of PRDM9, CO activities in a panel of 10 hot spots, including *SPRY3*, were investigated in men carrying a variety of *PRDM9* alleles (Berg *et al*., 2010). Data from these hot spots showed profound effects of *PRDM9* variation on CO frequencies and indicated that hot spot activity was primarily triggered by the PRDM9 A variant at all ten hot spots. Intriguingly, five of the hot spots studied lacked obvious matches to the predicted A-motif, casting further doubts on the proposed mechanism of interaction of PRDM9 with CO hot spots and the role of predicted binding motifs (Baudat *et al*., 2010). These questions can however, most likely only be answered through studies investigating interactions of the protein with hot spot targets *in vivo*.

Consistent with hot spot activation by the A allele, CO activities were significantly reduced amongst N/N men at all ten hot spots (P=5x10^{-16}). Furthermore, median CO frequencies amongst A/N men was nearly half that recorded for A/A men, suggesting that recombination activity at a hot spot was proportional to the number of activating *PRDM9* A alleles.

Data from Berg *et al*. (2010) also highlighted several instances wherein subtle changes in sequences encoding the PRDM9 ZnF array created hot spot inactivating alleles, without altering the predicted DNA-binding sequence of the corresponding protein. In addition to the example described for allele L9 at the *SPRY3* hot spot (see section 6.2.5), this was also seen for allele L13. This allele was non-activating for at least 3 hot spots, yet it encoded a protein that differed from the A variant by a single Ser → Arg substitution in the 11th ZnF. This change affected position -2, with respect to the ZnF α- helix and was therefore not expected to alter the DNA-binding specificity of the protein, relative to the A variant.

Conversely, subtle changes in the PRDM9 ZnF array also seemed to create hot spot *activating* alleles. This was best exemplified by hot spot MSTM1a, originally studied and characterised by Neumann & Jeffreys (2006). The earlier study had shown that this hot spot was active in only 3 out of 26 men studied and presented an example of on/off polymorphism that was independent of any local sequence variation (see Chapter 3, section 3.3). Analysis of *PRDM9* status of the three men active at the hot spot showed that they all carried either allele L9 or allele L24. These alleles differed from each other by a single synonymous base substitution and thus encoded identical ZnF arrays. This association between presence of alleles L9 and L24 and activity at the MSTM1a hot spot was significant (P=0.0004) and implied that these alleles specifically triggered recombination activity at this hot spot. Again, the DNA-binding specificity of proteins encoded by alleles L9 and L24 are not predicted to differ from that encoded by allele A, yet the latter failed to trigger CO activity at the hot spot.

Interestingly, data from Berg *et al.* (2010) also showed that alleles L9/ L24 were not associated with particularly high CO frequencies at neighbouring hot spot MSTM1b. MSTM1b lies ~ 2 kb-away from MSTM1a and was also characterised by Neumann & Jeffreys (2006). Whilst MSTM1b was active in all men studied, it showed a 75-fold variation in activity independent of local sequence polymorphisms. Berg *et al*. (2010) showed that the four men showing the highest CO frequencies at MSTM1b all carried *PRDM9* allele L20. This association was once again significant (*P*=0.0002) and suggested that allele L20 enhanced CO activity specifically at MSTM1b. Activation of closely located MSTM1a and MSTM1b hot spots by different PRDM9 variants was consistent with interactions between PRDM9 and CO hot spots being highly-localised and argued against hot spot activation via extensive chromatin remodelling by PRDM9 (Berg *et al*., 2010).

Interestingly, the *PRDM9* allele L20 also appeared to have the same effect at *SPRY3*, as indicated by elevated CO frequencies in d91 relative to A/A men assayed at the hot spot (see section 6.2.5). Curiously however, allele L20 did not active autosomal hot spots F and U, located on chr12 and chr9 respectively. Different effects of allele L20 on hot spots F/U and MSTM1b and also potentially on *SPRY3*, was again consistent with localised effects of PRDM9 and further suggested that the protein might interact differentially with individual hot spots. Activation of distinct sets of hot spot by different *PRDM9* alleles is explored further in Chapter 8.

Finally, the Berg *et al.* (2010) study also showed that variation in the ZnF array of PRDM9 also influenced minisatellite instability and also the frequencies of *de novo* rearrangements associated with genomic disorders. These findings implied that PRDM9 not only regulated activity at meiotic CO hot spots, but also influenced other forms of recombination-associated processes in the human genome and are discussed further in Chapter 9 of this work.

6.3.4. Evidence of additional influences on CO activity at the *SPRY3* hot spot

Despite their profound effects, *PRDM9* and 0442 alone did not account for all the variation observed at the *SPRY3* hot spot. Indeed, men carrying identical *PRDM9* and 0442 alleles showed ~100-fold variation in CO frequencies. Furthermore, CO frequencies appeared to be significantly reduced in some 0442 C/T men, relative to C/C men of the same *PRDM9* status (see section 6.2.6), again raising the possibility that COs in the former men were suppressed by additional *cis*- or *trans*-acting factors. In order to identify these, a number of potential *cis*- and *trans*-acting influences were investigated in detail since this work.

In order to identify any additional *cis*-acting regulators of recombination at *SPRY3*, haplotypes of men assayed at this hot spot, as determined by 31 SNPs, were analysed over an ~10 kb interval. In addition, variation within a purine-rich tandem repeat region, centred ~300 bp away from the hot spot centre and potential epigenetic effects mediated through differential DNA methylation at the 0442C allele, which creates a CpG doublet, were also investigated (R. Neumann; C.A. May).

Furthermore, three genetic loci *RNF212, UGCG* and *NUB1*, which were associated with genome-wide levels of recombination in human males (Kong *et al*., 2008; Chowdhury *et al*., 2009), were investigated as potential *trans*-acting regulators of CO activity (C.A. May). Findings from these studies are discussed in Chapter 9 of this work.

Chapter 7. A survey of non-crossovers at the *SPRY3* hot spot

7.1. Introduction

Cytogenetic studies in both mice and humans show a large excess of RAD51/DMC1 foci, which represent sites of DSB-formation, over MLH1 foci and/or chiasma, which represent sites of crossing-over (Vallente, Cheng & Hassold, 2006; Baudat & de Massy, 2007). This large excess of DSBs over COs predicts that a large number of DSBs, approximating 90% and 30% in mice and humans respectively, must be resolved through inter-homolog recombination without exchange, that is, through NCO (Baudat & de Massy, 2007). In mice, studies based primarily on the *Psmb9* hot spot have shown co-localisation of COs with NCOs, consistent with them being products of the same recombination initiating event, that is, SPO11-induced DSBs (Guillon & de Massy, 2002). Consistent with this, studies in *Spo11*$^{-/-}$ female mice have shown that both products are dependent on the protein (Guillon *et al.*, 2005).

In humans too, the two outcomes of meiotic recombination have been shown to co-localise at various hot spots, including *DNA3*, *DMB2*, *SHOX* and *NID1* (Jeffreys & May, 2004; Jeffreys & Neumann, 2005). Co-localisation of COs and NCOs was also observed at hot spot PAR1-P, described in Chapter 4 (Jeffreys & May, 2004; Jeffreys & Neumann, 2005). However, proportions of NCOs at these hot spots, with the exception of hot spot *DNA3*, have usually been lower than predicted from cytogenetic data (Jeffreys & May, 2004; Baudat & de Massy, 2007). The most extreme example of this is the human β-globin hot spot, where NCO:CO ratios are <1:12 (Holloway, Lawson & Jeffreys, 2006). This paucity of NCOs in comparison to COs raises the possibility that many DSBs are instead being repaired using inter-sister, rather than inter-homolog recombination events (Cole, Keeney & Jasin, 2010). However, it seems equally possible that true numbers of NCOs, whose detection is limited by the availability of polymorphic markers, were underestimated in these limited surveys.

In order to address this further and to evaluate in greater detail the true proportions of COs:NCOs at a mouse hot spot, Cole and colleagues (2010) investigated both types of events at the highly-active *A3* hot spot using combinations of various inbred mouse strains. Data from highly polymorphic F1 hybrids showed that true frequencies of NCOs at this hot spot were much closer to the expected 9:1 ratio and also provided a more detailed understanding of distributions of NCOs in this hot spot.

7.1.1. Frequencies and distributions of NCOs at the mouse *A3* hot spot

Cole and colleagues (2010) used high-resolution sperm assays to investigate COs and NCOs at the *A3* hot spot on chr1, which was previously characterised by Kelmenson *et al.* (2005). In order to maximise the chance of recovering NCOs, they used highly polymorphic F1-hybrids obtained from crossing inbred mouse strains, A/J (A), C57BL/6J (B), C3H/HeJ (C) and DBA/2J (D). Sperm assays in AxD F1-males showed that COs at the *A3* hot spot occurred in 0.22% of male meioses. On the other hand, NCOs affecting one or more of the 19 polymorphisms screened, located ~100 bp apart, occurred in 2.3% of male meioses. This implied that NCOs at the *A3* hot spot occurred >10-times more frequently than COs, at least in AxD hybrid mice. Notably, despite a high density of polymorphisms, 84% of NCOs at the *A3* hot spot involved only one marker, implying that conversion-tracts associated with NCOs were short. In fact, average minimum and maximum NCO tract lengths were estimated at 7 and 187 bp respectively.

Furthermore NCOs in AxD mice were distributed throughout a 1.5 kb-wide interval, within which 90% of COs occurred. Indeed, 42% of NCOs on the D chromosome occurred in regions flanking the two central markers showing the highest NCO activity. This implied that NCOs were distributed more widely than previously indicated by studies in a limited number of mouse and human hot spots. Given that NCO-tracts were short, they were more likely to approximate locations of recombination-initiating DSBs. Thus, data from the *A3* hot spot suggested that DSBs occurred within a certain preferred interval.

In addition to AxD hybrids, Cole *et al.* (2010) also investigated recombination in BxD and CxD hybrids. AxD hybrids differed from the other hybrids in that they showed strong TD accompanying COs and biased NCOs in favour of the A chromosome. BxD and CxD hybrids showed a CO-refractory zone at the hot spot centre. Despite these strain-specific differences in CO distributions, AxD and BxD hybrids showed many similarities in NCO profiles. Indeed, the frequency of NCOs greatly exceeded that of COs in both hybrids and majority of these NCOs (~86%) involved a single polymorphism. Estimates of maximum and minimum NCO tracts lengths were also remarkably similar. Furthermore, NCOs in BxD hybrids occurred at the hot spot centre, despite presence of a CO refractory zone and thus, showed significant overlap in distribution with those from AxD mice. Finally, 92% of NCOs involving co-conversion of markers, were located within a central region of ~100 bp in both hybrids. This ‘co-conversion’ zone also showed peak levels of overall NCO activity, consistent with a greater occurrence of co-conversions in regions receiving frequent DSBs.

In summary, the presence of a high density of polymorphisms, approximating 1 every 30 bp over the hot spot centre, allowed Cole and colleagues (2010) to investigate both COs and NCOs at the *A3* hot spot in great detail. In addition to revealing important features of distributions of NCOs, this study was the first to show a large excess of NCOs over COs, as predicted from cytogenetic data in mice (see above). Whilst it was, in theory, possible that the *A3* hot spot was unusual in this respect, it seemed far more likely that the high density of polymorphisms in the F1-hybrids studied allowed unprecedented scope for detecting true levels of NCOs. Indeed, Cole and colleagues (2010) predicted that the *A3* hot spot was representative of recombination in the rest of the mouse genome and that high-resolution analysis at other similarly polymorphic hot spots would also reveal high NCO:CO ratios. Finally, this work, along with that of Guillon *et al.* (2005), provided some insight into the likely pathways of NCO-formation, as described below.

7.1.2. Pathways of CO and NCO-formation in mouse meiosis

In yeast *S. cerevisiae*, COs and NCOs are thought to be formed using distinct pathways, as described in Chapter 1, section 1.2.4. Thus, COs are thought to be generated via resolution of a dHJ intermediate, whilst NCOs are thought to result from distinct intermediates of a different pathway, most likely SDSA (see Figure 1.7, Chapter 1). Work primarily based on *Psmb9* and *A3* hot spots suggest that the same may be true for mice (Guillon *et al.*, 2005; Cole, Keeney & Jasin, 2010). Some of this work was described previously in Chapter 1 (section 1.4.3) and is briefly recapitulated here.

First, data from AxD F1-hybrids at the *A3* hot spot, showed that COs recovered from either orientation were equivalent in rate, despite biased initiation of recombination and TD (Cole, Keeney & Jasin, 2010). This was consistent with reciprocal exchange, as predicted by the DSBR model of CO-formation (Szostak *et al.*, 1983). On the other hand, hardly any NCOs were recovered from the recombinationally suppressed haplotype, consistent with NCOs not altering the donor chromosome. The latter was a key tenet of the SDSA pathway for NCO-formation (Nassif *et al.*, 1994; Pâques & Haber, 1999).

Second, Cole and colleagues (2010) reported that gene conversion (GC)-tracts associated with COs at the *A3* hot spot were ~500 bp in length and were much longer than those associated with NCOs, which on average varied between 16-117 bp. Similar differences in GC-tracts associated with COs and NCOs were also reported by Guillon *et al.* (2005) from their work on the mouse *Psmb9* hot spot. These differences are consistent with GCs with and without exchange resulting from different intermediates.

Third, data from the *Psmb9* hot spot showed a dramatic reduction in levels of COs, to 5-10% of wild-type levels, in *Mlh1*$^{-/-}$ male and female mice (Guillon *et al.*, 2005). Fine mapping of COs in these mice also showed that Mlh1 was not only required for formation of the majority of COs, but also for mismatch repair of the hDNA formed in this process (Guillon *et al.*, 2005). However, Guillon *et al.* (2005) showed that NCOs in these mice remained unaffected and appeared with similar kinetics and at similar frequencies as in wild-type mice. Further, tract-lengths of NCOs also remained unaltered in *Mlh1*$^{-/-}$ male and female mice. Collectively, these data suggested that unlike COs, NCOs did not require MLH1 and further supported separate pathways for the two outcomes of meiotic recombination. It is worth mentioning here that work at the *Psmb9* hot spot was also instrumental in highlighting *trans*-regulation of not just CO, but also NCO-activity at the hot spot, as described below.

7.1.3. Regulation of NCOs at a mouse recombination hot spot

Data from Parvanov *et al.* (2009) and Grey *et al.* (2009) suggested that two overlapping loci on chr17, namely *Dsbc1* and *Rcr1*, regulated CO activity at several hot spots, including those located on other chromosomes (see Chapter 1, sections 1.4.2.5 and 1.4.2.6). Subsequent work showed that these loci reflected the effects of the meiosis-specific gene *Prdm9* (Parvanov *et al.*, 2010). Since then, data from several studies, including that described in Chapter 6, have firmly established PRDM9 as a global *trans*-regulator and specifier of human and mice CO hot spots (Baudat *et al.*, 2010, Myers *et al.*, 2010, Berg *et al.*, 2010).

Interestingly, analysis of NCO frequencies at the *Psmb9* hot spot on chr17, suggested that locus *Dsbc1* not only regulated COs, but also NCOs at the hot spot (Grey *et al.*, 2009). The same was true for *Rcr1*, which was shown to similarly influence CO, as well as NCO activity at hot spots *Hlx1* and *Esrrg-1* on chr1 (Parvanov *et al.*, 2009). Collectively, these studies implied that *Dsbc1*/ *Rcr1*, and thus *Prdm9*, most likely influenced recombination-initiation frequencies at the aforementioned mouse hot spots, rather than the CO/NCO decision and thus, had similar impacts on levels of both types of events.

7.1.4. This work

In humans, the frequencies and distributions of COs have been studied extensively. Indeed, COs at nearly 50 sperm hot spots have been directly investigated at micro-Morgan resolutions. Furthermore, >10 men were assayed per hot spot for at least 17 of these hot spots, thus providing detailed insights into variations in CO distributions and activities and *cis*- and/or *trans*-acting factors influencing these (as shown for the *SPRY3* hot spot in Chapter 6). In contrast, NCOs have only been investigated at six hot spots, including hot spot PAR1-P described in Chapter 4 (Jeffreys & May, 2004; Jeffreys & Neumann, 2005; Holloway, Lawson & Jeffreys, 2006). These studies have typically been based on the analysis of a limited number of NCOs at a limited number of polymorphic sites (usually <10). Moreover, only one or two men were assayed for NCOs per hot spot, leaving the extent of variation in NCO activity, if any, and factors influencing this largely unexplored.

In order to investigate in detail the distribution and frequencies of NCOs at a human recombination hot spot, these events were analysed using reciprocal sperm assays in 14 men at the PAR2 *SPRY3* hot spot. *SPRY3* contained a good density of polymorphisms across the hot spot, making it a suitable candidate hot spot for studying NCOs, which leave very localised signatures.

Men heterozygous at the central SNP 0442 showed TD and biased GC accompanying COs, as described in Chapter 6, section 6.2.3. This feature of the *SPRY3* hot spot provided scope for investigating whether the GC-bias seen in COs was reflected in NCOs, as would be expected if both events resulted from alternative resolution of a common intermediate. Limited analysis of NCOs in one man at the *NID1* hot spot showed that this was indeed the case with both GC-accompanying COs and NCOs showing virtually indistinguishable biases (Jeffreys & Neumann, 2005). Furthermore, TD at the *SPRY3* hot spot also provided the opportunity for making comparisons between tract lengths associated with GC with and without exchange. These comparisons can shed valuable light on likely pathways of CO and NCO formation in the human genome and enable important comparisons with other model organisms, in particular the mouse (Cole, Keeney & Jasin, 2010).

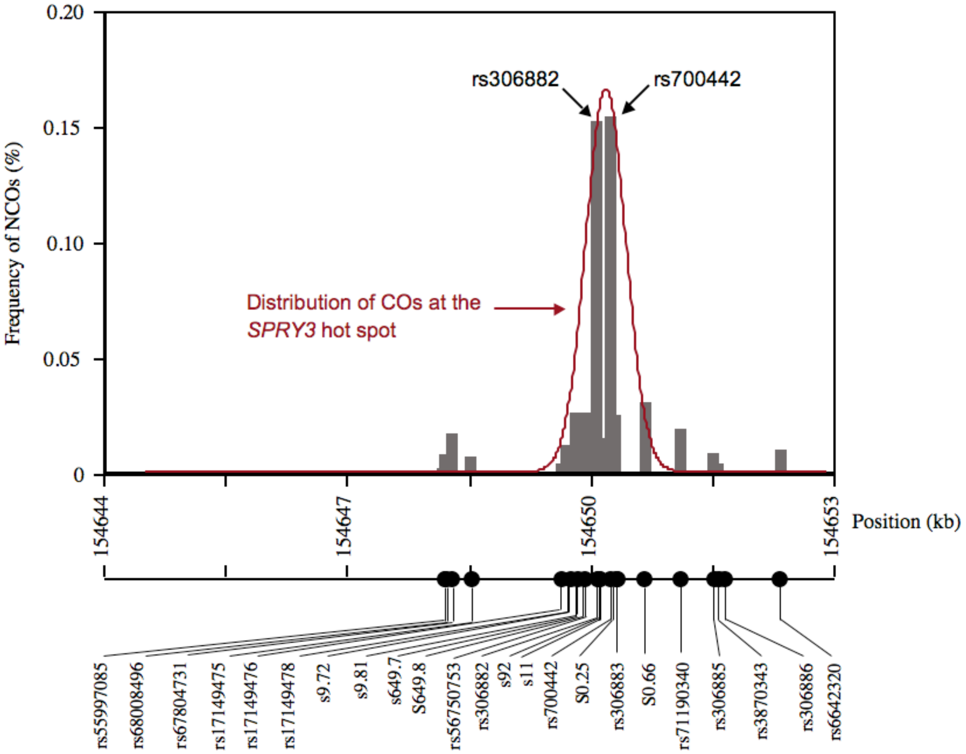
Work described in Chapter 6 (section 6.2.5) showed that CO activity at the *SPRY3* hot spot is regulated in *trans* by the meiosis-specific protein PRDM9. Data from mouse hot spots suggest that PRDM9 influences the initiation of recombination and affects NCOs in much the same way as COs (Parvanov *et al*., 2009; Grey *et al*., 2009). Testing NCO frequencies in *PRDM9* A/A, A/N and N/N men already assayed for COs at the *SPRY3* hot spot would enable this to be tested directly at a human recombination hot spot and potentially offer insight into factors regulating NCO formation in humans.

Finally, relative frequencies of COs and NCOs have been shown to vary substantially amongst the human hot spots tested to date (Jeffreys & May, 2004; Jeffreys & Neumann, 2005; Holloway, Lawson & Jeffreys, 2006). However, in almost all of these studies, NCOs were only investigated in one man so little is known about inter-individual variation in relative frequencies of COs and NCOs and factors influencing this balance. An extensive survey of NCOs, such as the one described below, allowed this to be addressed in detail.

7.2. Results

NCOs at the *SPRY3* hot spot were analysed in a total of 14 men using ‘assay 2’, which was described in detail in Chapter 6, section 6.2.1. All NCOs recovered using this assay were mapped and scored as described in Chapter 4, section 4.2.2.4. In total, >500 NCOs were analysed in this survey and ~80% of these involved SNPs rs700442 (0442) and/or rs306882 (6882). Both SNPs were close to the hot spot centre, located 45- and 119 bp-away from it respectively. This was consistent with co-localisation of NCOs with COs, as seen at other human and mice hot spots (Jeffreys & May, 2004; Guillon & de Massy, 2005; Cole, Keeney & Jasin, 2010). NCO data for each donor is summarised in Appendix 7.1. Overall distribution of NCOs across the *SPRY3* hot spot, determined using data from reciprocal assays in 14 men, is shown in Figure 7.1.

Figure 7.1 Distribution of NCOs at the *SPRY3* hot spot



Frequencies of NCOs for 23 polymorphisms at the *SPRY3* hot spot. Morphology of the CO hot spot, as described in Chapter 5, section 5.2.4, is indicated using a red line. It is important to note that all 23 polymorphisms were not typed in a single man. Instead data for each marker was combined for all men informative for that marker and the average thus obtained was included in the figure. Coordinates are as per chrX, NCBI36/hg18.

Whilst most NCOs occurred close to the hot spot centre, they were nonetheless spread over an ~4 kb-interval. Indeed, marker rs6642320, located ~2100 bp away from the hot spot centre, showed an average NCO frequency of 0.008%. This NCO-distribution was consistent with data from hot spot PAR1-P (Chapter 4), but wider in comparison to other human hot spots (Jeffreys & May, 2004). The possibility that these widely distributed, low frequency events were PCR mis-incorporations, cannot be discounted however.

The gradient in NCO frequencies at *SPRY3* was very steep and was comparable to those at hot spots *DNA3, DMB2* and *SHOX* (Jeffreys & May, 2004). Thus SNP rs56750753, located only 148 bp-away from SNP 6882, showed a >6-fold decrease in average NCO frequencies.

SNPs s92 and s11 located in between markers 0442 and 6882 showed unusually low average NCO frequencies. These SNPs were identified through re-sequencing across the hot spot centre and were heterozygous in just 3-4 men screened for NCOs (R. Neumann; *pers. comm.*). Thus, s11 was informative in d178, d244, d252 and d254, whilst s92 was informative in d240, d244 and d254. Interestingly, these men showed some of the lowest NCO frequencies, including at the central SNP 0442 (P=0.005 and P=0.01 for s11- and s92-heterozygous men respectively). Lower average NCO frequencies at s11 and s92 therefore most likely reflected suppressed activities in the men screened for these markers, rather than a local NCO-refractory zone within the hot spot.

Men assayed at the *SPRY3* hot spot showed striking variations in NCO frequencies, both at the level of individual haplotypes of any one man and also in overall NCO frequencies between different men. It is important to bear in mind that these comparisons were limited to detectable NCOs, which in turn relied on the presence of suitably informative markers close to the hot spot centre. Therefore, comparisons of NCO frequencies between men assayed were limited to events affecting central SNP 0442, which was informative in thirteen of the fourteen men analysed.

7.2.1. Analysis of NCOs at central SNP 0442

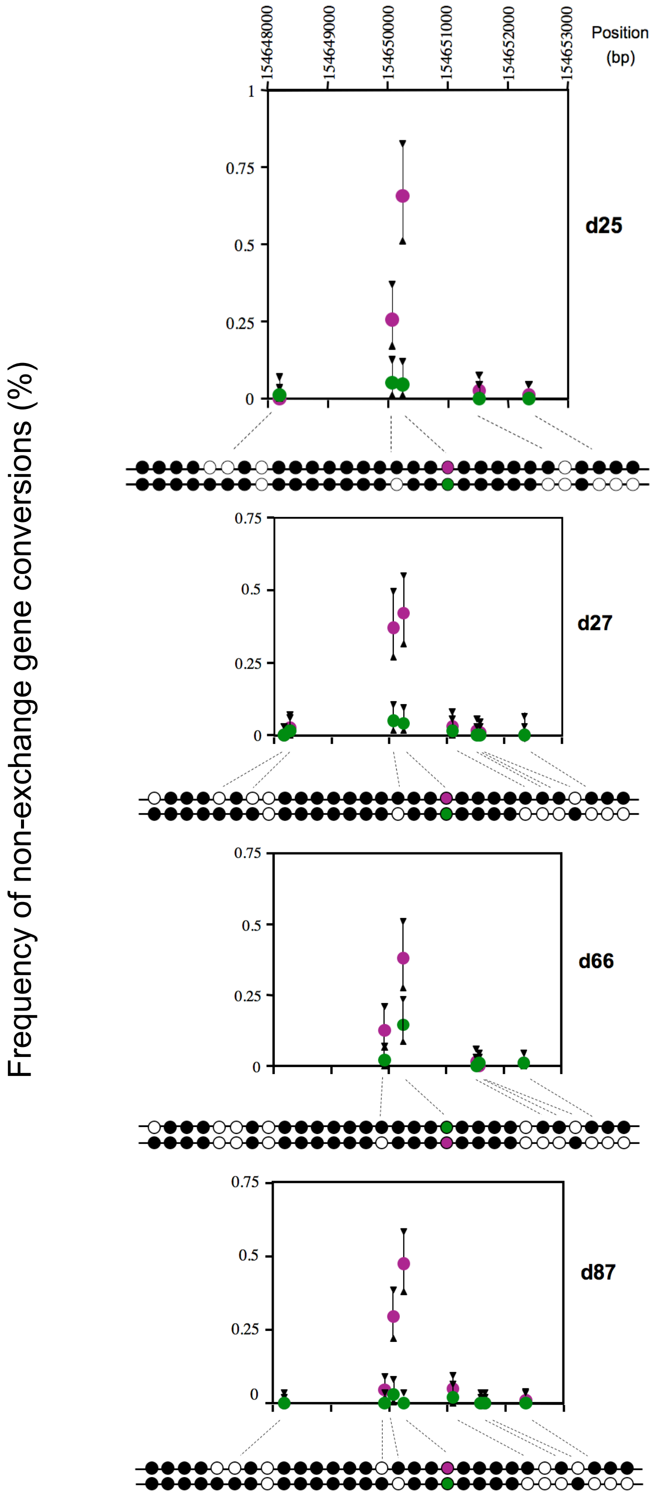
Both haplotypes of thirteen men heterozygous at 0442 were screened for NCOs. Between 5,246 and 17,815 molecules were screened per orientation and between 0-85 events recovered per orientation per man. In total, 368 NCOs were scored at this marker. All 0442 heterozygous men showed biased NCOs at the marker, with over-transmission of the C allele, as shown in Table 7.1 and Figure 7.2. Thus, 0442 T→C NCOs were more frequent than C→T events and this was significant for eight of thirteen men (*P*<0.003). Significant over-transmission of the C allele in NCOs was consistent with CO data at the hot spot and pointed to a higher frequency of recombination initiation on the haplotype carrying the active 0442T allele. It should be noted that the lack of significant bias in favour of 0442 T→C NCOs in the remaining 5 five men was most likely due to low numbers of events (<10) recovered in these men. Interestingly, the aforementioned bias in NCOs at 0442 varied rather substantially amongst men (Table 7.1 and in Figure 7.2). Thus in d87, 85 T→C and 0 C→T NCOs were recovered from screening equivalent numbers of molecules in both orientations. In contrast, the bias was far less extreme in d66, for whom 16 C→T versus 42 T→C NCOs were identified. Ratios of T→C versus C→T NCOs are listed in Table 7.1 for all men for whom both types of events could be recovered.

Table 7.1 NCOs at central SNP 0442

0442 NCO data for donors for whom such events were recovered from both haplotypes. Numbers of NCOs on either haplotype of a donor were normalised for estimates of T → C versus C → T bias. All values have been rounded-up to the nearest integer.

Donor name	C→T NCOs		T→C NCOs		Bias
	Molecules screened	No. of NCOs	Molecules screened	No. of NCOs	
d72	9070	1	9950	18	17
d25	8320	5	10900	71	14
d23	6460	1	6600	12	12
d27	12590	5	12370	53	11
d39	5860	3	5730	22	7
d274	14720	3	7050	19	6
d252	13300	2	13100	6	3
d66	11200	16	11150	42	3
d178	7240	3	5250	6	2
d244	5360	5	12290	5	1

Figure 7.2. Biased NCOs at SNP 0442



NCO frequencies are shown, using circles, for men for whom at least 50 0442-events were scored. Vertical bars represent 95% confidence intervals. NCOs on the haplotype carrying the 0442T and 0442C alleles are in purple and green respectively. In all cases, NCOs occurred more frequently on the haplotype carrying the T allele and most of these involved 0442.

NCO data from 10 men indicated that on average, there was a 7.5:1 bias in NCOs at 0442, in favour of the C allele. It should be noted that this was significantly different to the 1.94:1 bias observed for GCs accompanying COs. This difference is unprecedented and is discussed further in section 7.3.

As with COs, markers flanking 0442 also showed bias in NCOs (see Chapter 6, section 6.2.3 for CO data). Thus, significant bias in NCOs was seen at SNP 6882 in d25, d27 and d87 and also at SNP rs56750753 in d27 and d66. In all cases, alleles on the same haplotype as 0442T showed higher NCO frequencies, consistent with higher frequencies of recombination initiation on the haplotype carrying the T allele (Figure 7.2).

In addition to the aforementioned differences in NCO frequencies between the two haplotypes of the same individual, a ~50-fold variation in overall NCO activity at 0442 was also observed between various men assayed at the hot spot. This is highlighted in Table 7.2. Although allelic status at 0442 was sufficient to explain differences in NCO frequencies between haplotypes, it could not account for overall variation observed *between* different men. It therefore seemed possible that additional factors influenced NCO frequencies at the *SPRY3* hot spot. Data described in Chapter 6 showed that variations in sequences encoding the ZnF domain of PRDM9 had profound effects on CO frequencies at *SPRY3*, as well as at 11 other autosomal hot spots. Further, data from mouse hot spots *Psmb9*, *Hlx1* and *Esrrg-1* were consistent with PRDM9 influencing both CO and NCO frequencies at recombination hot spots (Parvanov *et al.*, 2009; Grey *et al.*, 2009; see section 7.1.3). Hence, any potential effects of variations in *PRDM9* status on NCO frequencies at 0442 were investigated next.

Table 7.2. Variations in overall NCO frequencies at 0442

Table shows average frequencies of C-›T and T-›C NCOs at 0442. Upper and lower 95% confidence intervals are also indicated. Men are listed in ascending order of NCO frequencies.

Donor name	NCO frequency (%)	Lower 95% C.I.	Upper 95% C.I.
d240	0.000	0.000	0.024
d254	0.007	0.000	0.040
d252	0.031	0.013	0.060
d244	0.039	0.016	0.082
d178	0.059	0.023	0.116

Donor name	NCO frequency (%)	Lower 95% C.I.	Upper 95% C.I.
d274	0.076	0.029	0.097
d72	0.098	0.062	0.158
d23	0.102	0.055	0.173
d39	0.211	0.134	0.309
d27	0.230	0.173	0.295
d87	0.237	0.236	0.364
d66	0.261	0.198	0.336
d25	0.350	0.307	0.488

7.2.2. Impact of *PRDM9* variation on frequency of NCOs at 0442

A comparison of NCO frequencies at 0442 amongst *PRDM9* A/A, A/N and N/N men revealed a major effect of *PRDM9* status (see Figure 7.3). Thus, median NCO frequency for A/A men was 0.22%, which was significantly higher than the 0.004% median for N/N men (*P*=0.026). This was consistent with activation of NCOs at 0442 by the *PRDM9* A allele. Moreover, median NCO frequency for A/N men was ~45% that of A/A men, consistent with a simple additive model wherein NCO activity was proportional to the number of *PRDM9* A alleles present.

Figure 7.3 Variation in frequencies of NCOs according to *PRDM9* status

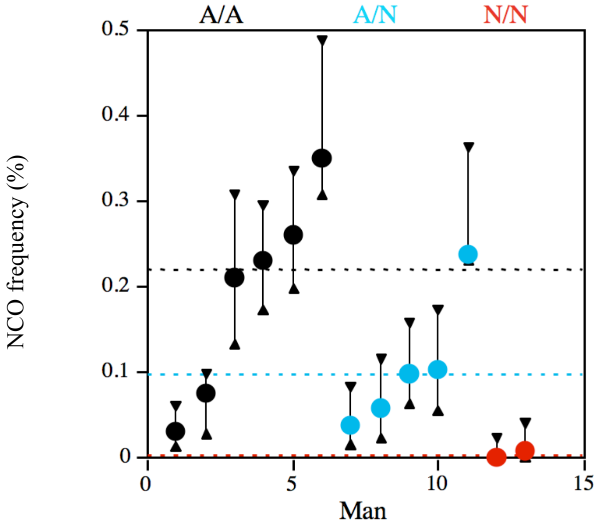
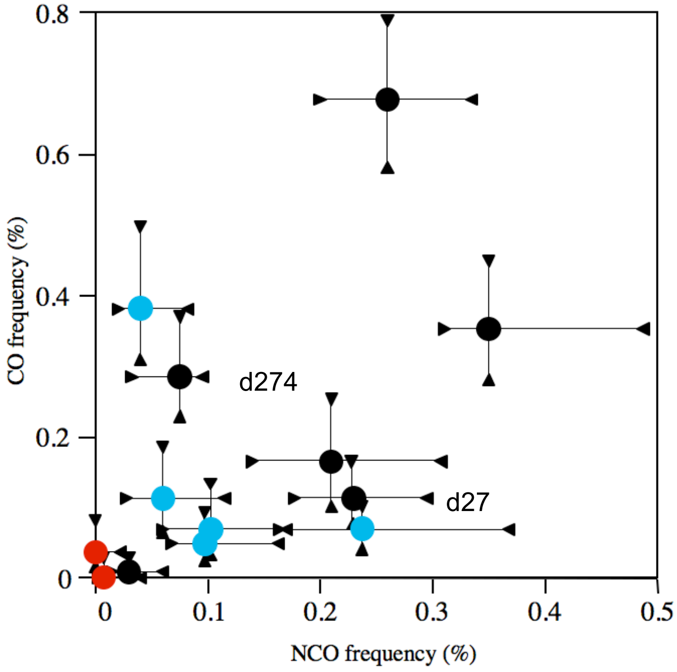


Figure shows variation in average frequencies (%) of NCOs at 0442. NCO frequencies of *PRDM9* A/A, A/N and N/N men are indicated using black, blue and red circles respectively. 95% confidence intervals are indicated using vertical bars. Black, blue and red horizontal dotted lines represent median NCO frequencies of A/A, A/N and N/N men and correspond to 0.220%, 0.098% and 0.004% respectively.

Overall, activation of COs and NCOs by the *PRDM9* A allele and significant suppression of both types of events in *PRDM9* N/N men implied that this *trans*-acting regulator influenced the frequency of initiation of recombination and acted upstream of the CO/NCO decision. This is further illustrated in Figure 7.4.

Figure 7.4 Comparison of CO- and NCO-frequencies at the *SPRY3* hot spot



The figure shows relative frequencies of COs and NCOs for 13 men assayed at the hot spot. NCO frequencies provided represent those NCOs that include marker 0442. *PRDM9* A/A, A/N and N/N men are represented in black, blue and red respectively. Vertical and horizontal bars represent 95% confidence intervals on CO and NCO frequencies respectively.

Analysis of relative frequencies of NCOs and COs, as shown in Figure 7.4, was consistent with *PRDM9* influencing recombination initiation, but not the CO/NCO decision ($P<0.05$). Thus, some *PRDM9* A/A men, such as d27, showed a clear preference for NCOs at the *SPRY3* hot spot, whereas others including d274, showed a preponderance of COs. Similar variations were also seen amongst *PRDM9* A/N men.

Overall, NCO:CO ratios at the *SPRY3* hot spot ranged from 4:1 - 1:10. This variation was within the range of 2.7:1 - 1:12 obtained from analysis of 1 or 2 men at several different human hot spots and suggested that factors influencing the CO/NCO decision at *SPRY3* and probably also at other hot spots were likely to be complex and/or prone to considerable stochastic variations. It is however worth mentioning that mean CO:NCO ratio at the *SPRY3* hot spot was ~2:1 and was thus similar to the 3:1 ratio expected from cytogenetic data (Baudat & de Massy, 2007).

7.2.3. Co-conversions and estimated conversion tract lengths

Data from other human hot spots had previously indicated that NCOs were highly localised, with average tract-lengths ranging between 55 and 290 bp. ~86% of NCOs recovered at the *SPRY3* hot spot affected a single SNP, despite a good marker density near the hot spot centre, consistent with these events having very localised effects. Maximum and minimum tract-lengths associated with these single-site NCOs are listed in Table 7.3.

It is worth mentioning that although estimates for maximum NCO tract lengths were large for some men, extending to nearly 6 kb, this was most likely a reflection of low marker density and non-uniform marker distribution in these men across the assay interval. Overall, there was no evidence to indicate that average NCO tract lengths at the *SPRY3* hot spot were significantly longer than estimated from previous studies at other hot spots (Jeffreys & May, 2004).

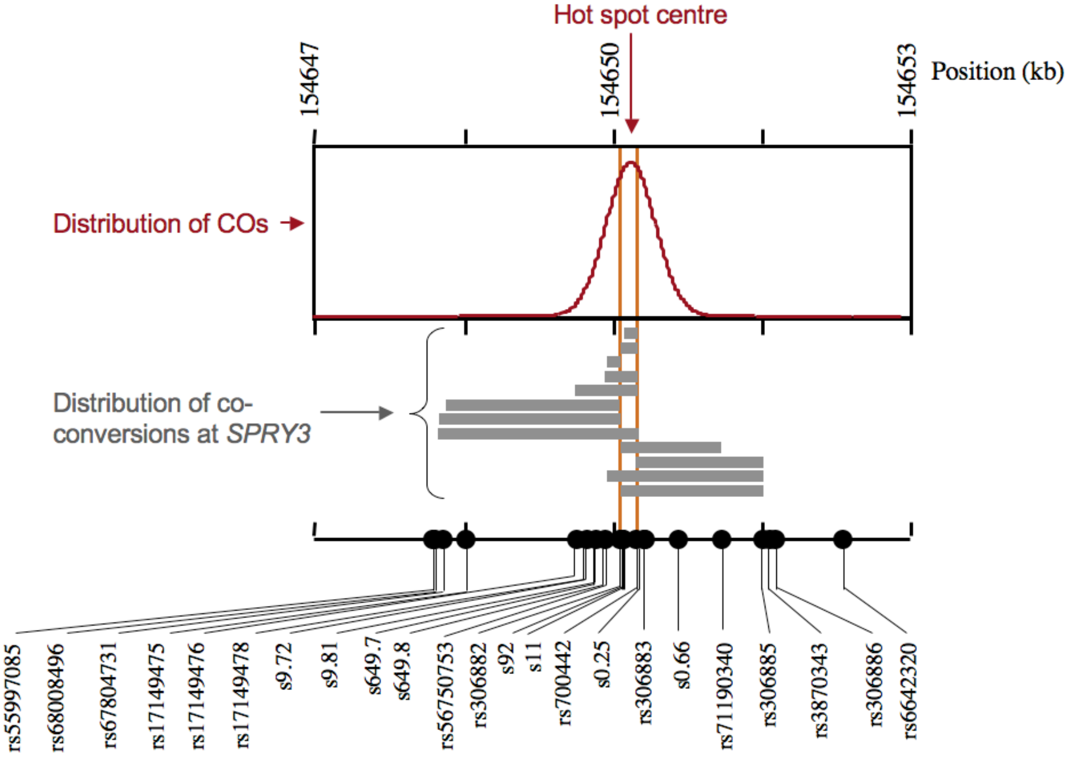
Table 7.3 Minimum and maximum tract lengths associated with single-site NCOs

Estimated tract-lengths for single-site NCOs in 14 men are listed below. Minimum tract-lengths are constant, since all events were only detectable at a single SNP. Estimates of maximum tracts lengths represent the distance between the SNP at which a NCO is detected and nearest 5′ and 3′ NCO-negative SNPs. These estimates are different both for various types of single-site NCOs and also for various men assayed at the hot spot. On account of these factors, a range of maximum tract-lengths is provided for each donor. Numbers of markers types are stated to provide an idea of how informative a particular man is over the assay interval.

Donor name	No. of markers typed	Minimum tract lengths (bp)	Range of estimated maximum tract lengths (bp)	
			Lower limit	Upper limit
d23	7	1	1501	2044
d25	6	1	1143	2095
d27	9	1	1031	1960
d39	5	1	1263	2827
d66	7	1	1337	5739
d72	8	1	1687	2044
d87	9	1	164	2830
d178	11	1	312	1395
d240	9	1	2592	2592
d244	7	1	190	4610
d252	10	1	142	1732
d254	8	1	141	229
d274	10	1	100	2666
d279	6	1	413	413

Whilst the majority of NCOs at *SPRY3* affected a single SNP, ~14% of events encompassed more than one marker. These events are henceforth referred to as ‘co-conversions’. All co-conversions detected at *SPRY3* involved either 0442 or 6882 or in most cases, both (see Figure 7.5). These markers also showed the highest overall NCO frequencies detected at the *SPRY3* hot spot. Since NCOs are generally good proxies for positions of DSBs on account of their highly localised nature, these findings indicated that co-conversions preferentially occur within and/or encompassed regions of frequent DSB-formation. This was consistent with observations at the mouse *A3* hot spot, as discussed in section 7.1.1 (Cole, Keeney & Jasin, 2010).

Figure 7.5 Distribution of co-conversions seen at the *SPRY3* hot spot



The figure shows distribution of NCOs that involve co-conversion of two/more polymorphisms. Minimum tract lengths associated with these events are indicated using grey bars. Orange lines highlight positions of SNPs 0442 and 6882. Distributions of CO at *SPRY3* and position of the hot spot centre are also indicated.

Whilst all co-conversions involved central markers 0442 and/or 6882, they extended up to ~2 kb upstream and ~1.3 kb downstream of the centre of the *SPRY3* hot spot (see Figure 7.4). It is worth noting however, that ~66% of all co-conversions observed, encompassed 0442 and 6882 only, which are located only 164 bp-apart. Estimates of maximum and minimum tract lengths associated with co-conversions are listed in Figure 7.4 for all donors assayed for NCOs.

Table 7.4 Minimum and maximum tract lengths associated with ‘co-conversions’

Estimated tract lengths for NCOs involving co-conversion of two or more markers are shown below. Different types of events scored for each donor are also indicated. Maximum and minimum tract length estimates depend on types of co-conversions, hence a range of values are provided for these estimates. Where lowest and highest estimates for minimum tract lengths are identical, only one type of co-conversions was detected. No co-conversions were detected in d39, d240, d244 and d274. The shortest minimum tract length and the longest maximum tract length are 21bp and 7012 bp respectively. Note: 8496 = rs68008496, 6885 = rs306885, 4731 = rs67804731, 0753 = rs56750753, 0340 = rs71190340, 9476 = rs17149476, 9478 = rs17149478.

Donor name	% of all NCOs	Type(s) of events	Minimum tract lengths (bp)		Maximum tract lengths (bp)	
			Lowest estimate	Highest estimate	Lowest estimate	Highest estimate
d23	33.3	6882 & 0442	165	165	2031	3381
d25	12.9	8496, 6882 & 0442; 8496 & 6882; 6882 & 0442; 0442 & 6885	165	2077	2076	3599
d27	21.2	4731 & 6882; 6882 & 0442; 6882, 0442, 0340 & 6885	165	1913	2076	3297
d39	0	-	-	-	-	-
d66	19.7	0753 & 0442; 0753, 0442 & 6885	313	1586	1649	7012
d72	13.0	6882 & 0442	165	165	3567	3567
d87	13.2	0753 & 6882; 0753, 6882 & 0442; 6882, 0442 & 0340	165	1032	1179	2911
d178	9.1	0442 & s11	123	123	1585	1585
d240	50.0	9476, 9478, s649.7 & s92	284	284	551	551
d244	0	-	-	-	-	-
d252	3.8	0442 & s0.25	21	21	989	989
d254	33.3	9476, s649.8, s92, s11 & 0442	352	352	2911	2911
d274	0	-	-	-	-	-
d279	0	-	-	-	-	-

Frequencies of co-conversions varied substantially amongst men assayed at *SPRY3*. Although differences in marker density and total numbers of NCOs analysed contribute towards this, they cannot account for all the observed variation. This is discussed further in the following section.

7.3. Discussion

This study constituted the largest survey of NCOs at any human recombination hot spot to date. It highlighted many similarities and differences between the *SPRY3* and other previously studied human hot spots, as well as revealing important factors regulating the frequencies of NCOs at this and most likely other human recombination hot spots.

First, COs and NCOs co-localised at the *SPRY3* hot spot, such that regions of peak NCO activity coincided with the centre of the CO hot spot. This co-localisation of COs and NCOs had been observed at other previously characterised hot spots and was consistent with both types of events resulting from DSBs occurring within the same initiation zone (Jeffreys & May, 2004).

Second, in contrast to other hot spots, NCOs at the *SPRY3* hot spot were spread over a wider zone of nearly 4 kb. Since GC-tracts associated with NCOs are very short, they are expected to correlate with positions of DSBs (Jeffreys & May, 2004; Cole, Keeney & Jasin, 2010). Data from the *SPRY3* hot spot was thus consistent with DSBs initiating over a wider zone than previously anticipated. It is however, also possible that some of these widely distributed NCOs result from HJs that have undergone extensive branch migration. Interestingly, wider distributions of NCOs, spreading over regions >1 kb, has also been observed at hot spot PAR1-P, described in Chapter 4 (section 4.3.3.1) and at the mouse *A3* hot spot (Cole, Keeney & Jasin, 2010; see section 7.1.1).

Third, despite a wider distribution, NCOs at the *SPRY3* hot spot showed steep gradients in activity, declining rapidly upstream of SNP 6882 and downstream of SNP 0442. This was similar to observations at previously characterised hot spots and was consistent with frequencies of DSBs declining on either side of a central initiation zone (Jeffreys & May, 2004).

Also, this extensive survey of NCOs showed for the first time large-scale variation in NCO:CO ratios at a single hot spot, extending from 4:1 to 1:10. This is comparable to the degree of variation previously documented *between* different hot spots and which was thought to reflect hot spot-specific biases in the resolution of recombination intermediates. Data from the *SPRY3* hot spot however suggests that the CO:NCO preference varies greatly amongst different men at the same hot spot and hints that this decision is likely subject to complex regulation and/or stochastic variations. It is interesting however, that the average NCO:CO ratio of 2:1 at *SPRY3*, is fairly similar to the 3:1 ratio suggested from cytogenetic data (Baudat & de Massy, 2007).

In addition to the above, this survey revealed that NCO-frequencies at this hot spot were regulated in-*trans* by meiosis-specific protein PRDM9 and in-*cis* by a central hot spot SNP 0442. The effects of these factors are described in detail below.

7.3.1. *PRDM9* status greatly influences frequencies of NCOs at the *SPRY3* hot spot

Work described in Chapter 6 showed that COs at the *SPRY3* hot spot were triggered by the PRDM9 A-variant, with *PRDM9* A/A men showing significantly higher CO frequencies than A/N and N/N men ($P<0.05$; see Chapter 6, section 6.2.5). The survey of NCOs at this hot spot showed a similar effect of *PRDM9* on frequencies of NCOs too. Thus, median NCO frequencies at 0442 were ~55-times higher for *PRDM9* A/A men compared to N/N men, who were highly suppressed for NCO activity at the *SPRY3* hot spot ($P<0.05$). Similar effects of PRDM9 on CO and NCO frequencies imply that this *trans*-acting factor influences recombination initiation, rather than the CO/NCO decision (see Figure 7.4). This is also consistent with data from the mouse hot spots *Psmb9* and *Hlx1* and further suggests that the role of the protein is conserved from mice to humans (Parvanov *et al.*, 2009; Grey *et al.*, 2009). It is however noteworthy that PRDM9 was not the only factor influencing frequencies of NCOs at the *SPRY3* hot spot. Indeed *cis*-acting SNP 0442 played a very significant role, as described below.

7.3.2. Central SNP 0442 influences frequencies of NCOs in *cis* at the *SPRY3* hot spot

CO data for the *SPRY3* hot spot suggested that the *cis*-acting SNP 0442 influenced recombination-initiation frequencies, such that these events were initiated more often on the haplotype carrying the active 0442T allele (see Chapter 6). NCO data at the hot spot revealed a similar picture, with T→C NCOs at 0442 largely exceeding C→T NCOs. This excess of T→C NCOs, on average led to a ~7.5:1 bias in NCOs at this marker, in favour of the 0442C allele.

Interestingly, this bias is far greater than the ~2:1 0442 GC-bias accompanying COs at this hot spot and enhances the level of over-transmission of the recombination-suppressing C allele from 66:34 to 74:26. At the population level, this is expected to strengthen the meiotic drive at this hot spot, to an extent that will guarantee fixation of the hot spot attenuating 0442C allele within the next 60,000 ± 16,000 years, assuming historical effective population sizes (A. Jeffreys, unpublished data).

Differences in the extent of bias in GC with and without exchange observed at *SPRY3* is unprecedented for a human recombination hot spot. Furthermore, the bias in NCOs at 0442 seems to vary substantially from one man to another, whereas the bias in GCs accompanying COs remains relatively constant (see section 7.2.1). Also, the extent of the bias in NCOs appears to be related to CO:NCO ratios. These observations have allowed new insights into the likely pathways of NCO-formation in humans and are discussed in detail in the following paragraphs.

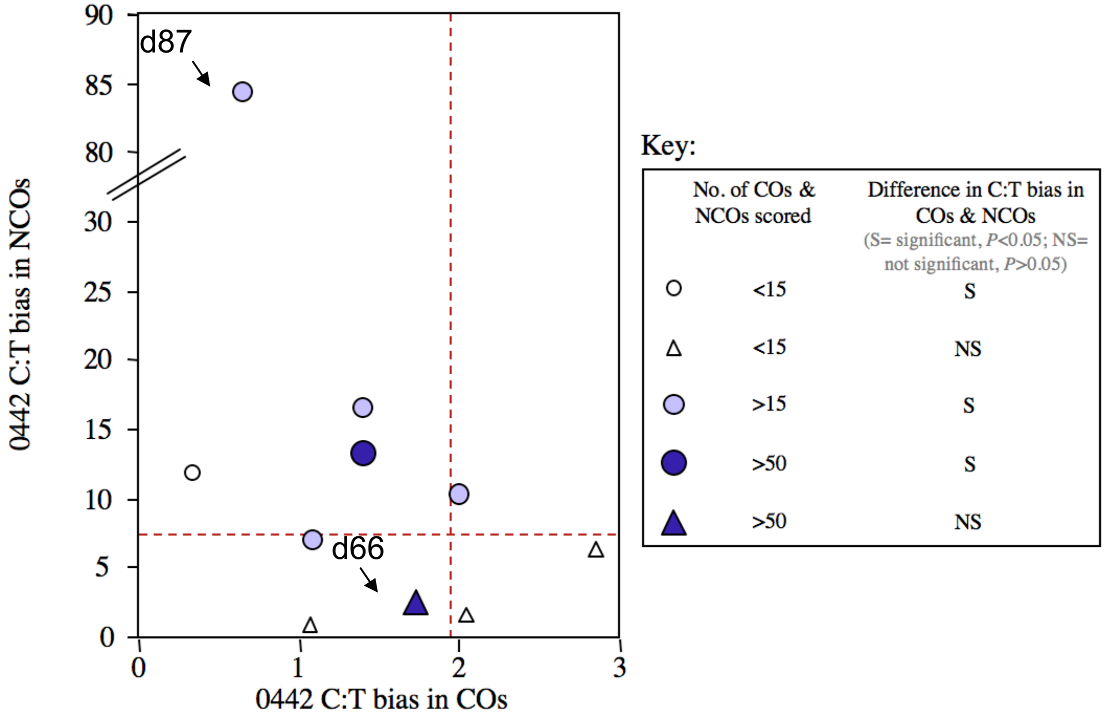
7.3.3. NCOs at the *SPRY3* hot spot may be generated via two distinct pathways

Previously, comparisons of extents of bias in GCs with and without exchange have only been made for one man and at one hot spot, namely *NID1* (Jeffreys & Neumann, 2005). Recombination activity at this hot spot was regulated in *cis* by a C/T polymorphism located

~70 bp away from the centre. Men heterozygous for this SNP showed TD and biased GC accompanying COs, in favour of the recombination-suppressing C allele. Analysis of NCOs in one man showed that these events too mirrored the bias observed in COs (2.3:1 and 2.7:1 respectively). This implied that the bias observed in both types of recombinants directly reflected the bias in recombination initiation frequencies on the active and suppressed haplotypes. Data from the *SPRY3* hot spot however, contradict the aforementioned scenario.

Comparisons of bias in GC accompanying COs and NCOs at the central *cis*-acting SNP 0442 could be made for 10 of 13 men assayed for both types of events. Two of the remaining men carried two *PRDM9* N alleles and were highly suppressed for both types of events and were thus excluded from analysis. COs were not recovered from both orientations in another man and he too was excluded from this analysis. The 10 remaining 0442 heterozygous men showed similar C:T bias in GCs accompanying COs, consistent with the average of 2:1 calculated in Chapter 6, section 6.2.3 (see Figure 7.6). The degree of C:T bias in NCOs however showed extensive variation ranging from 1:1 to >85: 1 and was significantly different to the bias in COs for 6 out of the 10 men included in the analysis (see Figure 7.7).

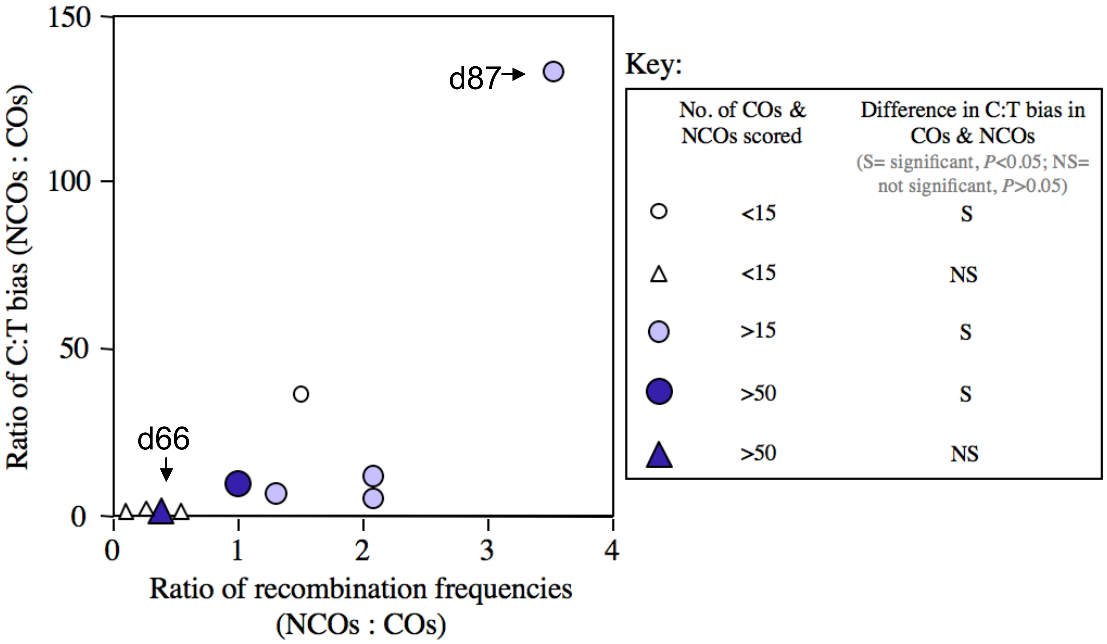
Figure 7.6 Comparisons of biased GC accompanying COs and NCOs at SNP 0442



Comparisons of 0442 C:T bias in NCOs vs. COs is shown for 10 men. The horizontal dotted line represents the median C:T bias in NCOs, whereas the vertical dotted lines represents the same for COs. Meaning of various symbols is indicated in the key.

Interestingly, men showing more extreme biases in 0442 C:T transmission ratios in NCOs, also showed an excess of these events in comparison to COs and this correlation was significant (*P*<0.01, Mann-Whitney U test). Thus, d87 who showed >85:1-C:T bias in NCOs as opposed to a ~2:1-C:T bias in GCs accompanying COs, also showed an excess of NCOs over COs (NCO:CO ⇒ 3.5:1). Conversely, d66, who showed similar C:T bias in COs and NCOs, showed ~2.6 times higher CO frequencies compared to NCO frequencies at 0442. This is further illustrated in the following figure.

Figure 7.7 Relationship between NCO:CO ratios and GC-biases



The figure shows relationships between NCO:CO ratios and GC bias in NCOs vs. COs in 10 men assayed at *SPRY3*. In general, men showing higher C:T bias in NCOs than COs, also show an excess of these events in comparison to COs. Thus d87, who showed the greatest difference in C:T bias in NCOs vs. COs (in favour of the former), also showed >3.5-fold higher NCO frequency than CO frequency. In contrast, d66 showed similar C:T bias in COs and NCOs and an >2.5 fold excess of COs, rather than NCOs.

If all NCOs were generated by a single pathway and if this pathway was the same as that used to generate COs, then all men assayed would have shown similar C:T bias in NCOs. Moreover, this bias would mirror the C:T bias seen in COs and reflect the initiation bias introduced via *cis* effects of 0442. This was however, not the case for all men assayed at *SPRY3* and some men showed significantly different GC biases in COs and NCOs (see Figure 7.6). Since the same men showed a large excess of NCOs over COs, it is tempting to speculate that a sub-set of NCOs in these men were generated by a second pathway of NCO-formation, which was more prone to biased GC. Indeed, data from *Mlh1*$^{-/-}$ mice are consistent with differential mismatch repair of heteroduplex DNA (hDNA) in NCOs, compared to COs, which are thought to be generated via a distinct pathway (Guillon *et al.*, 2005).

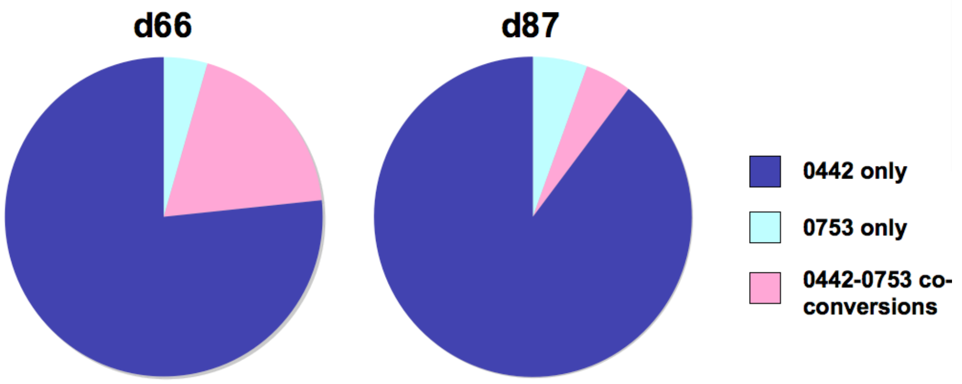
In order to investigate this further, GC tract lengths were compared between two men- one (d87) showing significantly greater C:T bias in NCOs compared to COs and the other (d66) showing remarkably similar GC-bias in the two types of events.

7.3.4. Comparisons of GC-tracts in two men showing very different extents of GC-bias in NCOs

In order to investigate GC-tracts in d87 and d66, numbers of co-conversions between two markers 0442 and 0753 (rs56750753), for which both men were heterozygous, were compared (Figure 7.8). d66, who showed similar GC-bias in COs and NCOs and greater numbers of COs, also showed significantly greater numbers of co-conversions than d87 (~19% and 5% of all NCOs at the two markers). Since 0442 and 0753 are 312 bp apart, this indicates a minimum NCO-tract length of 313 bp, which is not dissimilar to the 329 bp estimated for GCs accompanying COs. This is consistent with NCOs in this man being generated via the same pathway as COs, as predicted by the canonical DSBR model of recombination.

Significantly fewer co-conversions in d87 ($P<0.05$) is however consistent with shorter GC-tracts in this donor being generated by a pathway distinct from the DSBR pathway of CO formation. It is tempting to speculate that this second pathway of NCO-formation might be the SDSA pathway or one similar to it. It will be interesting to see if data from detailed surveys of NCOs at autosomal hot spots will be consistent with presence of two distinct pathways of NCO-formation or if this is a feature exclusive to recombination at pseudoautosomal hot spots.

Figure 7.8 Proportions of co-conversions observed in d87 and d66



The graph shows proportions of all NCOs at 0442 and 0753 that involve either or both markers. The difference in the proportion of co-conversions between the two men is significant. It should be noted that d87 is also heterozygous for SNP rs306882, which lies between 0442 and 0753. NCOs affecting this marker alone have not been considered above.

In summary data from this extensive survey has shed valuable light on the distributions and frequencies of NCOs at a human meiotic recombination hot spot and facilitated comparisons with the distributions and frequencies of COs at the same hot spot. The study has also shown two levels of regulation affecting NCOs, in the form of *trans*-acting PRDM9 and *cis*-acting SNP 0442. Interestingly, in contrast to COs, the aforementioned factors seemed to account for most of the variation in NCO activity seen at the *SPRY3* hot spot. Indeed, men with identical *PRDM9* and 0442 status showed at most, 11-fold differences in NCO frequencies. This was substantially lower than the ~100-fold residual variation in CO activity seen amongst *PRDM9* A/A, 0442 C/T men. This is suggestive of the existence of additional *cis*- and *trans*-acting regulators that may exclusively affect CO activity. If identified, these may shed valuable light on factors influencing the CO/NCO decision at human meiotic recombination hot spots.

Chapter 8. Identification and characterisation of an African-enhanced Xq/Yq pseudoautosomal hot spot

8.1. Introduction

Sperm analysis on *SPRY3* and on other LD hot spots identified from European population data analysed in the Berg *et al*. (2010) study showed that these were all activated by the common PRDM9 A variant. Further, men carrying two *PRDM9* N (non-A) alleles were highly suppressed for COs at all tested hot spots (Berg *et al.*, 2010). This finding raised several questions regarding the mechanisms of hot spot regulation in *PRDM9* N/N men, namely, are hot spots in these men regulated by a factor distinct from PRDM9 or are they simply tuned to a different set of *PRDM9* alleles. Two different studies provided some clues pertaining to this and are briefly discussed below.

8.1.1. Shifts in LD hot spot usage in Hutterites carrying the *PRDM9* 'I' allele

In order to investigate the influence of variation in the PRDM9 ZnF array, Baudat *et al*. (2010) analysed genome-wide LD hot spot usage in Hutterites carrying three different *PRDM9* alleles, namely, A, B and I. In contrast to *PRDM9* A and B alleles, the I allele encodes a protein variant that is not expected to recognise the 13-mer hot spot motif (Baudat *et al*., 2010). Individuals carrying this allele therefore provide a great resource for investigating the relationship between *PRDM9* allelic status and genome-wide LD-hot spot usage.

Comparisons of LD hot spot usage between *PRDM9* A/A and A/I individuals revealed significant differences, with A/I individuals showing a 70% decrease in genome-wide LD hot spot usage (Baudat *et al*., 2010). The authors suggested that this large disparity was a consequence of the *PRDM9* I allele activating a new set of hot spots, that were either too recent or too weak to leave their mark on population diversity data.

This study thus provided the first indications that individuals within the same population might be using different hot spots, depending on their *PRDM9* status. However, these shifts in hot spot usage could in theory be related to other hot spot specification systems whose effects are otherwise masked by that of the PRDM9 A variant (Berg *et al*., 2011). More direct analysis of COs occurring within hot spots activated specifically by PRDM9 A or I variants would be required to unequivocally assign shifts in hot spot usage to *PRDM9* allelic status.

8.1.2. Differences in PRDM9 zinc-finger array lengths influence recombination hot spot usage

Not long after the Baudat *et al*. (2010) study, Kong and colleagues (2010) published high-resolution sex-specific and sex-averaged recombination maps based on CO-events recovered from the analysis of 15,257 Icelandic parent-offspring pairs (also discussed in Chapter 1, section 1.5.2). The recombination maps were analysed in bins of 10 kb each and recombination hot spots defined where the genetic distance of a particular bin was 10-times greater than the genome average. Using this strategy, Kong *et al.* (2010) identified 4,762 male hot spots, within which 36.2% of genome-wide recombination events analysed occurred. They also reported 4,129 hot spots in females and 28% of all female recombination events recovered occurred within these.

Kong *et al*. (2010) also identified significant associations between multiple SNPs within *PRDM9* and the proportion of recombination events occurring within hot spots reported in the study. They further analysed PRDM9 ZnF-array lengths in 575 Icelanders and identified repeat lengths ranging from 12 to 15 units. Individuals carrying 12 or 13 ZnF-repeats showed similar levels of hot spot usage. Individuals carrying PRDM9 variants with 14 or 15 repeat units however showed a significant decrease in usage of hot spots identified by Kong *et al*. (2010). As with the Baudat *et al*. (2010) study, this finding also hinted that different PRDM9 variants, in this case the 12 or 13 variants and 14 or 15 variants, might activate different sets of hot spots.

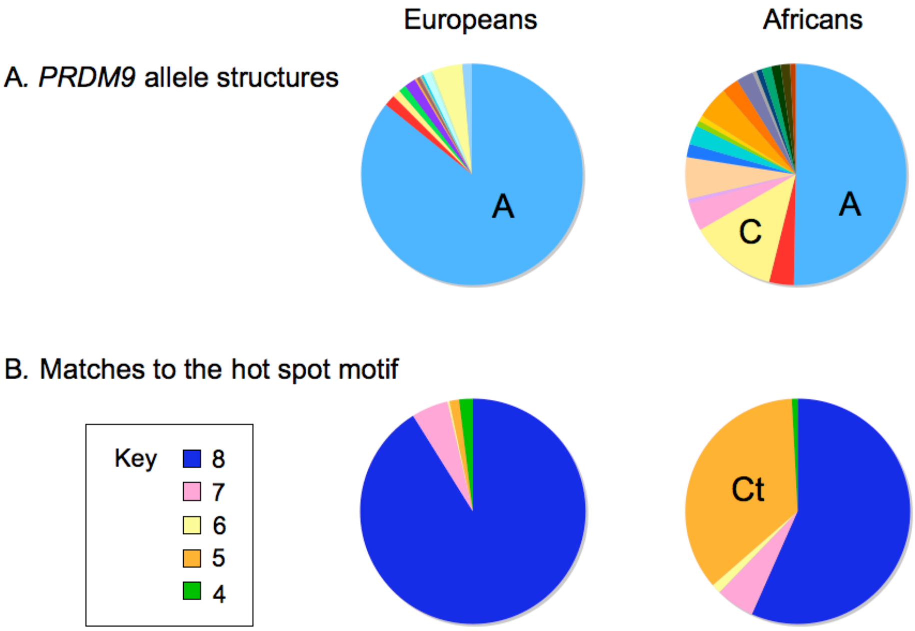
To investigate this further, Kong *et al*. (2010) constructed sex-specific recombination maps for carriers of PRDM9 variants with 12 or 13 repeat units and with 14 or15 repeat units. They reported that in carriers of 14 or 15 repeat-units, 18.3% of all recombination events occurred within hot spots identified in the study, as opposed to the 27% observed in carriers of 12 or 13 repeat variants. They suggested that reduced usage of hot spots identified in the Icelandic sex-averaged genetic map, in carriers of 14 or 15 repeat variants, was being compensated by use of other hot spots in these individuals. Indeed, 1,751 10 kb-bins that met the criteria for hot spots in the map for carriers of 14 or 15 repeat-units, did not do so for carriers of 12 or 13 units, suggestive of presence of different hot spots in carriers of different PRDM9 variants.

It should however be emphasised that many of the hot spots identified by Kong *et al*. (2010), which were specific to carriers of either 12 or 13 repeat units or 14 or 15 repeat units were defined on the basis of a single CO event. These may well represent a chance CO event, rather than a classical hot spot defined by the clustering of numerous of CO events in the same interval.

8.1.3. This work

Thus far, detailed sperm analysis has only been used to investigate PRDM9 regulation at European LD hot spots and these analyses have led to the in depth characterisation of hot spots activated by the common PRDM9 A variant. However, since Europeans show relatively low diversity of *PRDM9* alleles, they are a poor resource for investigating activation of different predicted hot spots by distinct *PRDM9* alleles (Berg *et al*., 2010). In contrast, African populations carry a wide array of *PRDM9* alleles, which encode variants with different predicted DNA-binding specificities relative to the A-variant. Particularly common amongst these non-A (N) alleles are the C allele and some 9 related alleles, all of which encode PRDM9 variants with identical predicted DNA-binding specificities (Berg *et* *al*., 2011; see Figure 8.1). These alleles are henceforth collectively referred to as C-type (Ct) alleles.

Figure 8.1 Diversity of *PRDM9* alleles and corresponding protein variants amongst Europeans and Africans



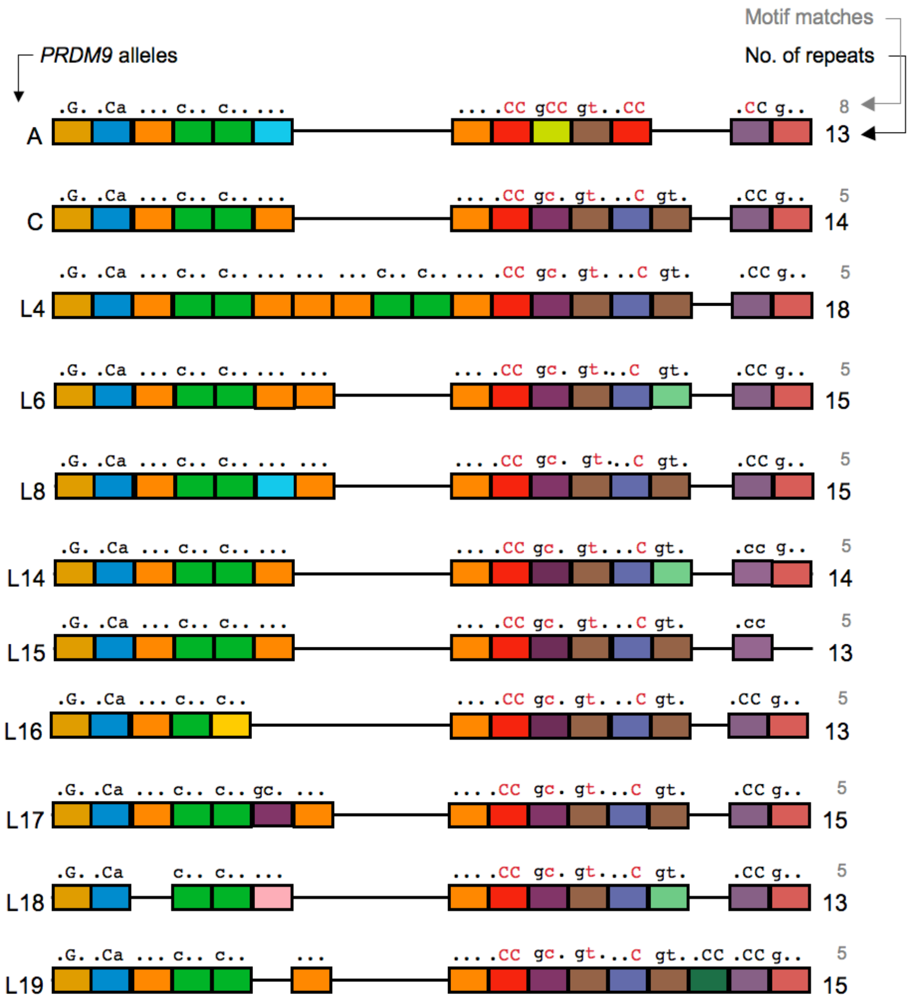
A. Diversity of *PRDM9* alleles in North European and African semen donor panels. The figure highlights greater allelic diversity amongst Africans than in Europeans. The *PRDM9* A allele is most common is both populations, but shows much higher frequency amongst Europeans. In Africans, the C and some C-type alleles also show high population frequencies (see B).

B. PRDM9 variants are classified according to the levels of matches of their DNA-binding motif to the 13-mer hot spot motif, CCNCCNTNNCCNC. The PRDM9 A-variant recognises all 8 underlined bases of the motif, whilst Ct variants recognise only 5 out of 8 bases.

Note: All four graphs are based on data from Berg *et al*. (2010) supplementary information.

PRDM9 Ct alleles all encode protein variants that are at best expected to recognise only 5 out of 8 bases (underlined) of the 13-mer European hot spot motif CCNCCNTNNCCNC, which is associated with ~40% of European LD hot spots (Myers *et al.*, 2008). Structures of PRDM9 Ct alleles and predicted DNA-binding specificities of corresponding protein variants are shown in Figure 8.2. Interestingly, the C allele and several C-related alleles encode protein variants containing 14/15 unit-long ZnF DNA-binding arrays. These 14/15 unit variants have been implicated in activating distinct sets of hot spots relative to the A-variant in the Kong *et al*. (2010) study. Semen donors who are carriers of these alleles are thus a valuable resource for directly verifying hot spot activation by *PRDM9* Ct alleles.

Figure 8.2 *PRDM9* Ct alleles observed in semen donor panel


Motif matches
No. of repeats
PRDM9 alleles
A .G. .Ca ... c.. c.. ... ... .CC gCC gt. .CC .CC g.. 8 13
C .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .CC g.. 5 14
L4 .G. .Ca ... c.. c.. ... ... c.. c.. ... .CC gc. gt. ..C gt. .CC g.. 5 18
L6 .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .CC g.. 5 15
L8 .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .CC g.. 5 15
L14 .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .cc g.. 5 14
L15 .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .cc 5 13
L16 .G. .Ca ... c.. c.. ... .CC gc. gt. ..C gt. .CC g.. 5 13
L17 .G. .Ca ... c.. c.. gc. ... ... .CC gc. gt. ..C gt. .CC g.. 5 15
L18 .G. .Ca c.. c.. ... ... .CC gc. gt. ..C gt. .CC g.. 5 13
L19 .G. .Ca ... c.. c.. ... ... .CC gc. gt. ..C gt. .CC .CC g.. 5 15

PRDM9 A and Ct alleles are shown. Various coloured boxes depict minisatellite repeat units, as indicated in the key. Sequences for all repeat units were previously described in Appendix 6.4. Each repeat encodes one ZnF; total numbers of ZnFs encoded by the allele are indicated alongside in black. Predicted DNA-binding specificities for each ZnF in the array are indicated above respective repeat arrays. Matches to the 13-mer hot spot motif CCNCCNTNNCCNC are shown in red and maximum numbers of such matches are indicated in grey to the right. Repeat units for all alleles are aligned with respect to allele A. Repeat types are as indicated previously.

The main aims of this study were to directly investigate whether carriers of PRDM9 Ct alleles use a distinct set of hot spots activated by specific Ct variants and to compare the characteristics of such hot spots with those activated by the common PRDM9 A variant. Once again, analysis of population diversity data was applied to identify suitable candidates for sperm analysis, as described below.

8.1.4. Strategy for identifying hot spots activated by non-A variants

Most LD hot spots in Europeans are likely to represent PRDM9 A-regulated hot spots, on account of the extremely high population frequency of this allele (Figure 8.1). In contrast, African LD hot spots are likely to represent ones regulated by other PRDM9-variants common in Africans, most notably Ct variants. Thus, regions harbouring an LD step in Africans, but not in Europeans, are likely to represent hot spots regulated by non-A PRDM9 variants, with Ct variants being the most likely activators.

In order to identify such 'African-specific' or 'African-enhanced' LD hot spots within the Xq/Yq PAR, HapMap phase II genotype data for four African populations as well as for an Indian and European population was analysed, as described in the next section. This work was part of a much larger survey aimed at identifying and characterising a panel of hot spots regulated by non-A variants (Berg *et al*., 2011). Key findings from this larger study will be discussed in section 8.3 of this Chapter.

8.2. Results

In an attempt to identify putative African-enhanced hot spots, LDU profiles across PAR2 were obtained from HapMap genotype data (release 27) for 6 populations. Thus, HapMap genotype data for 89 Utah residents of Northern and Western European ancestry (CEU), 83 individuals of African ancestry in SouthWest U.S.A. (ASW), 90 Yoruba from Nigeria (YRI), 89 Luhya from Kenya (LWK), 96 Maasai from Kenya (MKK) and 88 Gujarati Indians in Houston (GIH) were used to generate population-specific LDU maps. Metric LD maps were generated as described previously.

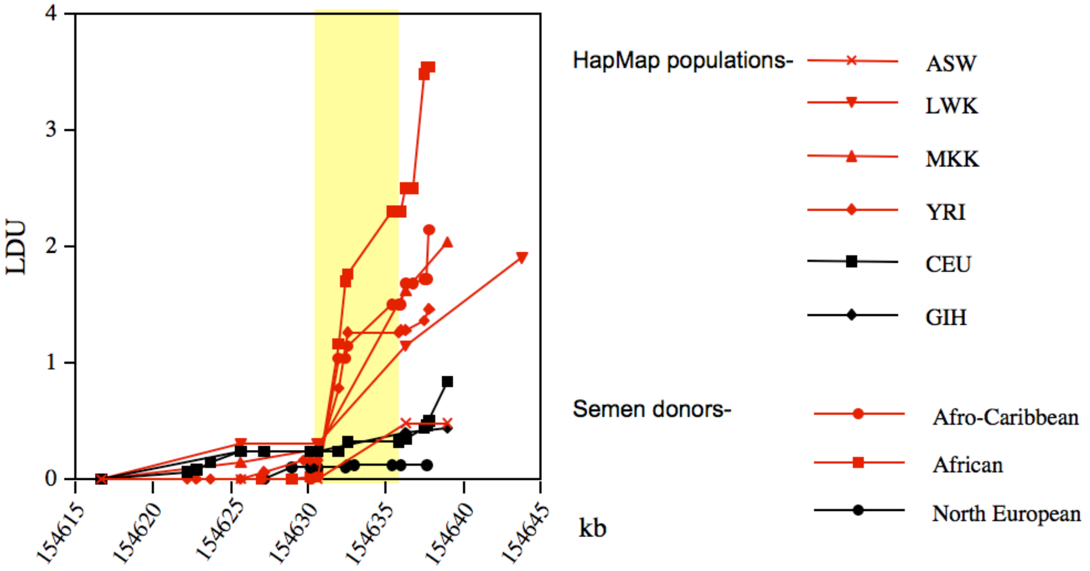
This analysis revealed an interval between 154,630 kb and 154,637 kb (ChrX; NCBI36/hg18) that showed the presence of an LDU step in at least three out of the four HapMap African populations, but not in the Indians (GIH) and the Europeans (CEU) (Figure 8.3). Since this region showed presence of historical recombination activity specifically in African populations, it was an ideal candidate 'African-enhanced' hot spot and was selected for further analysis in the semen donor panels.

8.2.1. High-resolution LD analysis in North European and African semen donor panels

The putative historical hot spot interval was first subjected to LD analysis in the semen donor panels (see Appendix 8.1 for the annotated sequence). To determine if the HapMap data could be replicated at higher resolutions in semen donor populations, which can be used for further hot spot characterisation. For this purpose, 32 SNPs between 154,627 and 154,638 kb (ChrX; NCBI36/hg18) were genotyped in 105 North European and 74 African semen donors. Genotyping strategy was similar to that previously described in Chapter 4 (section 4.2.1) and Chapter 5 (section 5.2.2). Details of PCR amplicons and ASOs used for genotyping are listed in Appendix 8.2 and 8.3 respectively. Genotypes for all informative markers for all semen donors typed are listed in Appendix 8.4.

Analysis of LD patterns in semen donor panels showed the presence of an LDU step of 3.5 LD units amongst men of African origin only (Figure 8.3) with, data from North European semen donors showing little evidence of historical recombination activity. This was consistent with data from the six HapMap populations studied and provided further evidence supporting the presence of an African-enhanced hot spot in the region.

Figure 8.3 LDU profiles across the target interval in HapMap and semen donor populations



LDU profiles across the interval of interest for six HapMap populations (ASW, LWK, MKK, YRI, CEU and GIH) and for men from two semen donor panels are shown. The African semen donor panel comprises 18 men of Afro-Caribbean origin and 56 men of Southern East-African descent (named 'African' in key). LDU profiles for these sub-sets are shown separately for clarity. All African populations are indicated in red. European populations are in black. Indians are similar to Europeans in terms of diversity of *PRDM9* alleles and are thus also indicated in black (Berg *et al*., 2011). The region of LD breakdown in Africans is highlighted in yellow. All coordinates are as per ChrX; NCBI36/hg18.

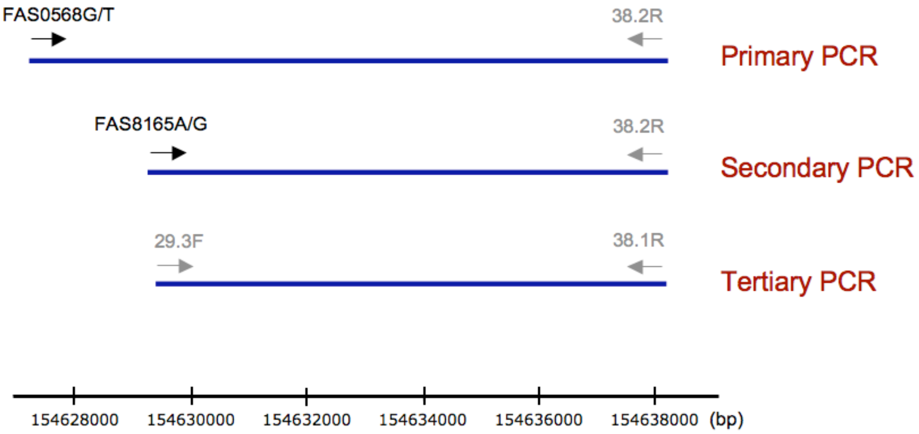
8.2.2. Analysis of the putative hot spot interval using sperm-based assays

In order to address whether the interval of higher historical activity thus identified represented a true contemporary African-enhanced hot spot and to establish whether the presence/absence of the hot spot in certain individuals or populations was linked to their *PRDM9* status, high-resolution sperm assays were carried out next.

LDU data from the African semen donor panel indicated breakdown of marker association from ~154,630 kb. However, the analysis failed to capture the 3′ end of the LDU step, making it difficult to select suitable selector sites 3′ of the LDU step. To overcome this issue, sperm assays were planned such that 5′ ASPs could be used in combination with 3′ universal primers. Using universal primers at the 3′ end would make it relatively easy to extend the assay further if necessary.

Two SNPs 5′ of the step, for which a large number of semen donors were heterozygous and which were present within a suitable genomic context, were chosen as 5′ selector sites. ASPs targeted towards these SNPs, rs5940568G/T and rs67398165A/G (since renamed rs495324), were designed and optimised as described previously (see Chapter 3, section 3.2.2). Next, a sub-set of men, prioritised on the basis of numbers of informative markers present in the assay interval and *PRDM9* status were chosen for analysis. Linkage phase of selector sites, as well as internal SNPs was determined as described in Chapter 4 (section 4.2.2.3). Finally, sperm assays were carried out, as shown in Figure 8.4, for a total of 25 North European and African men. Details of primers and PCR conditions used in sperm assays are listed in Appendix 8.2.

Figure 8.4 Strategy for a sperm assay across the putative hot spot interval



Strategy for the sperm assay used is shown. In the first two rounds of PCR, forward ASPs primers were used in combination with reverse universal primers to selectively amplify the chosen parental haplotype. A tertiary PCR was used to boost the yield of PCR products for dot blotting. All positions are as per ChrX, NCBI36/hg18.

This type of sperm assay was previously used to analyse recombinants at the PAR1-P hot spot (Chapter 4) and at the *SPRY3* hot spot (Chapter 6). Details of various steps of the assay, including recovery, mapping and analysis of recombinants can be found in the aforementioned Chapters (notably Chapter 4, section 4.2.2). Using this strategy, between 2,700 and 23,000 molecules were screened for recombinants per man in this study. CO data from the assay will be discussed presently; NCO data at two central SNPs will be discussed in the next section.

8.2.3. CO-breakpoints in the interval assayed cluster into a classical sperm CO hot spot

Of the 25 men assayed across this interval, 17 were assayed in both orientations. Between 0 and 88 COs were recovered per orientation per man, as listed in Table 8.1, with different men showing very different levels of CO activity. Distributions of COs in all men for whom sufficient numbers of events were recovered were consistent with the presence of a classical hot spot, wherein CO-breakpoints were normally distributed (Figure 8.5). Data from 6 men for whom at least 25 COs were recovered were used to calculate the least-squares best-fit values of hot spot centre and width, which were determined at 154,632,254 bp (ChrX) and 1.1 kb respectively. This hot spot will henceforth be referred to as the PAR2A hot spot.

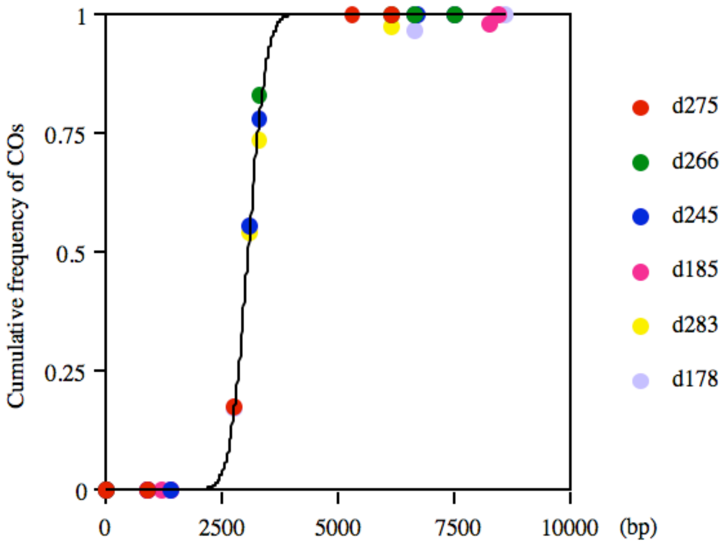
Table 8.1 CO data from sperm analysis in 25 men

CO data from 25 men are shown. 'Eu' indicates European and 'Af' indicates African. d178, d245, d262, d266, d275, d279, d281 and d282 were analysed in one orientation only. Remaining men were analysed using reciprocal assays.

Man	Origin	PRDM9 alleles	Molecules screened	Corrected COs	Frequency of COs (%)	Lower 95% C.I.	Upper 95% C.I.
d275	Af	C/A	4460	86	1.921	1.564	2.373
d266	Af	L14/A	5500	75	1.361	1.089	1.704
d245	Af	L14/A	4370	48	1.092	0.831	1.454
d185	Af	C/L12	2550	26	1.034	0.761	1.310
d283	Af	L4/C	13100	88	0.672	0.540	0.744

Man	Origin	PRDM9 alleles	Molecules screened	Corrected COs	Frequency of COs (%)	Lower 95% C.I.	Upper 95% C.I.
d279	Af	C/A	10010	63	0.633	0.493	0.804
d178	Af	L6/C	5730	30	0.525	0.368	0.734
d265	Af	C/A	3550	6	0.173	0.098	0.295
d262	Af	L19/C	4440	4	0.091	0.037	0.231
d242	Af	C/A	4690	3	0.065	0.023	0.187
d243	Af	B/L15	3560	2	0.056	0.017	0.202
d281	Af	A/A	2700	1	0.037	0.009	0.206
d238	Af	L22/A	3560	1	0.028	0.003	0.078
d244	Af	L21/A	4210	1	0.024	0.007	0.086
d261	Af	L4/A	4440	1	0.023	0.003	0.063
d274	Af	A/A	4660	1	0.022	0.003	0.060
d282	Af	A/L25	5730	0	0.000	0.000	0.064
d278	Af	A/A	3700	0	0.000	0.000	0.050
d260	Af	A/A	4440	0	0.000	0.000	0.042
d248	Af	A/A	4170	0	0.000	0.000	0.044
d247	Af	A/L11	4470	0	0.000	0.000	0.041
d239	Af	L6/A	3450	0	0.000	0.004	0.081
d181	Af	A/L11	3830	0	0.000	0.003	0.073
d105	Eu	A/A	3300	0	0.000	0.000	0.056
d50	Eu	A/A	3190	0	0.000	0.000	0.058

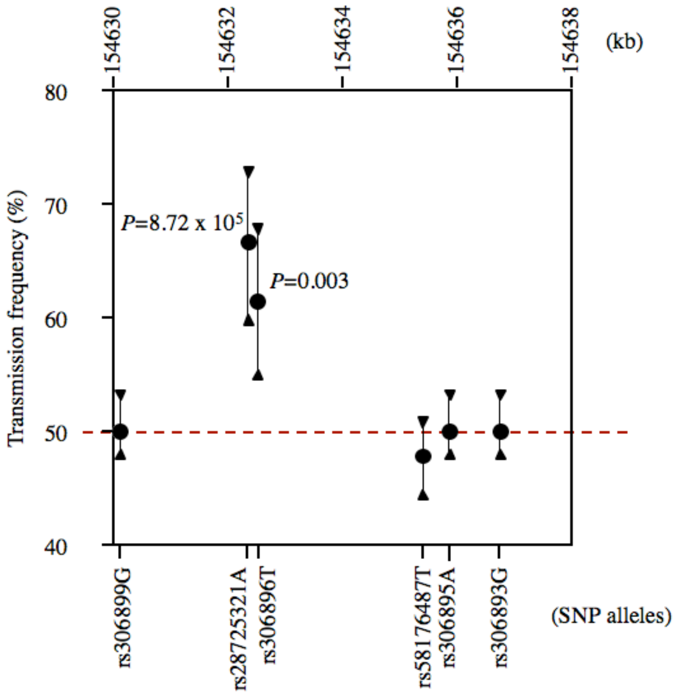
Figure 8.5 Cumulative frequency distribution of COs at the PAR2A hot spot



Data points of various colours show cumulative distribution of COs in the six men used to define morphology of the PAR2A hot spot. The black line represents the combined least-squares best-fit cumulative frequency distribution of COs for the PAR2A hot spot. All coordinates are relative to the assay interval, such that ‘0’ equals 154629253 bp (ChrX; NCBI36/hg18).

Interestingly, d283, but not d185, showed reciprocal CO asymmetry and significant TD at two central SNPs rs28725321 and rs306896, with A and T alleles being over-transmitted respectively (Figure 8.6). Unfortunately, not enough COs were recovered from other men assayed who were heterozygous at these SNPs to analyse this further. Also, 14 additional men carrying appropriate SNP heterozygosities could not be assayed owing to time constraints. Nonetheless, possible implications of this observation are discussed briefly in section 8.3 of this Chapter.

Figure 8.6 TD of alleles in d283

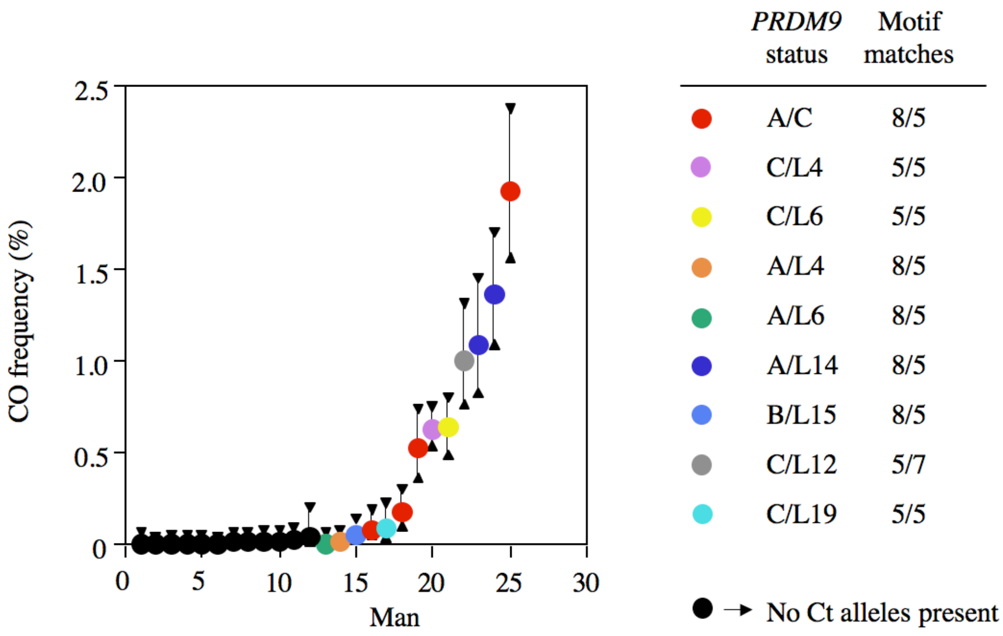


Transmission frequencies of markers from the over-transmitted haplotype into recombination progeny are shown. ASPs directed towards SNP alleles rs5940568G and rs67398165T were used to amplify the haplotype shown. SNPs rs28725321 and rs306896 both show significant TD. All positions are as per ChrX; NCBI36/hg18.

8.2.4. CO activity at hot spot PAR2A is triggered by PRDM9 Ct variants

CO frequencies at hot spot PAR2A showed >80-fold variation, ranging from 1.921% to <0.022%. Classifying men assayed at the hot spot according to their *PRDM9* status, showed profound influence of the *trans*-acting factor on CO frequencies. Thus, all men showing high CO frequencies at hot spot PAR2A carried at least one Ct allele and this correlation between hot spot activity and the presence of Ct alleles was significant (P=1x10^{-5}; Mann-Whitney U test). Furthermore, mean CO frequencies in *PRDM9* Ct carriers were >50-fold higher compared to men who lacked such alleles (Table 8.1 and Figure 8.7).

Figure 8.7 Correlation between hot spot activity at PAR2A and presence of *PRDM9* Ct allele(s)



Men assayed at hot spot PAR2A are classified according to the presence/absence of Ct alleles. CO frequencies for carriers of these alleles indicated in various colours shown in the key, others are represented in black. Vertical bars indicate upper and lower 95% C.I.s. For CO frequencies for all men and their PRDM9 status, see Table 8.1. Median CO frequency for carriers of Ct alleles was 0.525%, whereas that for men lacking such alleles was 0.000%.

Interestingly however, not all Ct alleles activated the hot spot to the same extent. Thus, whilst men carrying *PRDM9* C or L14 alleles showed the highest CO frequencies at the hot spot, Ct alleles L4 or L6 failed to activate the hot spot. Hot spot activation by distinct sub-sets of Ct alleles was also observed at autosomal hot spots analysed as part of this survey and is discussed further in section 8.3 (Berg *et al*., 2011).

In addition to differences in hot spot activity amongst men carrying various Ct alleles, variation in CO frequencies were also observed amongst men carrying the same *PRDM9* alleles. Thus, CO frequencies amongst three *PRDM9* C/A men assayed at PAR2A varied between 0.173% and 0.633%.

It is possible that additional *cis*-acting factors contribute towards this variation, but these could not be readily identified as haplotypes of the three C/A men differed at several positions. The semen donor panels include at least 10 additional *PRDM9* C/A men. Sperm analysis in these men might have shed additional light on this, but could not be carried out owing to time constraints.

8.2.5. NCOs at the PAR2A hot spot are also regulated by PRDM9 and activated by Ct variants

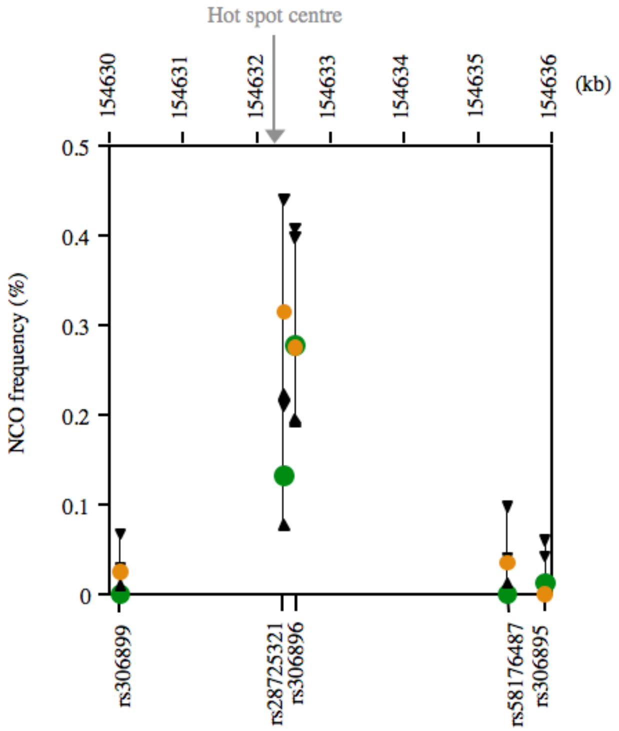
In addition to COs, hot spot PAR2A was also active for NCO. NCOs co-localised with COs in all donors who were informative for markers near the hot spot centre and for whom sufficient numbers of events could be recovered. Moreover, markers closest to the centre showed the highest NCO activity, which is consistent with both types of events resulting from recombination-initiating DSBs occurring within a central initiation zone. Also, frequencies of NCOs showed a steep decline in activity, consistent with data from other hot spots, including *SPRY3*. This sharp gradient in NCO-activity is evident from the profile shown in Figure 8.8, which represents data from >100 NCOs recovered from d283. Data from d283 also suggested that most NCO tracts at hot spot PAR2A are short. Thus, only 40-50% of NCOs at two central SNPs rs28725321 and rs306896, located just 157 bp apart, involved co-conversion of both SNP sites, with the remaining events affecting one or the marker only.

Unfortunately however, a significant number of men assayed were not very informative over the hot spot centre, which is crucial for the effective detection of NCOs. Moreover, in order to make valid comparisons between donors, they must informative at the same sites.

Hence, analysis of NCOs at PAR2A was limited to two central SNPs, which were each informative in 9 men assayed at the hot spot.

Both SNPs, rs28725321 and rs306896, were very close to the hot spot centre, located 122 and 279 bp away from the centre respectively. NCOs at rs28725321 were analysed using reciprocal assays for 8 out of 9 men heterozygous for the marker. d283, who showed significant TD in COs at this SNP, also showed biased NCOs at the marker (*P*=0.01; see Figure 8.8). Similar to COs, the A allele of the SNP was over-transmitted in NCOs as well. Biased NCOs was not observed in any of the other donors analysed, however, this was probably due to the low sample sizes of NCOs recovered in these men. This is discussed further in section 8.3 of this Chapter.

Figure 8.8 Biased NCOs at SNP rs28725321 in d283



Distributions and frequencies of NCOs in d283 are shown. Orange and green circles are used to distinguish NCOs from the two parental haplotypes. Frequency of C→A NCOs at rs27825321 (orange) is significantly greater than the frequency of A→C NCOs (green), leading to over-transmission of the A allele in NCO products (*P*=0.01). The centre of the CO hot spot is arrowed. All positions are as per ChrX, NCBI36/hg18.

NCOs at rs306896, located 157 bp-away from rs28725321, were analysed using reciprocal assays in 4 out of 9 men informative for the marker. No evidence for biased NCOs was observed for this SNP, even in d283 for whom 57 events were analysed at this site. Absence of biased NCOs at rs306896 in d283 suggested that NCO tracts at this hot spot (at least in this man) were extremely short, such that rs306896 was not passively swept into biased NCOs occurring at rs28725321, located just 157 bp away.

The *PRDM9* status of semen donors influenced NCO frequencies at both rs28725321 and rs306896, with carriers of Ct alleles showing significantly higher NCO frequencies than men lacking these alleles (*P*<0.01; see Table 8.2 and Figure 8.9).

Table 8.2 Frequencies of NCOs in men heterozygous for central SNPs rs28725321 and rs306896

NCO data, including numbers of events scored, molecules screened, frequencies of NCOs and upper and lower confidence limits on these are shown for two central hot spot SNPs. Carriers of *PRDM9* Ct alleles are indicated in red, whilst men lacking these alleles are in black. Men are listed in descending order of NCO frequencies.

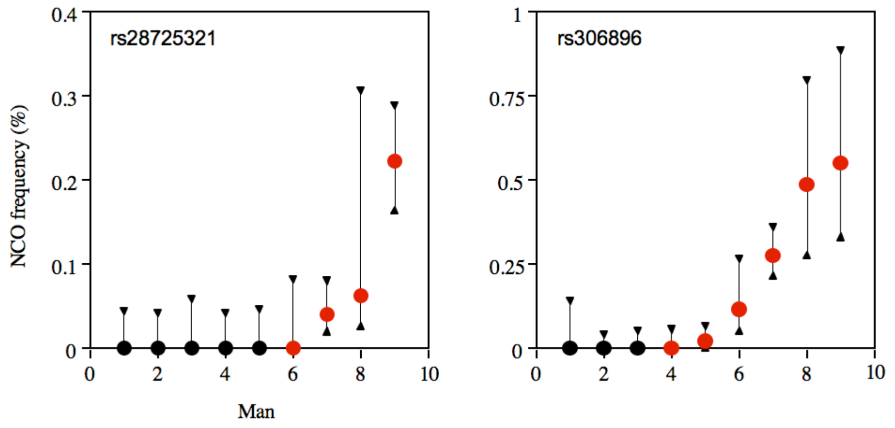
a. NCOs at rs27825321

Donor	*PRDM9* status	Corrected NCOs	Molecules screened	Freq. of NCOs (%)	Lower 95% C.I.	Upper 95% C.I.
d283	L4/C	48	22130	0.222	0.164	0.288
d245	L14/A	2	3260	0.062	0.026	0.305
d279	L6/C	7	18150	0.040	0.019	0.079
d244	L22/A	0	8370	0.000	0.000	0.044
d239	L4/A	0	6900	0.000	0.004	0.081
d261	L21/A	0	8870	0.000	0.000	0.042
d50	A/A	0	6370	0.000	0.000	0.058
d247	A/A	0	8950	0.000	0.000	0.041
d248	A/A	0	8300	0.000	0.000	0.045

b. NCOs at rs306986

Donor	*PRDM9* status	Corrected NCOs	Molecules screened	Freq. of NCOs (%)	Lower 95% C.I.	Upper 95% C.I.
d266	L14/A	15	2780	0.549	0.329	0.887
d245	L14/A	13	2780	0.483	0.275	0.797
d283	L4/C	57	20490	0.276	0.215	0.360
d262	L19/C	5	4400	0.117	0.050	0.266
d282	L6/A	1	5730	0.018	0.000	0.064
d281	A/A	0	2670	0.000	0.001	0.138
d239	L4/A	0	6860	0.000	0.000	0.054
d247	A/A	0	8950	0.000	0.000	0.041
d278	A/A	0	7400	0.000	0.000	0.050

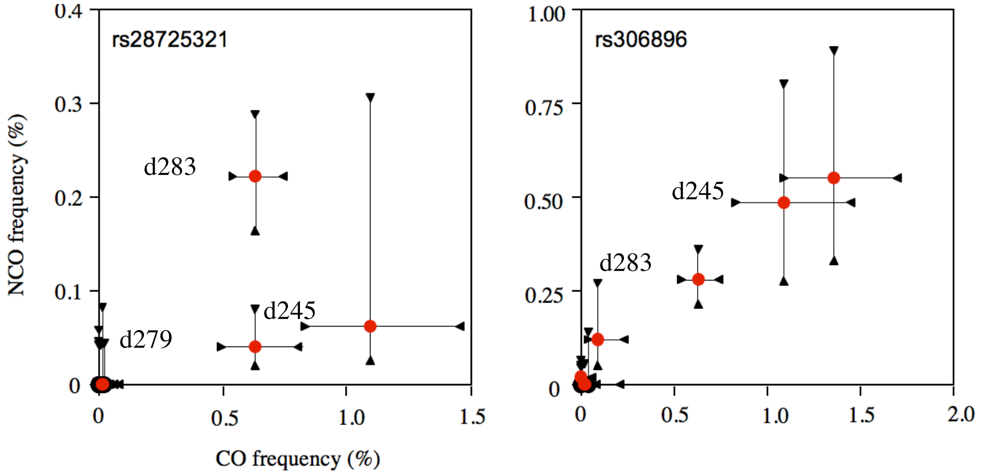
Figure 8.9 Variation in NCO frequencies at central SNPs with *PRDM9* Ct status



Frequencies of NCOs at two central hot spot SNPs are shown. Carriers of Ct alleles are indicated in red and men lacking these alleles are represented in black. Frequencies of NCOs at both markers correlate with *PRDM9* Ct status, with carriers of Ct alleles showing higher NCO frequencies, compared to men who lack these (*P*=0.048 and 0.119 for rs28725321 and rs306896 respectively). Vertical lines represent 95% C.I.s. Note: Man 1-9 are different for the two SNPs.

Two men, namely d245 and d283, who were assayed for NCOs at both rs28725321 and rs306896, showed higher NCO frequencies at the latter (Figure 8.10). This was curious considering that rs306896 was located ~150 bp further away from the centre, compared to rs28725321 and suggested that the true centre of the recombination hot spot in at least these men was 100-200 bp more proximal than estimated previously.

Figure 8.10 Comparisons of CO and NCO frequencies at the PAR2A hot spot



Relative CO and NCO frequencies are shown for 9 men assayed at the hot spot. It should be noted that NCO frequencies provided here represent those at rs28725321 and rs306896 respectively. Carriers of Ct alleles are shown in red and men lacking such alleles are in black. Vertical and horizontal bars represent 95% confidence intervals on CO and NCO frequencies respectively.

As at hot spot *SPRY3*, NCO frequencies at PAR2A varied proportionately with CO frequencies, with carriers of C and L14 *PRDM9* alleles showing the highest CO as well as NCO frequencies. This correlation was particularly clear at SNP rs306896 and less so at rs28725321. The latter was however, most likely a consequence of low numbers of NCOs (<10) analysed for d245 and d279 (indicated in Figure 8.10). Moreover, men lacking Ct alleles were suppressed for COs and NCOs including markers rs306896 and rs28725321, consistent with PRDM9 influencing the frequency of initiation of recombination events, rather than the CO/NCO decision.

8.3. Discussion

This Chapter describes the identification and characterisation of the first pseudoautosomal *PRDM9* Ct-regulated African-enhanced hot spot from analysis of population diversity data. The hot spot, named PAR2A, lies ~18 kb proximal to the *SPRY3* hot spot and shares many characteristics with its A-regulated counterpart.

8.3.1. Hot spot PAR2A shares many features with the *SPRY3* hot spot

Mapping of COs at hot spot PAR2A showed that CO-breakpoints at this hot spot clustered into a 1.1 kb-wide interval. This hot spot width is very similar to that of the *SPRY3* hot spot (1 kb) and indeed to previously characterised autosomal PRDM9 A-regulated hot spots. Likewise, distributions of CO-breakpoints that were compatible with normal distributions were also akin to that seen at *SPRY3* and other classical hot spots. These observations indicate that whilst different hot spots might be activated by distinct PRDM9 variants, downstream processing of recombination intermediates is likely the same for all of them.

This view is further supported by similarities in the distributions of NCOs between hot spots *SPRY3* and PAR2A. Although NCOs at PAR2A were not studied as extensively as at *SPRY3*, key features of NCO distributions, including co-localisation of peaks in CO and NCO activity and steep gradients of NCO frequencies (see Figure 8.8), appeared to be similar at the two hot spots.

Whilst sufficient data was not available at PAR2A to estimate NCO tract lengths, limited data from d283 suggested that most NCO-tracts were short, with ≤ 50% of NCOs at SNPs rs2875321 and rs306896 involving both of these markers, which were located just 157 bp-apart. More detailed analysis of NCOs in additional men is required to obtain a more robust and accurate NCO profile for the hot spot.

Aside from similarities in the distribution of recombination events, data from PAR2A also hinted that similar forces of hot spot evolution might be operating at this and at the *SPRY3* hot spot. As described in Chapters 6 and 7, *cis*-acting SNP 0442 caused biased gene conversion in COs as well as in NCOs at the *SPRY3* hot spot, with over-transmission of the recombination-suppressing 0442C allele. This resulted in a form of meiotic drive at the hot spot that guarantees its extinction within the next 60,000 ± 16,000 years (Sarbajna *et al.*, submitted). TD in COs and biased NCOs at the PAR2A hot spot observed in d283, suggested that a similar phenomenon might be operating at this hot spot too, with recombination events initiating more frequently on the haplotype carrying the rs28725321A allele (see Figure 8.6 and Figure 8.8). However, at the time of writing, sufficient data was not available to investigate this further. It should be noted however, that extents of bias seen in COs and NCOs in d283, were very similar to each other and to bias in COs seen at hot spot *SPRY3* (~2:1). It will be interesting to see if data from other men is consistent with this and if indeed, hot spot *SPRY3* is unique amongst pseudoautosomal hot spots, in terms of possessing two potential pathways of NCO-formation.

In contrast to SNP 0442 at hot spot *SPRY3*, the over-transmitted allele of SNP rs28725321 at the PAR2A hot spot, did not interrupt the predicted Ct motif CCNCNNTNNNCNTNNC. Indeed, no significant matches to the Ct motif were detected anywhere within the PAR2A hot spot. It should be noted however, that autosomal Ct hot spots containing perfect matches to the Ct-motif near their hot spot centres, have been identified (Berg *et al.*, 2011). It thus appears that like the A-variant, PRDM9 Ct-variants can activate hot spots either containing or lacking matches to their predicted DNA-binding motif and raises further questions regarding the precise mechanism of hot spot activation by PRDM9.

Finally, data from both *SPRY3* and PAR2A hot spots are consistent with PRDM9 influencing frequencies of recombination initiation, rather than the CO/NCO decision. Thus, *PRDM9* A/A men are suppressed for both COs and NCOs at PAR2A, whilst *PRDM9* C/A and L14/A men are active for both types of events (as shown in Figure 8.10).

8.3.2. Hot spot PAR2A is activated by at least two different *PRDM9* Ct alleles

As opposed to the *SPRY3* hot spot, which was activated specifically by the PRDM9 A-variant, hot spot PAR2A appears to be more broadly tuned to multiple, but specific, *PRDM9* Ct alleles. Thus, *PRDM9* C and L14 alleles activate the hot spot, whilst alleles L4 and L6 fail to do so. It should be noted however, that alleles L4 and L6 were designated as non-activators on the basis of observations in one man. As such, it remains possible that recombination activity in these men were suppressed by *cis*-acting influences. Additionally, it is also possible that other Ct alleles, not investigated in this study, also activate the PAR2A hot spot. A larger survey involving men carrying a wider spectrum of *PRDM9* alleles is necessary to address this further.

Hot spot activation by distinct *PRDM9* Ct alleles was also observed at autosomal Ct hot spots, with different alleles activating different hot spots (Berg *et al*., 2011). Thus, in contrast to PAR2A, hot spot 12B (on Chr12) was activated by allele L6 in d282. Similarly, allele L4 in d239 activated hot spot autosomal hot spot 12D, but not PAR2A (Berg *et al*., 2011). Since all Ct- variants have the same predicted DNA-binding specificities, differential hot spot activation by these variants is intriguing. Berg *et al.* (2011) suggest that local chromatin environments specific to individual hot spots might alter the DNA-binding specificities of PRDM9 variants, such that only a sub-set of these retain the ability to bind/activate the hot spot in question. Alternatively, it is also possible that other proteins might interact with PRDM9 via its KRAB domain to modulate its ability to bind and activate hot spots (Berg *et al*., 2011).

In summary, work described in this Chapter and in Berg *et al*. (2011) has shown that different recombination hot spots are tuned to different sets of *PRDM9* alleles. Thus, individuals in the same population, but carrying different *PRDM9* alleles (for instance, A/A and C/C) are likely to use entirely different sub-sets of hot spots. Conversely, populations containing a more diverse array of *PRDM9* alleles (*eg*. Africans) are likely to use a broader repertoire of hot spots, compared to others containing a more limited array of alleles (*eg*.

Europeans). Data from recently generated African-American genetic maps indeed show this appears to be true on a genome-wide scale (Hinch *et al*., 2011; Wegmann *et al*., 2011).

Analysis of population diversity data has been instrumental in the identification of these hot spots regulated by common PRDM9 variants. It will however, be a challenge to identify hot spots regulated by rarer variants. Nonetheless, availability of genetic maps for more diverse, recently admixed populations are likely to reveal additional interesting candidates for further analysis (Hinch *et al*., 2011; Wegmann *et al*., 2011).

Chapter 9. Discussion

Despite the important role of the human Xp/Yp PAR in male meiosis, relatively little was known about the fine-scale distribution of recombination events in this region. Indeed, prior to this work, recombination in PAR1 had only been analysed at high-resolution in the *SHOX* region, which led to the characterisation of the first and the only known PAR hot spot, namely the *SHOX* hot spot (May *et al*., 2002).

Work described here has led to the identification of three additional CO hot spots in PAR1, including a second, albeit less active hot spot SHOX-5, located ~18 kb away from the original *SHOX* hot spot (May *et al*., 2002; see Chapter 3). The first pseudoautosomal double hot spot was also identified in this survey, following on from analysis of HapMap population diversity data across the entire region (Chapter 4). Surveys on autosomes have shown that double hot spots are a relatively common feature of the autosomal recombination landscape (Webb, Berg & Jeffreys, 2008). Identification of the active CO hot spots PAR1-P and PAR1-D (Chapter 4) indicate that these features are prevalent in the Xp/Yp PAR and point towards common underlying recombination processes operating on a genome-wide level. It is at present unclear why hot spots exist in clusters. However, hot spot clustering would be expected if PRDM9 can only modify chromatin containing a particular mark/modification, which extends over very short distances.

Cytogenetic data for the Xq/Yq PAR or PAR2 had previously indicated that X-Y pairing in this region rarely persisted into mid- to late-pachytene, raising questions regarding the nature of genetic exchange in PAR2 and whether these occurred in the form of COs or NCOs (Chandley *et al.* 1984, 1987; Speed & Chandley 1990, Armstrong *et al.* 1994). Whilst linkage analysis in CEPH pedigrees and single-sperm typing studies did provide some evidence of reciprocal exchange in this region, these studies were limited in resolution and inferences were based on very few events (Freije *et al*., 1992; Schmitt *et al*., 1994; Li & Hamer, 1995; Lien *et al*., 2000; Flaquer *et al*., 2009). Thus, prior to this work, it was largely unclear whether this genomic region also harboured hot spots of recombination and if these were similar to their autosomal and indeed PAR1 counterparts.

Work described in Chapters 5, 6, 7 and 8 has provided unequivocal evidence supporting the presence of recombination hot spots in PAR2, with the two hot spots identified located just 18 kb-apart. Furthermore, both *SPRY3* and PAR2A are amongst the most highly active sperm hot spots identified to date, with CO frequencies as high as 0.91% and 1.92% respectively.

Analysis of COs and NCOs at all five PAR hot spots have provided valuable comparisons with their autosomal counterparts. They have also shed important light on the *cis*- and *trans*-acting factors that regulate recombination frequencies in these regions. The broader implications of the work described in this thesis will be considered in this Chapter.

9.1. Features and distributions of pseudoautosomal hot spots

Analysis of HapMap population diversity data revealed very different LD profiles for autosomes compared to PAR1, with the latter showing extensive LD breakdown across the entire region and very little in way of clear LD block structure (see Chapter 4, section 4.1.2). This raised questions regarding the distribution of recombination events in PAR1 and whether the need for an obligatory CO in this region had modified PAR1 recombination processes, such that COs were randomly distributed. Data from the extended *SHOX* survey, described in Chapter 3, showed that the extreme breakdown of marker association in PAR1 was likely to be the consequence of reduced hot spot spacing and increased levels of inter-hot spot recombination rates in this region, rather than the result of entirely randomly distributed recombination events in the history of the population. Thus, distance between hot spots *SHOX* and SHOX-5 was merely 18 kb, as opposed to the ~50 kb spacing estimated for autosomal hot spots (Myers *et al.,* 2005; Jeffreys, Kauppi & Neumann, 2001). Moreover, recombination activities in the interval separating the two hot spots were 5-12 cM/Mb (C.A. May, unpublished data), which was much higher than the 0.1

cM/Mb estimated for equivalent regions on autosomes (Jeffreys, Kauppi and Neumann, 2001; Jeffreys *et al*., 2005).

Interestingly however, none of the four PAR1 hot spots identified to date are as active as either of the two PAR2 hot spots. Given the increased hot spot density and inter-hot spot CO activity observed in the *SHOX* region, it is tempting to speculate that the need for an obligatory CO in PAR1 has modified the recombination landscape in this region such that it harbours a greater density of relatively less active hot spots, which also show a substantial leakage of recombination activity. In contrast, the LD profile in PAR2 is reminiscent of that on autosomes, with only one peak in historical recombination activity identified, at least in Europeans (Chapter 5). This suggests that PAR2 most likely harbours fewer, more active hot spots in comparison to PAR1. Nonetheless, the possibility that the most active PAR1 hot spots have simply escaped detection cannot be excluded either. Moreover, the limited nature of the surveys of recombination at PAR1 hot spots (in comparison to *SPRY3* and PAR2A) means that the full range of CO activities at these hot spots is currently not known and men who show much higher CO frequencies at SHOX-5, PAR1-P and PAR1-D may well exist. Preliminary analysis at hot spot PAR1-D in fact indicates that it shows considerable polymorphism in CO activity (discussed further in section 9.3.1).

In addition to potentially different distributions of recombination events in PAR1, data from mice also suggested that recombination in this region might be under distinct temporal and genetic control (Kauppi *et al*., 2011). Thus, Kauppi and colleagues (2011) showed that initiation of DSB-formation and chromosome pairing in the mouse PAR was delayed with respect to autosomes. Moreover, these late-forming DSBs were initiated by the *Spo11*α isoform, rather than the *Spo11*β isoform, which was required for DSB-formation in the rest of the genome (Kauppi *et al*., 2011). Since Spo11-isoforms are conserved between humans and mice, it seemed plausible that recombination in the human PARs, especially in PAR1, was also under distinct temporal and genetic regulation.

Striking similarities between the five PAR hot spots characterised as part of this work and autosomal hot spots, however suggested that even if DSBs in human PARs are generated by a distinct *Spo11* isoform, processing of these breaks were likely to be the same as on autosomes. Thus, both PAR1 and PAR2 hot spots showed very similar hot spot widths compared to autosomal hot spots, with the widest hot spot (SHOX-5) being ~2.2 kb-wide. Distributions of CO breakpoints for at least four out of five hot spots were also consistent with a normal distribution, as has been observed for autosomal hot spots. Moreover, in addition to similarities in hot spot morphology, this work also showed that recombination at PAR hot spots was regulated by the same mechanisms as on autosomes.

9.2. Pseudoautosomal hot spots are subject to the same forces of hot spot attenuation as their autosomal counterparts

Despite being nearly identical to chimpanzees at the sequence level, humans do not appear to share recombination hot spots with their closest evolutionary relatives (Winckler *et al*., 2005; Ptak *et al*., 2005). The apparent lack of hot spot conservation between the two species is suggestive of very high rates of hot spot evolution. This rapid turnover of meiotic CO hot spots is perhaps best exemplified by the S2 hot spot (Jeffreys & Neumann, 2009; see Chapter 1).

A discussed in Chapter 1, Jeffreys and Neumann (2009) showed that the S2 hot spot rose through an A→G mutation event that occurred ~70,000 years ago and was only active in carriers of this allele. However, sperm CO data at the hot spot revealed strong transmission distortion in A/G heterozygous men, with over-transmission of the recombination-suppressing A allele. Owing to this, ~0.04% of the hot spot-activating G alleles would be lost from the population per generation, in turn guaranteeing hot spot extinction within the next 120,000 ± 150,000 years.

Similar hot spot-attenuating influences were also identified at the *SPRY3* hot spot. Thus, sperm CO data from 28 men showed significant TD in COs, with over-transmission of the recombination-suppressing ‘C’ allele in men heterozygous for the *cis*-acting SNP 0442 (Chapter 6). In addition to COs, biased GC at 0442 was also seen in NCOs, resulting in a

C:T bias of 72:26 amongst recombinants. Combined with the high recombination frequencies at this hot spot, this bias resulted in the strongest meiotic drive observed at a human hot spot to-date. Indeed, simulation data suggested that this meiotic drive will not only guarantee fixation of the recombination-suppressing C allele ($P>0.999$), but will do so within ~60,000 ± 16,000 years, which is ~1/10th of the time required for the fixation of a non-driven mutation (Sarbajna *et al*., submitted).

Biased GC in COs and NCOs was also observed in one man at hot spot PAR2A (Chapter 8). Moreover, sperm CO data from donor 10 (d10) at hot spot SHOX-5 suggested the presence of borderline TD accompanying COs in this man. Given these observations, it is tempting to speculate that reciprocal CO asymmetry and TD of hot spot alleles are as common amongst PAR hot spots as they are on autosomes. This would in turn imply that despite their unique histories and the role of PAR1 in male meiosis, hot spots in these regions are under the influence of the same forces of hot spot evolution as their autosomal counterparts. Identification of PRDM9 as the global regulator of human recombination hot spots further supports this notion (Berg *et al*., 2010, 2011).

9.3. PRDM9 regulates CO frequencies in *trans* at PAR hot spots

In 2010, three different studies by Parvanov *et al*., Baudat *et al*. and Myers *et al*., provided evidence implicating PRDM9 as the global regulator of recombination hot spots in humans and mice. Direct analysis of effects of variation in sequences encoding the ZnF array of PRDM9 on sperm CO frequencies revealed profound effects at both autosomal, as well as PAR hot spots. Data from *SPRY3* and 11 other hot spots showed that all of these hot spots identified through analysis of European population-diversity data were activated by the PRDM9 A-variant common in Europeans (Berg *et al*., 2010), which was expected to bind the 13-mer sequence motif enriched within ~40% European LD hot spots (Myers *et al*., 2010). Further, analysis of sperm COs and NCOs at African-enhanced LD hot spots, including PAR2A, revealed a second class of hot spots tuned to specific subsets of *PRDM9* Ct alleles, common in Africans (Berg *et al*., 2011), as discussed below.

9.3.1. Activation of hot spots by the A-variant common in Europeans

Data from the *SPRY3* hot spot showed that this too was activated by the PRDM9 A-variant, in a manner similar to other autosomal hot spots identified from analysis of European population diversity data (Berg *et al*., 2010). Furthermore, CO activity was proportional to the number of A alleles carried by an individual, with *PRDM9* A/A and A/N men showing significantly higher CO frequencies than N/N men ($P \leq 0.05$; Chapter 6).

In addition to demonstrating hot spot-activation by specific *PRDM9* alleles, data from *SPRY3* and other hot spots studied by Berg *et al.* (2010) also brought to light an unexpectedly complex relationship between PRDM9-variants and their predicted DNA sequence binding motifs. *SPRY3* was one of the five hot spots in the study that contained nearly perfect matches to the predicted binding sequence of the PRDM9 A-variant near their centres. On account of this, it seemed possible that the A-variant activated these hot spots by directly binding its predicted sequence motif, as suggested by Baudat *et al*. (2010). Intriguingly however, data from five other autosomal hot spots indicated that the A-variant also activated hot spots lacking matches to its predicted binding motif. Equally perplexing was the fact that several other *PRDM9* alleles encoding protein variants with the same predicted DNA-binding specificities as the A-variant, were nonetheless incapable of activating specific sperm CO hot spots (Berg *et al*., 2011). Instances of this were also observed at hot spot *SPRY3* (see Chapter 6). Thus, a man carrying *PRDM9* alleles A and L9, which encoded a protein with the same predicted binding-specificity as the A-variant, showed CO frequencies >6-fold lower than the median for A/A men (with the same 0442 status) at the *SPRY3* hot spot. Needless to say, these observations raised several intriguing questions with regards to the nature of interactions between PRDM9 and hot spot sequences and the role the 13-mer binding motif, if any. This is discussed further in the context of African-enhanced hot spots in the next section.

Since work on both PAR1 hot spots was carried out prior to the identification of PRDM9 as a global regulator of meiotic recombination events in humans and mice (Baudat *et al*., 2010; Parvanov *et al*., 2010; Myers *et al*., 2010), the influence of PRDM9 at these hot spots were not investigated in detail. However, available data suggests that these hot spots too are regulated by PRDM9 and activated by the PRDM9 A-variant.

First, work described in Chapter 3 showed a 7-fold difference in CO frequencies between the two men assayed at the SHOX-5 hot spot. Analysis of *PRDM9* status of these men revealed that d10, who showed higher CO frequencies, carried two *PRDM9* A alleles, whereas d62 carried only one copy of this allele, in combination with allele L20. Lower CO frequencies in the *PRDM9* A/N man was consistent with hot spot-activation by the PRDM9 A-variant, in a manner similar to other hot spots analysed by Berg *et al*. (2010). However, differences in CO frequencies between the two men could in theory, also be the result of *cis*-acting influences on CO activities or consequences of differences in DNA quality and/or PCR efficiency of the two sperm assays. Analysis of COs in additional *PRDM9* A/A, A/N and N/N men are undoubtedly required in order prove a statistically significant effect of *PRDM9* status on CO frequencies at this hot spot.

Second, preliminary data for hot spot PAR1-D from work carried out subsequent to that described in Chapter 4 was consistent with the PRDM9 A-variant activating this hot spot. Thus, CO data from 6 PRDM9 A/A men, 7 A/N and 1 N/N at the PAR1-D hot spot, showed highly suppressed CO activity in the N/N individual ($P<0.07$; Mann-Whitney U test). Further, median CO frequencies for A/A men were twice as high as those for A/N men. These data were however based on agarose gel electrophoretic analysis of secondary PCR products from a classical CO assay (described in Chapter 3) and were not validated through CO mapping. It was also unclear whether the low CO frequencies in the single N/N men assayed at PAR1-D were due to the inability of *PRDM9* N alleles to activate the hot spot, since *cis*-acting influences or CO interference from high activities at the neighbouring PAR1-P hot spot could just as well have caused this suppression. Analyses of COs in additional *PRDM9* N/N men are required to shed further light on this. It is further important to map the observed COs to determine whether these are consistent with the hot spot identified in d74 (Chapter 4). Finally, it will be interesting to see if CO frequencies at the two neighbouring hot spots vary in relation to one another as a consequence of CO interference and also if both are activated by the same *PRDM9* alleles.

Suppression of all twelve recombination hot spots studied by Berg *et al*. (2010) in *PRDM9* N/N men raised questions regarding mechanisms of hot spot-regulation in these men, that is, were they were also regulated by PRDM9 but activated by a distinct set of alleles. Analysis of COs and NCOs at African-enhanced LD hot spots showed that this was indeed the case, with identification of several hot spots that were tuned to specific sub-sets of C-type alleles.

9.3.2. Activation of African-enhanced hot spots by Ct variants

The second Xq/Yq PAR hot spot (PAR2A) characterised as part of this work, was identified in a screen for intervals that showed evidence of LD breakdown African populations, but not in Europeans and Indians, in a bid to identify hot spots activated by non-A PRDM9 variants common in Africans, but rarer in the other two populations (Berg *et al*., 2011). Thus, in contrast to hot spot *SPRY3* and possibly SHOX-5 and PAR1-D, hot spot PAR2A was not activated by the PRDM9 A variant (Chapter 8). Instead, this hot spot was activated by a distinct set of *PRDM9* Ct alleles, as possibly expected from analysis of population diversity data.

Mean CO frequencies amongst carriers of Ct alleles were >50-fold higher compared to men who lacked the same. Interestingly however, hot spot PAR2A lacked an obvious match to the predicted Ct-sequence motif. This again suggested that the role of predicted DNA-binding motifs in PRDM9-mediated hot spot activation, if any, were likely to be complex as well as subtle. Also rather unexpectedly, not all Ct alleles activated the hot spot, despite encoding *PRDM9* variants with identical predicted DNA binding specificities. Thus, while alleles C and L14 activated the hot spot, alleles L4 and L6 failed to do so. The inability of allele L6 to activate was particularly surprising, since the C-terminal region of the ZnF array encoded by this allele (which is predicted to play an important role in hot spot activation) was remarkably similar to that encoded by hot spot-activating allele L14. It is worth mentioning however, that L4 and L6 alleles activated other hot spots in the same men (Berg *et al*., 2011).

It is at present largely unclear why PRDM9 variants with identical DNA-binding specificities differ to such great extents in their ability to activate specific hot spots. Detailed *in vitro* binding assays will be instrumental in addressing this further. It will also be interesting to see if future studies reveal any influence of local chromatin environment at hot spots on the ability of PRDM9 to bind/activate them and whether proteins that interact with PRDM9 also modify its ability to activate hot spots, as has been suggested by Berg *et al*. (2011).

9.4. Additional regulators of CO activity at pseudoautosomal hot spots

Only 34.4% of the variation in CO frequencies observed at hot spot *SPRY3* could be explained by 0442 and *PRDM9* status (Sarbajna *et al*., submitted). Indeed, ~114-fold variation in CO frequencies was observed amongst 0442C/T, *PRDM9* A/A men (Chapter 6). It thus seemed likely that additional *cis* and/or *trans*-acting factors influenced CO frequencies at this hot spot. Subsequent to the work described in Chapters 5 and 6, further analysis was carried out in order to identify these factors (Sarbajna *et al*., submitted), as described below.

9.4.1. Additional *cis*-acting influences on CO frequencies

In addition to analysing SNP haplotypes for the 28 men assayed at *SPRY3*, variation within a purine-rich tandem repeat region, centred ~300 bp away from the hot spot centre, was also investigated as a potential *cis*-acting regulator of CO activity (R. Neumann, C.A. May). Poly-purine tracts of >15 bp have previously been associated with genome-wide sites of DSB-formation in *S. cerevisiae* (Bagshaw, Pitt & Gemmell, 2006). Moreover, purine-rich sequences have the potential to form non-B DNA, which have been known to show increased propensity for DSB-formation (Chuzhanova *et al*., 2009). Hence, the repeat region located distal to the centre of *SPRY3* was first investigated for potential correlations between purine-content and CO activity. Re-sequencing this region in all 28 men revealed repeat regions ranging in length from 250- 298 bp. However, these did not vary substantially in their purine-content, which only ranged between 94.2- 95.5% and did not appear to be associated with CO activity at the hot spot.

Second, the more variable distal end of this repeat region showed higher order repeat structures, with each repeat unit being ~18 bp-long, but ranging in length between 12 and 21 bp. Based on this, 26 different repeat alleles were identified. However, no significant associations were observed between any repeat allele and CO activities at the *SPRY3* hot spot.

Finally, in the context of potential *cis*-acting influences on recombination, it is also important to consider potential epigenetic effects. The C allele at *SPRY3* SNP 0442 creates a CpG doublet, it is thus possible that epigenetic effects were mediated through differential DNA methylation at this site. Interestingly however, limited bisulphite sequencing over an ~200 bp-long interval encompassing 0442 revealed a complete absence of DNA methylation in this region, arguing against the existence of such effects (R. Neumann; *pers. comm.*).

9.4.2. Other potential *trans*-regulators of CO-activities

In addition to the aforementioned potential *cis*-acting factors, three genetic loci previously associated with variation in genome-wide levels of recombination in human males (Kong *et al.*, 2008; Chowdhury *et al.*, 2009) were also investigated as possible *trans*-acting regulators of COs at the *SPRY3* hot spot (Sarbajna *et al.*, submitted). Two key studies by Kong *et al.* (2008) and Chowdhury *et al.* (2009) that led to the identification of these factors were previously discussed in Chapter 1, section 1.5.2.4 and are briefly reiterated here.

The only locus that showed significant genome-wide association with male recombination levels in both studies was *RNF212* (Kong *et al*., 2008; Chowdhury *et al*., 2009). The locus was first reported by Kong *et al.* (2008), who tested 309,241 SNPs for association with recombination rates in 1887 Icelandic males and 1702 females. They found three SNPs, rs3796619, rs1670533 and rs2045065, all mapping within the gene *RNF212* on chr4p, to be significantly associated with recombination levels in males. Two of these SNPs also showed significant association in females, however, opposite SNP alleles were associated with elevated recombination levels in the two sexes. Moreover, SNP rs2045065 was considered a surrogate for rs1670533, hence, the study mainly focussed on the effects of SNPs rs1670533 and rs3796619. Kong *et al.* (2008) identified only three out of four possible rs3796619-rs1670533 haplotypes, indicating strong LD between the markers. Further, SNP haplotype C,T was specifically associated with high male recombination levels, whilst haplotype T,C was associated with high female rates. It should be noted that significant association between *RNF212* SNP haplotypes and genome-wide recombination levels was also reported in AGRE/FHS datasets (Chowdhury *et al*., 2009).

In order to test for an effect of the two *RNF212* haplotypes on CO frequencies, rs1670533 and rs3796619 genotypes were first determined for all men in the two semen donor panels (E. Lamb, C.A. May). In order to maximise the chances of detecting a significant effect, CO data was examined as follows for all *PRDM9* A/A men assayed at any of the 10 A-regulated hot spots studied in Berg *et al.* (2010) (C.A. May). First, CO frequencies for A/A men assayed at a particular hot spot were normalised to median frequencies for that hot spot. Then, normalised CO frequencies for A/A man were averaged across all hot spots for which they were assayed. Following these adjustments, A/A men were ranked according to overall CO frequencies. Of all PRDM9 A/A men assayed at one/more hot spots, 22 were homozygous for each of the two *RNF212* haplotypes. Further, only one of these men carried the T,C haplotype associated with high female-/ low male-recombination levels. Interestingly, this man showed the lowest CO frequency amongst the 22 men tested (*P*=0.135). Although this association was not significant, it does not necessarily discount *RNF212* as a possible *trans*-regulator at the hot spot. Indeed, it is possible or even likely, that effects more subtle than PRDM9 will go undetected in studies with modest sample sizes, such as the one described presently.

In addition to SNPs in *RNF212*, Chowdhury *et al.* (2009) also reported significant association between SNP rs7863596 and male recombination rates in AGRE/FHS datasets. This SNP lay ~30 kb upstream of gene *UGCG* on chr9q. In order to test for an effect of this SNP on CO frequencies, the same analysis was repeated, as described above for *RNF212* (C.A. May). Of all *PRDM9* A/A men assayed at any of the 10 A-regulated hot spots, 34 were homozygous at rs7863596. Moreover, only one of these men was homozygous for the G allele, previously associated with high recombination rates. CO frequency for this man only ranked 9th amongst the 34 analysed. Although this analysis did not show a significant effect of rs7863596 status on CO frequencies, this could once again very possibly have been on account of insufficient statistical power to detect any subtle effects of the potential *trans*-acting factor on recombination activity.

Finally, Chowdhury *et al.* (2009) reported a significant association between SNP rs11764733 and male recombination rates, with the G allele of the SNP specifically associated with higher genome-wide recombination levels. This SNP lay ~ 1 kb downstream of gene *NUB1* on chr7q. Of the 18 rs11764733 homozygous, *PRDM9* A/A men assayed at one/more A-regulated hot spot, 11 carried two G alleles. However, no significant associations were detected between rs11764733 status and CO activity (P>0.05) (C.A. May).

In summary, the aforementioned analysis did not reveal any significant effects of potential *trans*- acting factors reported by Kong *et al*. (2008) and Chowdhury *et al*. (2009) that could account for the extensive variation seen amongst 0442 C/T, *PRDM9* A/A men at the *SPRY3* hot spot. However, given the sample sizes of this study, the possibility of any of these candidates having subtle effects on CO frequencies cannot be discounted either.

It should be mentioned that in addition to the above, potential effects of fertility on CO frequencies at the PAR2 *SPRY3* hot spot were also investigated in some detail (C.A. May).

This was primarily in view of data from cytogenetic studies, which have shown rare instances of persistent pairing at the distal ends of X and Y-chromosomes in infertile men, resulting in the formation of the so-called 'ring bivalent' (Chandley *et al*., 1987; see Chapter 1, section 1.6). This analysis was based on using sperm DNA yields from semen as a proxy for sperm counts and thus fertility and did not reveal any significant association between 'fertility' and variation in *SPRY3* CO frequencies amongst 0442 C/T, *PRDM9* A/A men (C.A. May). DNA yields of all men assayed at *SPRY3* are listed in Appendix 6.7.

Whilst none of the *cis*- or *trans*-acting factors described here accounted for the substantial variation in CO activity amongst 0442 C/T, *PRDM9* A/A men, it is important to bear in mind that new candidate regulators of recombination have and will most likely continue to be identified through various avenues of research. For instance, in a study published towards the end of 2010, Murdoch *et al*. (2010) analysed numbers of MLH1 foci in 192 F2 male mice derived from a cross between CAST/EiJ and C57BL/6J parental strains. They then utilised the observed variations in mean numbers of MLH1 foci to uncover genetic loci associated with these. Using this strategy, they identified significant associations between seven different loci, on chromosomes 2, 3, 4, 14, 15, 17 and X, and genome-wide recombination levels in male mice. It will be interesting to see if any of these loci have an impact of levels of crossing-over in human males and if so, do they account for variations in activity observed at hot spots like *SPRY3*.

9.5. Insights into NCO activity at recombination hot spots

In addition to providing valuable insights into the distribution and frequencies of COs at PAR hot spots and factors regulating the same, this work has also provided unprecedented insights into NCO activity at human sperm recombination hot spots. In fact, the survey of NCOs at the *SPRY3* hot spot constituted the single largest analysis of this type of recombinant at any one hot spot to date.

Data from this *SPRY3* survey provided important information regarding the frequencies and distribution of NCOs at meiotic recombination hot spots and factors influencing these. The survey also shed light on the relative balance between CO and NCO outcomes of recombination at this hot spot and provided important evidence supporting the presence of a second active pathway of NCO-formation at the hot spot. In the next few sections, these findings and how they relate to available data from humans, as well as other model organisms, will be discussed in detail.

9.5.1. Distribution of NCOs at *SPRY3* and other pseudoautosomal hot spots

NCO data from hot spot *SPRY3*, as well as hot spots PAR1-P and PAR2A showed evidence of co-localisation with COs. These observations were consistent with previous data from autosomal hot spots (Jeffreys & May, 2004) and suggested that both types of events resulted from recombination-initiating DSBs occurring in the same region. Furthermore, peak NCO activities at all three hot spots coincided with the centre of the CO hot spot. For instance, at hot spot *SPRY3*, ~80% of NCOs involved central SNPs 0442 and 6882, located only 45 and 119 bp away from the centre of the hot spot (Chapter 7). Increased frequencies of both COs and NCOs near the centre of the hot spot pointed towards the presence of a central initiation zone within which most recombination-initiating DSBs occurred. Indeed, sharp declines in NCO frequencies were observed outside of this zone at all three hot spots. Also, most if not all co-conversion events observed within these hot spots also occurred within this central zone. Thus, all co-conversions at hot spot *SPRY3* involved one or both central SNPs 0442 and 6882 (Chapter 7). It is worth reiterating however that NCOs, particularly at hot spots *SPRY3* and PAR1-P did occur, albeit at low frequencies, across ~4 kb-intervals encompassing the CO hot spot (Chapters 4 and 7). This was contradictory to data from autosomal hot spots, which suggested that NCOs were more localised in nature (Jeffreys & May, 2004). It is worth reiterating however that low frequency events detected in these assays could also have been early PCR misincorporations.

It is noteworthy that these features of NCO distributions at PAR hot spots are remarkably similar to observations at mouse hot spots, particularly hot spot *A3*, which too was extensively studied for these types of recombinants (Cole, Keeney & Jasin, 2010). Thus, co-localisation of COs and NCOs, with most NCOs occurring within a central 'initiation zone' within which most co-conversions were also observed are all features shared by human and mouse hot spots. Further, similarities between the two species also extend to the nature of NCOs observed at these hot spots, as exemplified by similar tract-lengths and frequencies of co-conversion events. Analysis of NCOs at the *A3* hot spot in highly-polymorphic mouse strains, wherein SNPs were present on average <100 bp apart, revealed that ~86% of NCOs affected a single SNP. Further, NCO-tracts were very short, on average extending between just 16 and 117 bp. Data from an extensive survey of NCOs involving data from 23 markers and 14 men at the *SPRY3* hot spot, also revealed short NCO tracts ranging from 55 to 290 bp in length. Moreover, only ~14% of the >500 NCOs analysed at hot spot *SPRY3* were co-conversions, despite the presence of a good density of polymorphisms near the hot spot centre. Limited datasets for hot spots PAR1-P and PAR2A meant that accurate estimates of NCO tract lengths could not be made at these hot spots. Nonetheless, NCO data from a single man assayed at PAR2A for whom a large number of events were scored suggested that >50% of NCOs at two central SNPs located merely 157 bp-apart affected one or the other site only, consistent with NCO-tracts at the hot spot being short.

9.5.2. Relative frequencies of COs and NCOs at pseudoautosomal hot spots

Cytogenetic data from analysis of human spermatocytes is consistent with the presence of three times as many RAD51 foci, which represented sites of recombination-initiating DSBs, as numbers of CO or chiasmata in male meiosis (Baudat & de Massy, 2007). These data thus predicted a NCO:CO ratio of ~3:1 at sperm recombination hot spots. Previous data from autosomal hot spots, based on the analysis of COs and NCOs in one or two men per hot spot however showed, in all but one case, an excess of COs rather than NCOs. The most extreme example of this is a >12-fold excess of COs at the β-globin hot spot (Holloway, Lawson & Jeffreys, 2006). Even data from a limited survey of NCOs at hot spot PAR1-P were consistent with CO:NCO ratios of 6:1 (Chapter 4). Similarly, comparisons of CO:NCO ratios at PAR2A in a single man for whom >50 NCOs were scored, showed twice as many COs as NCOs (Chapter 8). If this paucity of NCOs at sperm recombination hot spots was genuine, it would indicate that a large number of DSBs were being repaired using sister chromatids, rather than homologous chromosomes, as templates.

The large survey of NCOs at the *SPRY3* hot spot, however, showed that average frequencies of NCO:CO at the *SPRY3* hot spot was ~2:1, which was fairly similar to the ratios predicted from cytogenetic data (Chapter 7). Since NCOs leave very localised signatures, it seems likely that lower frequencies of these events recovered from other more limited surveys at autosomal hot spots were the consequence of low marker densities, coupled with smaller sample sizes. However, in light of the identification of large-scale inter-individual variations in CO:NCO ratios (Chapter 7), it is also possible that men analysed at autosomal hot spots showed a genuine preference of COs over NCOs. Indeed, the analysis of NCOs at the *SPRY3* hot spot revealed CO:NCO ratios ranging from 4:1 to 1:10 (Chapter 7). Such variations had previously only been observed between hot spots and were thought to reflect hot spot-specific biases in the resolution of recombination intermediates (Holloway, Lawson & Jeffreys, 2006). Large-scale variations in CO:NCO ratios at a *single* hot spot however suggest that factors influencing the CO:NCO decision at human hot spots are likely to be more complex than previously anticipated.

9.5.3. Regulation of NCOs at pseudoautosomal hot spots

Data from mouse hot spots *Psmb9*, *Hlx1* and *Esrrg-1* suggested that loci *Dsbc1* and *Rcr1*, which were subsequently shown to represent the effects of *trans*-acting factor Prdm9, influenced both CO and NCO activities at these hot spots (Grey *et al.*, 2009; Parvanov *et al.*, 2009, 2010). Data from *SPRY3* and PAR2A showed that this was indeed the case for human hot spots, with *PRDM9* status profoundly influencing NCO frequencies at both hot spots. However, different *PRDM9* alleles activated NCOs at the two hot spots. Thus, NCOs at the *SPRY3* hot spot were triggered by the A-variant, such that NCO frequencies amongst *PRDM9* A/A and A/N men were significantly higher than those amongst *PRDM9* N/N men ($P<0.05$; Mann-Whitney U test). Further, both COs and NCOs were suppressed in N/N men, suggesting that PRDM9 influenced frequencies of recombination-initiation, rather than the CO:NCO decision (Chapter 7). At hot spot PAR2A, NCOs (like COs) were activated by specific *PRDM9* Ct alleles, C and L14. Carriers of these alleles showed significantly higher NCO frequencies compared to men lacking such alleles, most of whom were A/A homozygotes (Chapter 8). As at *SPRY3*, data from hot spot PAR2A was also consistent with PRDM9 acting upstream of the CO:NCO decision (Chapter 8).

In addition to *trans*-regulation of NCOs by PRDM9, data from the two PAR2 hot spots also suggested the presence of *cis*-acting influences on NCO activity. At hot spot PAR2A, significant bias in NCOs was only observed in a single man at central SNP rs28725321. This man also showed biased GC accompanying COs, with over-transmission of the same SNP allele, that is allele A. Further, extents of biased GC observed in the two types of recombinants were similar. Indistinguishable GC-bias in COs and NCOs had also been observed in a single man assayed at the autosomal *NID1* hot spot (Jeffreys & Neumann, 2005). Similar GC-biases in both types of recombinants were consistent with them resulting from the alternative resolution of a common intermediate, with the bias reflecting haplotype-specific biases in initiation of recombination. However, a more detailed analysis of NCOs at hot spot *SPRY3* revealed an entirely different picture.

Recombination frequencies at hot spot *SPRY3* were regulated in *cis* by central hot spot SNP 0442, such that recombination events initiated more frequently on the haplotype carrying the C allele of the SNP, compared to the haplotype carrying the T allele (Chapters 6 & 7). Data from ten men assayed at the *SPRY3* hot spot showed that while all men showed a more or less constant 0442 C:T bias of 2:1 in COs, these values ranged from 1:1 to >85:1 in NCOs. In fact, for six of these men, C:T bias in COs was significantly different to the bias observed in NCOs (Chapter 7). This observation was not only unprecedented at a human sperm recombination hot spot, but it also pointed for the first time to the presence of a distinct pathway of NCO-formation in humans, as discussed below.

9.5.4. Multiple pathways of NCO-formation at the *SPRY3* hot spot

Significantly different biases in COs and NCOs were consistent with the presence of factors that specifically increased the GC-bias in NCOs relative to that introduced at initiation by *cis*-effects of SNP 0442. Interestingly, these men also showed significantly higher numbers of NCOs than COs, a correlation that was statistically significant ($P<0.01$; Chapter 7). This increased propensity of NCOs, but not COs, towards elevated GC-biases in men showing high NCO:CO ratios, also pointed towards alternative processing of at least a subset of NCOs in these men.

Several lines of evidence from yeast and mice suggest that COs and NCOs in these model organisms are most likely generated via distinct pathways that diverge soon after the initial RAD51/DMC1-mediated strand-exchange reaction (see Chapter 1, section 1.2.4). In yeast *S. cerevisiae* evidence in favour of distinct pathways of CO and NCO-formation comes from ZMM mutants, which show dramatic reductions in frequencies of COs, but not NCOs (Lynn, Soucek & Börner, 1997; Shinohara *et al*., 2008). Further, data from Allers and Lichten (2001), based on the physical detection of recombinants and the analysis of *ndt80* mutants, were consistent with COs resulting from a JM (joint molecule) intermediate and NCOs resulting from a distinct pathway that did not involve long-lived dHJ intermediates. Also, mismatch proteins Mlh1 and Mlh3 were specifically required for the formation of COs but not NCOs in the budding yeast (Hunter & Borts, 1997; Wang, Kleckner & Hunter, 1999). Similar observations were also made from analysis of *Mlh1*$^{-/-}$ mice, which showed that the protein was specifically required for CO-formation, as well as for the repair of hDNA generated in this processes (Guillon *et al*., 2005). Further evidence for different pathways of CO and NCO-formation in mice came from analysis of GC-tracts associated with the two types of recombinants. Thus, data from *Psmb9* and *A3* hot spots, showed that GC-tracts associated with NCOs were typically very short (~100 bp), while those associated with COs were much longer, extending over ~500 bp (Guillon *et al*., 2005; Cole, Keeney & Jasin, 2010).

Analysis of NCOs at the *SPRY3* hot spot also showed evidence of significantly shorter NCO-tracts in a man showing greater biases in NCOs than COs as well as high NCO:CO ratios, compared to a man showing similar biases both types of recombinants and larger numbers of COs than NCOs. Longer NCO-tracts identified in the latter were very similar in length to GC-tracts associated with COs (313 bp and 329 bp respectively), consistent with both types of events resulting from the same pathway. Data from the other semen donor however, are more consistent with at least a sub-set of NCOs resulting from a distinct pathway of NCO-formation, defined by high-GC biases and shorter NCO tracts (Chapter 8; Sarbajna *et al*., submitted). It will be interesting to see if similarly large surveys of NCOs at other PAR and autosomal hot spots also show presence of a separate pathway of NCO-formation or if this is a feature restricted to the *SPRY3* hot spot.

9.6. Future considerations

This work had led to the identification of three new PAR1 hot spots, as well as the first two PAR2 hot spots. While data from these and other hot spots have greatly increased our understanding of the fine-scale distribution of recombination events and factors influencing these, a large number of important questions still remain. It is, for instance largely unclear if and how PRDM9 interacts with hot spot sequences to trigger recombination activity and what roles, if any, predicted DNA sequence-binding motifs play in this process. It will also be a challenge to identify statistically significant effects of additional *trans*-acting regulators on CO activities at hot spots, although a number of potential candidates have been and continue to be identified, including those identified by Murdoch and colleagues (2010).

Work at the *SPRY3* hot spot has also revealed remarkable variation in relative frequencies of COs and NCOs amongst various men assayed at the hot spot. It is however as yet unknown what factors cause this inter-individual variation in the CO/NCO decision. Finally, it remains to be seen if an additional CO-independent pathway of NCO-formation does indeed exist at other hot spots and if so, what determines the relative frequencies of DSBs that are processed by one pathway or the other. It will be interesting to see how future research unravels the answers to these questions, which will no doubt bring us a few more steps closer towards a more holistic understanding of meiotic recombination in the human genome.

Chapter 2 Appendices

Appendix 2.1

- **Suppliers of chemical used**

2M trisma base, 2-mercaptoethanol, ammonium sulphate, dithioreitol, glycerol, sodium chloride, sodium lauryl sulphate, spermidine trichloride, tetramethyl-ammonium chloride, water were obtained from Sigma-Aldrich (Poole, U.K.). Bovine serum albumin (BSA) was obtained from Ambion (Huntington, U.K.). Deoxy-nucleotide triphosphate(s) (dNTPs) were obtained from Promega (Hampshire, U.K.). Ethidium bromide was obtained from BioRad (Hemel Hempstead, U.K.). Ethylenediamine tetra-acetic acid (EDTA), magnesium chloride, polyvinylpyrrolidone (PVP) and sodium hydroxide were obtained from Fisher Scientific (Loughborough, U.K.). Ficoll 400 was obtained from GE healthcare (Little Chalfont, U.K.).

- **Suppliers of plastic-ware**

Most plastic-ware (including PCR tubes, 96-well plates, self-adhesive seals) were obtained from ABgene (Surrey, U.K.). PCR master mixes were set up in Protein LoBind Tubes obtained from Eppendorf Scientific (Hamburg, Germany). Plastic pipette tips were obtained from Starlab Ltd. (Milton Keynes, U.K.) and Sarstedt (Germany).

Chapter 3 Appendices

Appendix 3.1

Table A. Universal primers used in the study

- **Primers used in donor re-sequencing project**

Primer Name	Forward (F) / Reverse (R)	Primer sequence (5′ to 3′)
3D	F	CTTTGGGCCGCTCCAAAGACG
CHMR6	R	CTACCCCTCACCGCCTGCCGA
35B3	F	ACCACCCTGAGTTTCTCTG
35A	R	GCCCGGCCAAAGCTGAATAG
35AR	F	GGTTAAACTATTCAGCTTTGGC
35D2	R	AACGACAACTCCAAGAAC

- **Primers used in ASP optimisations** (Note: ASP sequences are listed in Table B)

Primer Name	Forward (F) / Reverse (R)	Primer sequence (5′ to 3′)
SEQ35B	F	TAACCGAAATGAAGCCGAG
SEQ35A	R	GCCCGGCCAAAGCTGAATAG

Table B. Primer combinations and PCR conditions used in phasing reactions

Note: The SNP against which ASPs are designed is shown in red. Also, same selector sites were used for both men.

- **Primary forward ASP and secondary reverse ASP combinations**

Primers	**PCR conditions**
Forward primers: AS341A2 – 5' GCTGTGTTCCGATGTAAC 3′ OR, AS341T – 5' ACGTGGGCTGTGTTCCGT 3′ Reverse primers: RAS4673A- 5' cACTGTGGTTTTGGCCAT 3′ OR, RAS4673G- 5' cACTGTGGTTTTGGCCAC 3′	96°C for 1min 96°C for 20 sec 63°C for 30 sec } x 31 cycles 65°C for 5 min 65°C for 5 min

- **Secondary forward ASP and primary reverse ASP combinations**

Primers	PCR conditions
Forward primers: AS663C – 5' GGATGTATCGGATGTTGC 3′ OR, AS663G – 5' GGATGTATCGGATGTTGG 3′ Reverse primer: RAS9931C2- 5' ATTTCAGCACAGGATG 3′	96°C for 1min 96°C for 20 sec 54°C for 30 sec 65°C for 5 min } x 31 65°C for 5 min
Forward primers: AS663C – 5' GGATGTATCGGATGTTGC 3′ OR, AS663G – 5' GGATGTATCGGATGTTGG 3′ Reverse primer: RAS9931T- 5' ccccTTTCAGCACAGGATA 3′	96°C for 1min 96°C for 20 sec 50°C for 30 sec 65°C for 5 min } x 31 65°C for 5 min

- **Determining internal phase**

Primers	PCR conditions
Forward primers: AS663C – 5' GGATGTATCGGATGTTGC 3′ OR, AS663G – 5' GGATGTATCGGATGTTGG 3′ Reverse primers: RAS4673A- 5' cACTGTGGTTTTGGCCAT 3′	96°C for 1min 96°C for 20 sec 54°C for 30 sec 65°C for 5 min } x 14 65°C for 4 min

Table C. Primer combinations and PCR conditions used in man 1 and man 2 crossover assays.

Note: Same selector sites were used for both men; they also had the same phasing at selector sites.

- **Primary PCRs:**

Primers	PCR conditions
Orientation A- AS341A2 – 5' GCTGTGTTCCGATGTAAC 3′ RAS9931C2– 5' ATATTTCAGCACAGGATG 3′	96°C for 1min 96°C for 20 sec 56°C for 30 sec x 27 cycles 65°C for 6:30 min
Orientation B- AS341T – 5' ACGTGGGCTGTGTTCCGT 3′ RAS9931T– 5' ccccTTTCAGCACAGGATA 3′	96°C for 1min 96°C for 20 sec 50°C for 30 sec x 27 cycles 65°C for 6:30 min

- **Secondary PCRs:**

Primers	PCR conditions
Orientation A- AS663C – 5' GGATGTATCGGATGTTGC 3′ RAS4673G– 5' cACTGTGGTTTTGGCCAC 3′	96°C for 1min 96°C for 20 sec 61°C for 30 sec x 2 cycles 65°C for 5:30 min 96°C for 20 sec 60°C for 30 sec x 22 cycles 65°C for 5:30 min
Orientation B- AS663G – 5' GGATGTATCGGATGTTGG 3′ RAS4673A– 5' cACTGTGGTTTTGGCCAT 3′	96°C for 1min 96°C for 20 sec 60°C for 30 sec x 2 cycles 65°C for 5:30 min 96°C for 20 sec 59°C for 30 sec x 22 cycles 65°C for 5:30 min

- **Tertiary PCRs:**

Primers	PCR conditions
Orientation A and B 3D – 5' CTTTGGGCCGCTCCAAAGACG 3′ 35D2– 5' AACGACAACTCCAAGAAC 3′	96°C for 1min 96°C for 20 sec 54°C for 30 sec x 36 cycles 65°C for 5 min

Appendix 3.2.

Sequences of ASOs used to determine internal SNP status in d10 and d62 CO assays.

The SNP itself is indicated in red. Many of these SNPs were identified through re-sequencing efforts in the lab (C.A. May, unpublished data) and were therefore named according to a 'local' system. However, some of these SNPs were subsequently added to dbSNP and 'rs' numbers are shown where available.

SNP name	'rs' number	Allele	ASO sequences
rs3995663	rs3995663	- AGGG	CATCCCGCCAAAGTCCAG CATCCCGAGGGCCAAAGT
3/1035	rs2213234	A G	CAGCCAGACCCCCGAAAT CAGCCAGGCCCCCGAAAT
3/1076	-	G A	GTTTCTGTTCTGGTCTTG GTTTCTATTCTGGTCTTG
3/1248	rs34668574	A G	TGTCCTCACCTCTGGGTT TGTCCTCGCCTCTGGGTT
35/1400	-	C T	GACATAGCGAGGATGGCA GACATAGTGAGGATGGCA
35/1652	rs28408272	C G	GTATGTGCATTTGTTATT GTATGTGGATTTGTTATT
35/2287	rs2187644	A G	GGTACCCAGTGTTGCCCC GGTACCCGGTGTTGCCCC
35/2637	rs2156998	A G	CACCACCACGCCCGGCTC CACCACCGCGCCCGGCTC
Indel103	-	T G	CCAGCCTTGACCAAGCTG CCAGCCTGTTTTTTAGGG
rs34995914	rs34995914	G A	GGAAGCTGAGGCAGGAGA GGAAGCTAAGGCAGGAGA

Appendix 3.3

Data summary from reciprocal crossover assays on d10 and d62

Table shows calculation of crossover frequency per unit length of DNA for individual marker intervals for each donor. Note: Number of molecules (mols.) screened for each interval = Input per pool x number of pools scorable for COs of a particular type. Correction factor = total number of mols. screened in the assay/number of mols. screened per interval. Corrected number of crossovers (COs) = Poisson-corrected COs x correction factor. Recombination activity per interval (cM/Mb) = (corrected number of COs /length of interval, bp) x 10^8.

d10, Orientation A:

Marker interval	rs3995663 - 1035	1035-1076	1076-2287	2287-2637	2637-103	103-rs34995914
Uncorrected COs	0	1	57	12	9	1
Poisson-adjusted COs	0.00	1.00	71.79	12.50	9.27	1.00
Total mols. screened	31890.65	31890.65	41319.19	41596.5	43260.36	43260.36
Correction factor	1.36	1.36	1.05	1.04	1.00	1.00
Corrected no. of COs	0	1.36	75.16	13.00	9.26	1.00
Length of interval (bp)	15	42	1336	350	806	256
cM/Mb	0	75	130	86	27	9

d10, Orientation B:

Marker interval	rs3995663 - 1035	1035-1076	1076-2287	2287-2637	2637-103	103-rs34995914
Uncorrected COs	0	1	72	15	27	1
Poisson-adjusted COs	0.00	1.00	98.68	15.80	29.70	1.00
Total mols. screened	37714.16	37714.16	41419.19	40209.95	41873.81	41873.81
Correction factor	1.15	1.15	1.05	1.08	1.03	1.03
Corrected no. of COs	0.00	1.15	103.32	17.00	30.68	1.03
Length of interval (bp)	15	42	1336	350	806	256
cM/Mb	0	63	179	112	88	9

d62, Orientation A:

Marker interval	rs3995663 - 1035	1035-1076	1076-2287	2287-2637	2637-103	103-rs34995914
Uncorrected COs	0	4	1	5	52	0
Poisson-adjusted COs	0.00	4.10	1.00	5.15	68.22	0.00
Total mols. screened	131307	133307	133307	138640	173302	173302
Correction factor	1.33	1.31	1.31	1.26	1.01	1.01
Corrected no. of COs	0.00	5.36	1.31	6.49	68.75	0.00
Length of interval (bp)	15	212	101	429	1792	256
cM/Mb	0	11	6	6	16	0

d62, Orientation B:

Marker interval	rs3995663 - 1035	1035-1076	1076-2287	2287-2637	2637-103	103-rs34995914
Uncorrected COs	0	4	6	1	64	1
Poisson-adjusted COs	0.00	4.08	6.21	1.00	95.01	1.00
Total mols. screened	141305	144638	143305	139972	169302	165969
Correction factor	1.25	1.22	1.23	1.26	1.04	1.06
Corrected no. of COs	0.00	4.97	7.62	1.26	98.75	1.06
Length of interval (bp)	15	212	101	429	1792	256
cM/Mb	0	10	32	1	23	2

Chapter 4 Appendices

Appendix 4.1

Annotated sequences, ChrX: 1686960-170399 (Target ‘A’)

Centre of LD step as estimated from LDU map: 1695000bp.

~9 kb of sequence on either side, i.e. ~18 kb of sequence flanking this centre was chosen for study.

Note:

- Genomic locations, according to NCBI36/hg18 (in bp), annotated at the end of each line indicate the position of the first base of that line. Local numbering, on the other hand, indicates the position of the last base of that line.
- SNPs are highlighted in blue; variant bases are indicated above each SNP. ‘rs’ numbers are written alongside respective SNPs. SNPs identified through donor re-sequencing are indicated in **purple** and named according to their positions, followed by ‘seq’.
- All repeats except Alus are indicated by lower case letters coloured red. These include LINEs, LTRs, simple repeats, low complexity repeats and MER elements. Alus are indicated by lower case letters coloured green. All repeat types are indicated near the start positions of respective repeats.
- Primer sequences are underlined, primer names are written above the sequence and primer direction (forward/ reverse) are indicated using arrows. If two primer sequences overlap, one of them is highlighted in yellow.

```
                                 rs6644628 g                   NCBI36   Local
cctgggcgacagagcgagactccatctcaaaaaaagtaaaacaataaagaGTATAGAAAA   1686960   60
    A1687.0F ->
ACATGCATGCATGGGGGTATAGGAGAGGTCCCAGCCCTCAGCCGTAGCCATGGATGTCC   1687020   120
```

```
A1687.08F ->
AGGTGGTGATGTGTTGCCCTGGCCCGCCCCGCAGGCCTCCAGCATCTCCAACAGCGGCTT   1687080   180
                      rs28698568 a                     FAS2605A/G ->
CCCAGGGGATGCTGAAGCTGCGGCCGGACCGGCCAGCTCGGGCAGGGGTGAGCGGAGCTG   1687140   240
 rs28612605 g
CATCACCCAGGCATCCCGTCTGGGACCTCCTCCCATGCCTCCTTCCTGGGCACCTGCCCC   1687200   300
                                                 FAS0198 ->
TCCTCACTAGCTCATCCTCCTTCCCACCTGGTGGAGGACACGCACACCCTCCTGGTGGCG   1687260   360

 rs28640198    rs28412948 rs28681628    rs28582979
     g             c        t                t
TCCTCAAGCCCCGTCCACCTGCCCAGGAGCTGGGGCATCCTGTGTCTCCCGCCTTGCTCT   1687320   420
CCTGCACAGCCTGACTCAGAGCGGGAGCTCGACCCCCTACCCTCCCCGCTGTACGCTGCC   1687380   480
              rs28677958 t    g rs28516063
CCAGCAAACCTCCCGAGTTTGCACCCGGCCACTGGATCGGCCTGCACAGGGGCTGCCGCC   1687440   540
TCCGTCCCAGATGGCAGGGGCTTTTCCCTGCACCCCACCTCACTCAGCTCATGCCTCTGC   1687500   600
TCCTGGTGAAAAACTTTTCCCATCACTCCGTCCCGTCCCCTCTCTGCAGCAAATGACTCT   1687560   660
                               rs5949017 g
TGTCCAGGCTGTCCAGGCTGCCAGAGACCACCGAGGCACTCAGTCCGGAAGCCCATCTGT   1687620   720
GCCAGGGTCATTGGGAGGTAAAGGCAACAGGCTTTACCTCTCCAGGCCACCCCTGCGCCT   1687680   780
         rs28753733 t        c rs28651728
ACTCCAGCTAAGACTCCAGGCCACCCCTGACCCTACTCCAGCCAAGCCTCCAGGTCACCC   1687740   840
   t rs28645134                                  AluY (SINE)->
CTGCCCCTACTCCAGCCAAGCCTCCAGGTCACCCCTGTCCCTACTCCAGGggctgagcgc   1687800   900
ggtggctcacgcctgtaatcccagcactttgggactccgagttgggcggatcacgaagtc   1687860   1020
aggagatcgagaccatcctggctaacacggtgaaatgccgtctctactaaaaatacaaaa   1687920   1020
acaaaattagccgagtgtggtggcaggtgcctgtagtcccagctactggggaggctgagg   1687980   1080
rs28609866 a                    rs28476694 g
caggagaatggcgtgaacccgggaggcggagcttgcagtgagccaagatcacgccactgc   1688040   1140
          rs28447248 t             -/a rs35799222
actccagcctgggtgacagagcgagactccgtctcaaaaaaaaaaaaaaaaaGCAACAGT   1688100   1200
ATGTTCCCCTGGGAGAACGGAGGGGGCGGCTGTTTACAGTTGCCCCAGGGCAGGACCTGC   1688160   1260
ACACTGGACCATCCAGGGCCAGTCACCCCACAGGCATCTCAGACTCCACACATCCCCGCC   1688220   1320
                                                         g rs5989848
AAACCCCAGCCCTGGGTCTTCTGCATTCTGGGGACATGGGGACACCACGCTGGTCGGTAG   1688280   1380
                                             rs5989849 c
GGTCCCAGGCTCTTCTCAATCTCTCTCGGGCCCCTCCTCTCCTGCTGCCGCTGCCGCGGT   1688340   1440
                   A1688.4F -> A1688.40F ->
GAGACTCCGAGGCCTGTAGGAAGCTTCAGGTGCCTGAGCCTCCAGTACAGCCCTTCCCCT   1688400   1500
      g rs5989850
ACTCCACCCAAGCCTCCAGGCCACCCCTGTTCCTACTGCAGCCAAATCTCCAGGCCGCCC   1688460   1560
            A1688.5F ->
CTGTCCCCTACTCAAACCAAGCCTCCAGACCGCCCCTGTCCTCTACTCCAGCCAAGCCTC   1688520   1620
```

```
                                                 C-rich-
                             rs28762782 c    c rs28854338
CAGGCCGCCCCTGTCCTCTACTCCAGCCAAGCCTCCAGGCTGCCCGTGTcccctactcca   1688580   1680
 rs10523083
Large deletion/-                       g rs28781588
Gccaaagctccaggccgcccctgcccctactccagccaaagctccaggccacccctgtcc   1688640   1740
cctactccacccaagcctcCAGGAGGCCCTGTCACCTACTCCAGCCAAGCCTCCAGGCCA   1688700   1800
GTCCTTCCTTTACTCCAGCCAAGCCTCCAGGCCACCCCTGTCCCTACTGCAGCCAAACCT   1688760   1860
                                          C-rich->
CCAGGCCGCCGCTGTCCCCTACTCAAACCAAGCCTCCAGGCCgcccctgctcctcctcca   1688820   1920
      rs28414351 g                     rs28368906 t
gccaagcctccaggccacccctgtcccctgctccagccaaagctccaggccgccccgccc   1688880   1980
                      rs28409457 t   c rs28548651      a 28464375
ctgctccagccaaccctccaggccgcccctgccccctgttccagccAAGCGTCCAGGTCGC   1688940   2040
                                           LTR2C ->
CCCCGTTCCTACTCCAGCCAAACCTCCAGGCCACTCCTGCCCCtgtagcaggactagccg   1689000   2100
t rs28481943               a rs28724312
cggataaaacccctcagacaccaggtttaggaaggttttggctttattcggccggaagcg   1689060   2160
ttggcggactcacgtctcgagaaccgggctctccgaagacagagctcctggccctttgaa   1689120   2220
aggcttacaactctaaggggttccacgtgaaagggttctgatagattgagagcacgtgtg   1689180   2280
gTTAGAGTCGGGGGTTAATGTTTTAACCTCAGGCCTGGTCAGGGGTGCCGGCTGGTCTTG   1689240   2340
CCACTGACTTCATTCCTGTTGTTTTTCAACTTTTACTTCCTCCTTCTCTTCAGAGACAGG   1689300   2400
AGATAGTAAAATAAATGGCCTCTCTCCCCACCCCGACTCCAGCCAAGCCTCCAGGCCGCC   1689360   2460
                    rs28391749 a
TCTGACCCTGCTCCAGCCAAACCTCCAGGCCTCCCCTGCCCCTACTCCAGCTAAGACTCC   1689420   2520
   L2 (LINE)-> rs5949018 t
AGGccacccctgaccctactccagccaaaacttcagcccgtcccactcacccaagttata   1689480   2580
Caaaagctcctgggctttccggccaactctgtgtcccaaacatcccatccccaaaatggc   1689540   2640
                                            rs28517389 c
agcttctgactgtgtccagaccatgcaggcctctcctgaaaccgtcctgccagacgctgt   1689600   2700
   rs28565489 a  g rs28372415
ctcaaggtggtgagggtttccgtgcaatgtccctgagggcttttaaagccaccctgtctc   1689660   2760
ctgtccatcggcccccctcccctcctctccctctgtctttccagatgctccacccatgcag   1689720   2820
               rs28696100 t
ggttgttgtctcttctttgaacagcccggacgcactgccgctaccccacggcctccgcac   1689780   2880
          -/t rs34639351             rs5989851 a
ctgctgttccctctggcagagacggtctttcctgggagcacccacccggtccctcctgcc   1689840   2940
                                     rs7877401 t
tcggatttcaaagacttccaaaaccaccttcaccccaaaatccagctctccctcccctcc   1689900   3000
             A1689.9F ->                 AluSx (SINE)->
ccttctgtccacacccccAAATCACCTTCTAACATCTTCAtctttttttttctttgtttt   1689960   3060
gagatggagtctccttctgtcgcccaggctggagtgcagcggcgcgatctcggctcactg   1690020   3120
caaccgccacctcccgggttcaagcgattctcctgcctcagcctcccaagttgctaggat   1690080   3180
```

```
t rs5989836                       rs5989837 c    a rs28638243
aacaggcatgcactgccacgcccagctaatttttgtatttgtagaagagatggggtttc    1690140   3240
rs28464926 c                        rs28649445 c
accatgttggtaggctggtctcaaactgctgacctcaaatgatccagctgtcttcgcctc    1690200   3300
rs28560383 g                               THE1B (LTR)->
ccaaactgcccagatgacacatgtgagccaccacacctggccCagatctgatggttttat    1690260   3360
      rs28550202 c               A1690.3F/R ->
aaatgggaggtcccctgttgcacacgccctctggcctgtcaccaggtaagacatgccttt    1690320   3420
gcttctccttggctttcctgccatgagtgtgatatgtatttatcagcaacatgaaaatgg    1690380   3480
                                                         AluY (SINE)->
actaatacaatatgcttgtcagccacgcgtgtgtcttctttttgagaagtgtctgttggc    1690440   3540
cgggcgcggtggctcactcctgtcatcccagcactttgggaggccgaggcgggtggatca    1690500   3600
cgaggtcaagagatcgagaccatcctggctaacacggtgaaaccctgtctctactgaaaa    1690560   3660
     rs5949020 c g rs28866032
tacaaaaaattagctggctgtggtggctggtgcctgtagttccagctactctggaggctg    1690620   3720
                                             rs5949021 c
aggcaggagaatgacacgaacccgggaggcggagcttgcagtgagctgagatcgagccac    1690680   3780
                                                       THE1C (LTR)->
tgcactccagcctggtgacagagcgagactccatctcaaaaaaaaaaaaaaaaaattacc    1690740   3840
                                                     LIP (LINE)->
caaaattacccagtctcaggtatgtcttatcagcagcatgaaaatggactaatatgcttg    1690800   3900
                                      AluY (SINE)->
tcagtcacgtgtgtgtctttttttggagaagtgtctgttggccgggcgcggtggctcacg    1690860   3960
                         rs4933062 t
cctgtcatcccagcactttgggaggccgaggcgggcggatcacgaggtcaagagatcgag    1690920   4020
      rs6644630 t             rs6644631 ga rs6644771
accatcctggctaacacggtgaaaccccgtctctactaaaattacaaaaaattagccagg    1690980   4080
                      rs6644632 c
tgtggtggtgggcgcctgtagtcccagctactggggaggctgaggcaggagaatctcttg    1691040   4140
aacccgggaagcggagcttgcagtgagctgagatcgtgccactgcactccagcctggtga    1691100   4200
                             A1691.1F/R -> THE1C (LTR)->
cagagcgagattccatctcaaaaaaaaaaaaaaTTACCCAaaattacgcagtctcagata    1691160   4260
              rs28477469 c        LIP (LINE)->
tgtctttatcagcagcatgaaaacggactaatacaatatgcttgttggccacgtgtgtgt    1691220   4320
                    AluSx (SINE)->
catctttggaaaagtgcctgttggctgggttcagtggctcatgcctattatcccagcact    1691280   4380
                          rs6644633 a
ttgggaggctgaggcagacggatcacctgaggtcaggagttcgagatcatcctggccaat    1691340   4440
atagtgaaaccctgtctctacgaaaaatacaaaaattagccaggtatggtggcaggtgct    1691400   4500
tgtaatcccagctactcgggaggctgaggcagaagaatcgcttgaacctggaaggtggag    1691460   4560
gttacagtgagccaagattacaccactgcactccagcctggtgacaaagcgagactctgt    1691550   4620
ctcaaaaggaaaaagaagaaaaaagaaaaaGTGTCCGTTCATGTCCTTTGTAAGACTCCA    1691580   4680
CTCTCTAGTAACCGTGTTCCTGCCCTTCTCAATTCCTAAGGTAATCGCCTTAACTCAGAT    1691640   4740
```

```
CCCTGAGAAATCCCCACAGCAGGCAGAGATTTGGTCTCCTGACAGTAGCTTTGGGATTCC   1691700   4800
                                A1691.7F ->
ATTAAGCAAATATGATTGGGGGCCAGCCCCTTCTGTCTCCAGGTAAGGTGTCCCCAAGGA   1691760   4860
                 LIMC4 (LINE)->
A1691.7F ->               -/tctc rs34332882
CCCTTCATCGCTCTGAActtcaaggatttcttttctttcttgctttccttctttctttc   1691820   4920
                       rs6644772 a              rs28477873 t
tttccttctttcttccttcttttctttttcttccttccttctttcttcctttcttttc   1691880   4980
 rs28592889 t           rs28514348 t       c rs28547988
tttctttctttccttctttttcttccttccttcctccctccctttccttctttctttcttt   1691940   5040
                     rs28640167
       rs28522722 c      t  c  rs6644773
ctgtatctctctctctcttttctttcctttctttctttccatttctgtctctctctcttct   1692000   5100
ttcttttctcttctttctttctttctttctttctctctctctctctcctCTCCTCTTT   1692060   5160
rs34673148 g AluSx (SINE)->              rs5948999 a
CTCTCTCTCTGtttctctctttctttctttctttttgatacagacctcactggtgttgccc   1692120   5220
aggctggagtgcaatggcccaatctcggctcactgcaatctctcccttccaggttcaagt   1692180   5280
gattctcctgcctcagcctcccaagtagctgggatgacaggcacctgccaccatgcccgg   1692240   5340
ctaattttgtatttttagtagagacggggtttcaccatgttgaccaggctgatcttgaa   1692300   5400
ctcctgacctcaagtgatcctcgcacctcagcctcccaaagtgctgggattacagacatg   1692360   5460
                AluSq/x (SINE)->
agccaccttacctggctaattttatatttttagtagagtcggggttacaccgtgttagc   1692420   5520
caggctggtctcgatctcctgacctcatgtgatctgccccacattgctgggatgacaggc   1692480   5580
                     A1692.5F ->      AluSp (SINE)->
TGTGTTTATATTTTTCCCCTAGAAAATCACTCTGGAggctgggagtggtggcttacgcct   1692540   5620
  g rs6644775
gtcatcccagcaatttgggaggccgaggtaggcggatcagctgagatcaggagttcaaca   1692600   5680
                        rs5949000 t        rs6644776 t
ccagcctgaccaacatggagaaaccccgtatctactaaaaatacaaaattagccgggcgt   1692660   5740
ggtggtgggtgcctgtaatctcaactactgggagcctgagacagaagaatcacttgaacc   1692720   5800
         rs5989838 t              rs6644777 a

tgggaggcagaggttgcggcgagACATGCATCCTCAGAAATGGGGTAGAGACCTGGCACG   1692780   5860
     AluY (SINE)->
GTCACtttttttttttttgagacagattctcgctctgtcgcccaggctggagtgcagtg   1692840   5920
gcgcgatcttggctcactgcaacctcagcctcccgggttcacacccattctcttgcctca   1692900   6000
rs7878620
g     g rs4372180
gcctcccaagtagctgggactacaggcacccgccaccacgcccagctaatttttttgtatt   1692960   6060
                           rs4442313 t
tttagtagagactgggtttcactgtgttagccaggatggtctcgatctcctgaccttgtg   1693020   6120
atccgcctgccccggcctcccaaagtgctgggatgacaggctaccacgcccagccCTGGC   1693080   6180
                   A1693.1F/R ->
```

```
ACGGTCACATTTTAATTATGTGCAGGCATTTAGATTGTCCAAATAAACGTTGCAAGATAT   1693140   6240
  AluJo (SINE)->
Ctaggctgaagtgggaggatctcttgaagtcaggagttcaagaccagcctgggcaacata   1693200   6300
                                              AluSx (SINE)->
gtgagacacccccccccacccccgatctctacaaagaattaaaaaattagtcaggcacat   1693260   6360
                  rs7472200 t            rs6588801 t
tggctcatgcctgtcttcccagcactttgggaggccgaggtgggtggatcaccttaggtc   1693320   6420
     c rs6644778
aggagttggagaccagcctggccaacatggtgaaaccctgtctctactacaaatacaaaa   1693380   6480
                                            -/t rs34973570
attagccaggtgtggtggtgtgtgcctatagtcccagccactcgagaggctgaggcagga   1693440   6540
gaatcgcttgaacccaggaggcagaggttgcagtgagctgagatggtgccaccgcactcc   1693500   6600
rs34505693 -/a                            A1693.56F ->
agcctgacgacagagcgagactccatctaaaaaaaaaaaaaaGGGGTCTCACTATGTTGC   1693560   6660
         MLT1C (LTR)->
CCAGGCTGGcctgctggcatcttgatgttgaacttccagcctccagaagtgtgagagaac   1693620   6720
tggtttcttgagccacccagtttgcagaaattcgcacgacccccagaaatgaatgcaGGA   1693680   6780
A1693.7F ->                      rs4446909 a
CCTGCTCAATCCATAAGACGATTAATGAGTGCTAATGATGGGTGAGCATATTGGTTTGGG   1693740   6840
       A1693.8F ->
ATCAGTTCTACAGCTGATGAGTCAAGGGCTCGGGGGGGCCTGATGCAGACTATTTTAGGG   1693800   6900
                      rs5989681 c
CTGCGGAGGGGCGCCCAGAAAGCCAGGGAATCGCCACACGCTTTGTGATCCCCTAATACG   1693860   6960
3920seq
A   RAS3920C/T <-
GTTAATGCAAATTTCTAGAGTGATCCGTCTTTGTTTGAGTACTGGGCAGGCAGCAGGGAG   1693920   7020
    rs6644635 c
AGTCAGGCAGCAGCTGTGAGCGGGTGGCTCTTCCCCACCTTGCCAGCAGGCTCTGTGCTC   1693980   7080
    A1694.0R <-
CTTGAAGCAAGCGCTCCAGAGGCTCCGGAAGCCACGGCTGGATTGGAGACAAGATGGGAT   1694040   7140
                                rs17149149 a     <- A1694.16R
CCTCAGAGGACCAGGCCTATCGCCTCCTTAATGACTACGCCAACCGGCTTCATGGTGTCCC   1694100   7200
             <- A1694.1R
AGGTAGGATACGCTCTGTGGGACAAGGGGGAATAGACTTCCGTTCATCACACTCACGTTC   1694160   7260
MER4B (LTR)->
CATTGTTGtgtcagtcgagaaaaatgatgagtctccatcattttaggaggtttattttcc   1694220   7320
   a rs28408621
accgtgaaagacgtacacccacgatgtagcctcaggagttcctgaggacatgtgcccagg   1694280   7380
                        4340seq c   rs5948839 a
gtggtctggggacagcttggtttgatacattttacagagacatgagtcatcaatcaatgt   1694340   7440
              AluSx (SINE)->
        rs35621427 -/c
                     t rs4575022
```

```
atgtaagtacattgggccagtcgtggtggctcacacctgtgatcccagcactttgggagg   1694400   7500
ctgaggcgggcggatcacctgaggtcgggagttcgagaccagcctgacccacacagagaa   1694460   7560
                                            rs35690135 -/c
accccatctctactaaaaatacaaaattagccaggcttggtggtacatgcctataatccc   1694520   7620
agctacttgggaggctgaggcaggagaatcgcttgaacctgggaggcggaggttgcagtg   1694580   7680
               rs7391804 a
agccgagatcgcaccactgcactcccgcctggggacacagcaagactctgtgtcaaaaa   1694640   7740
                               Charlie4 (DNA, MER1_type)->
                             rs28633458 a
aaaaccaaacaaacaacaaaaaaaaAGAAACaaagcttaagatgaaatacataggattag   1694700   7800
aaacaagacaatgtttattgaactgcaagtgttgaaatttttaaaacatttcatgataga   1694760   7860
gtagtatatgtggttgtttattaacccagtaaataaTTACACTGTGCGTTTATCATAGAG   1694820   7920
                  A1694.8F/R            AluSx-> rs5989852 c
GTGAGTTTTGAGGGTTCACCCGTAGAGGAGCAACACCTCAttcctttttttttattttttt   1694880   7980
gagctggagtctcactctgtaacccaggctggagtgcagtggcacggtcttggctcactg   1694940   8040
caacctccgcctccagggttcaggctattctcctgcttcagcctcccgagtatctgggat   1695000   8100
tacaggcacccaccaccacgcccggctaattttgtatttttagtagagatggggtttca   1695060   8160
ccatgttggccaggctggtctcgaactcctgacctcaagtgattcacctgcctcggcctc   1695120   8220
A695.1F/R ->      L1MB5 (LINE)->
ccaaggCCTCATTTCTTCTtctcacagaataacatcgtgctcagtgaaggaaggctgacg   1695180   8280
   rs5989853 g
tgaaaggccccatatggtatgacttcatttatatgaaatgtgcagagcaggcaaatcctt   1695240   8340
cgaggtggaaagtggactggtggtttgcagaggctgggcaggactgcggagggagtgttt   1695300   8400
aacggagacagtgtccctttttggggtgatgcaaatgttctggaacccggcggaagtgct   1695360   8460
ggctatacaacatcacgaatgtactaagtgcccctgaattatgcttgaaaagggttaatG   1695420   8520
  AluJo (SINE)->
ATggccgggctcagcggtgacacctgtcatcccaacggtttggggtccgaaacgaggga   1695480   8580
atcgctttagcccaggaattcgagaccagtctgggcaacatagtgagacccccccatttct   1695540   8640
     AluY (SINE)->              t 5600seq
acaaattctttttctttttgagacagagtctcgctctgtcgcccaggctggagtgcaat   1695600   8700
ggcacgatctcggctcactgcaagctccacctcccaggttcacgccattctcctgcctca   1695660   8760
                rs28831989 g
gcctcccgagtagctgggattacaggcacccgccaccacgcccggctaatttttgtatt   1695720   8820
tttagtagagacggggtttcaccgtgttagccaggatggtctcgatctcctgacctcatg   1695780   8880
atccgcccgcctcggcctcccaaagtgctgggattacaggcgtgagccactgtgcctggc   1695840   8940
 AluJo->         c rs6644779
caaaaatattttaaaaattagccaggtggggtggtgcacacctgtagtcccagctactca   1695900   9000
          rs28749802 a
ggaggctgaggcgggaggatcgcttaagcctaggagatgaaggctgcagtgagctatgat   1695960   9060
tgcaccactgcactccagcctggatgacagagcaatactgagtctctttaaaaaataaaC   1696020   9120
AluSx (SINE)->
ggcagggcgcggtggctcatgcctgtcatcccagcacttttggggtccaaggtgggcgg   1696080   9180
atcacttgaggtcaggagttcaagaccagcctggccaacatggtgaaaccccgtctctac   1696140   9240
```

```
tagaagtataaaaattagccaggcatgatggtgggcacctgtcatcccatttactcggga   1696200   9300
ggctgaggcaggagaattgcttgaacccggaggaggtagaggttgcagtgagccgagatt   1696260   9360
atgccactgcactccagcctgggtgtcgcagtgagactctgtctcagaacaaaaaccaaa   1696320   9420
  A1696.3F->                      Charlie4 (DNA, MER1_type)->
aaCCAGCATCATCGTGAAAGGGGGTCGTGTGCTGacgatattttgagatgtctgcagcag   1696380   9480
                                                        AluSp->
ctctcgtgtgacttgaaaatatctgtggtttctctccgtcagcaagccaccggttttttt   1696440   9540
tgtttgtttttttcctttttttttttcagacagagtttccctcttgtcgcccaggctgga   1696500   9600
gcgcagtggcgctatctcggctcactgcaacctctgcctcccgggttcaagcgattctcc   1696560   9660
tgcctcagcctcccgagtagctgggagtacaggcgcgttccaacacgcccagataatttt   1696620   9720
                rs6644636 t
gtatttttagtagagatggggtttctctatgttggtcaggctggtctcgaactcccaacc   1696680   9780
                Charlie4 (DNA, MER1_type)->
tcaggtgatcctcctggcccccacactttcttctatcatggctgtggtttgccagtacgc   1696740   9840
tcatggttgaagggaacaccacatcttaggaagaggtagtgaaaattagagtgaagtctc   1696800   9900
      c rs7061961
ttctcctttcctggtttgtagagcccgtgaattttacctctaggcccccaagggttaaaaa   1696860   9960
        A1696.9R <-          AluSq (SINE)->
tacctATTCCCCAGCCACTGATACGCAGAgcggcctccgcctacggggttcaagcgattc   1696920   10020
tcctgcctcagcctcctgagtacctgggattacaggtgcctgccaccaagcctggctaac   1696980   10080
     AluSq (SINE)->
ttttatttttatttttatttttatttatttttctgatatggagtttcactgttgttgccc   1697040   10140
aggctggagtgcagtggcatgatctcagctcactgcaacctctccctcccaggttcaagc   1697100   10200
aattcccctgccttagcctcctgagtagctgggatttcaggcgtccaccaccacgcccgg   1697160   10260
ctaatttttgtatttttggcagagacggggtttcatcatgttggccaggctggtgtcgaa   1697220   10320
ctcctgacctcgggcgatctgcctgcctcggccttccaaagtgctaggattacaggtgtg   1697280   10380
                 AluSq/x (SINE)->
agctactgcctctggctaatttttttgtagttttggtagagatgaggtttcatcatgttgg   1697340   10440
ccaggctggtcttgaactcctgaccttgggcaatctgcctgcctcagcctcccaaagtgc   1697400   10500
                 A1697.4F->          FLAM_C (SINE)->
taggatgacgggcgtgagccactgACATCCAGCTAATtttttgtatttttggtagagacg   1697460   10560
gggtttcatcatgttgtccaggctggtctcgaactcctgacctcgggtgatctgcctgcc   1697520   10620
  g 7580seq
tcagcctcccaaagtgctaggattataggcatgagccaacatgtctggcTAATTTTTCTA   1697580   10680
TTTTTTAGTAAAGGTGGGATTTCTGTTACTGTCCCAGACAACAGTAACTTGCTACCTTTG   1697640   10740
                      A1697.72R <-              A1697.7R <-
CCTCCTTTGACCGCAGGCCTCCTGCTGGGGCTATTGACTGCTCATTCAAAGACTGTCGCT   1697700   10800
GGTTCACAGCGGTGTTCAGGGAGGTGGTCTGGTGGGTTCTAAATCTTGGAAGAGCAGTCA   1697760   10860
CTTTAAGGCACATGATACCATGTCATGGGAGTTGTCAGAAGTTTCACAGAACCAGAGTTA   1697820   10920
ATGCAGAGTTAATACCACCTTAAGAGGATCAAACTTAGTACACCCTTAACAGAAACAATT   1697880   10980
AGCTAATGCAGCTTTCATAGACCAAGAGTGAATATCGCCTTAAGAAGATAACATTTAGTA   1697940   11040
       AluSg (SINE)->
TACCCTTggccaggcacagtggctcacgcctgtaatcccagcactttgggaggccgaggc   1698000   11100
```

```
              rs28567974 a
gggcggatcacaaggtcaggagttcgagaccagcctggccaatatggtgaaatcccatct   1698060   11160
ctactaaaaatacaaaaattacctgggcgtggtggcgggcacctgtagtaccagctactt   1698120   11220
gggaggctgaggcacgagaatggcgtgaacccgggaggtggaggttgcagtgagccgaga   1698180   11280
tcacgccactgcactccagccttgggggcagagcgggaccctgtctcaaaaaaaaaaaaa   1698240   11340
    L1PA8 (LINE)->
aaaagtgggggctggggagggagagcactaggacaaatacctaatgtaggtgacgggttg   1698300   11400
acaggtgcagcaaaccaccatggcacatgtatacctgtgcaacaaacctgcacgttctgc   1698360   11460
        simple repeat (taa)n->                         rs34564551
              rs10648789 -/ataata   rs10579148 -/ata     -/aat
acatgtaccccagaacttaaagtataataataataataataataataataataataataa   1698420   11520
   L1PA8 (LINE)->
taaaaagtctttgggagcagaaaaaaaaaaaGATCAAAGTTAGCGTACCCTTAACAGAAA   1698480   11580
              A1698.5F ->  A1698.54F ->
CAGTTAATGCACCTTTCATAGACCAAGGGTGAACTCAGCCATAAGAAGGTAACATTTAAT   1698540   11640
                           A1698.6F ->       rs36094992 -/t
ATATGCCCTTAAAAGAACAATTAATTGAGACAACCCCAAAAGGAATGAATACAACTTTAA   1698600   11700
                                           AluY (SINE)->
TGAATAAAACAAAACAGAACGCTAATTTAAAAAGGTTAATACGggccgggcgtggtggct   1698660   11760
catgcctgtaatctgagcactttgggaggccgaggaaggtggatcaggaggtcaggagat   1698720   11820
tgagaccatcctggctaacacggtgaaaccccgtctctactaaaaatattttaaaaatta   1698780   11880
          -/g rs34209505
gccgggcgtggtggtgggcgcctgtagtttcagctactcgggagactgaggcaggagaat   1698840   11940
cccttgaacccgggaggcggagcttgctgtgagccgagatcgcaccactgcactccagcc   1698900   12000
tgggtgacagagccagactccgtctcaaaataaataaataaataaataaataaaAAAGAA   1698960   12060
                                        AluSx (SINE)->
AAAGGTTAATACAACATTAAAGAACAAGAATTAATATAGCtttttttttttttttttttt   1699020   12120
tttgagacctagtctcactctgttgcccaggctggagggcagtggcatgatctcagctca   1699080   12180
ctgcaacctccgcctcccgggttcaggcgattctcctgcctcagcctcccgaataggtgg   1699140   12240
gattacaggcgtgcaccaccatgcctacttttttgtatttttagtagagacggggtttcac   1699200   12300
tatgttggccaggctggtgtcgaactcttgacctcaagtgatctgcctgccttggcctcc   1699260   12360
caaagtgctgggatgacaggtgtgagccactgcacctggccTAATATAGCTTTAATAGAA   1699320   12420
AAGGAGCTAATACAAGCTTAGAATAATAAACTTTATGCAGACTTACTGAAACAAACAGTC   1699380   12480
    A1699.4F/R ->
AATGCAACCTTAACCAAACAGGAATTGATAACAGTTTAATAAAAAAGCTAATACCACTTG   1699440   12540
                                              AluSc (SINE)->
AATAGAATATATCTTACTACAACCTTCATAAGAAAAACAACTAATAcactttgggagacc   1699500   12600
gaggcaggcagatcacgaggtcaggagatcgagaccatggtgaaaccccgcctctactaa   1699560   12660
                         t 9620seq
aaatacaaaaaaattagcctggcatggtggcgggcacctgtagtcccagctagtcgggag   1699620   12720
gctgaggcaggagaatggcttgaacctgggaggtggaggttgccgtgaggcgtgatcgtg   1699680   12780
ccactgcactccagcctggcgacagagcgagactctgtctcaaaaaaataatgataataa   1699740   12840
                           AluSg/x (SINE)->
```

```
                    rs5901182 -/a
taataaacaaaTAAAGATACCAAAAAaaaaatacaaaaatttagccgggtgtggtggcac   1699800   12900
  rs7473718 t
gtgcctgtaatcacagctactcgggaggctggggcaagagaatcacttgaacgcgggagg   1699860   12960
rs4308913 a
cgtaggttgccgtgagccgagattgcaccgctgcactccagcctgggtgacagagcgaga   1699920   13020
                                A1700.01R <-
ctccgtctcaaaaaaaagaacaaacaaaaagaTAGTGGTAGGTGTAATCTGTGTTTCTTT   1699980   13080
                                      A1700.0R <-
CCCCATGCTGTTGAAATAAAGCTAACTGTAAGTATCCTGGAAGACACGTGGCCTGAAGAA   1700040   13140
ACCATCTTTCCCTGTGGTTTGTTTATGACTTCCTGTCACTAAGGTTGCCTGAGCCTTGGC   1700100   13200
                                            AluSx (SINE)->
ACCCATCCAAAACGATGTGTGCAAGAAAAGAAATGCATGGATGAtttattcttgttttat   1700160   13260
tttttttttgagatggagtctcgctctgttgcccaggctggagtgcagtggcgcgatctt   1700220   13320
ggctcactgcaatctccacctccaggattcaagcaattctcctgcctcaccctccccagt   1700280   13380
agctgggactacaggctcacaccaccatgcccagctaatttttatacttttagtagagat   1700340   13440
ggggtttcaccatgttggccaggttagccctgaactcctgacctcaggtgatccgcccac   1700400   13500
ctcggcctcccagaatgttgggattacaggcgtgagctaccacgcctggctGATTTATTC   1700460   13560
                            1700.5F/R ->                AluSp ->
ATTTTTTATAGTTTGTGTATTATTTGAAGGCATTACCTGATTACTGCTTTATTTTAtttt   1700520   13620
tatatgtatttatttatttatctattgagacggagtttcgctcttgttgcccaggctgga   1700580   13680
gtgcaatggcacgatctcggctcaccacaacctctgcctcccgggtgcaagtgattctcc   1700640   13740
tgcctcagcctcccgagtagctgggattacaggcatgcaccaccacaccaggctaatttt   1700700   13800
                rs6644637 c
gtatttttagtagagatggggtttctgcacgttggtcaggctggtcacgaactcacaacc   1700760   13860
   rs28812599 g
tcagttgctccacctgcctcggcctcccaaagtgctgggatgacaggcgtgagccaccgc   1700820   13920
          AluSx->
acccggccTAtttattttttttgagacggagtctcactctgtccagcccaggctggagtg   1700880   13980
cagtgatgcgatcttggctcactgcaacctctgcctcccgggtttacgccattctcctgc   1700940   14040
ctcagcctcccgagtagctgggatacaggtgcgcaccaccatgcctggctaattttttgta   1701000   14100
tttttagtagagacggagtttcaccatgttggccaggctggtctcgaacccctgacctca   1701060   14160
agtgatctaccctcctaggcctcccaaagtgctgggattacaggcgtgagccaccacgcc   1701120   14220
                AluSx->
cggtcCTCTGTCTCTGtttttgagacagagtcttactctgtcacccaggcttggagtgca   1701180   14280
gtggtgagatctcggctcactgcaacctccgcttagcaggttcaagcgattctcctgcct   1701240   14340
                                  rs4933103 t
cagcctcccaagtagttgggattacaggcatgagccaccacgcccggctaattttttgtat   1701300   14400
ttttagtagagatggggtttcaccatgttggccaggctggtctcaaactcctgacctcag   1701360   14460
gtgatccacccacctcccaaagcgcagggattataggtatgagttaaagcacccagccAG   1701420   14520
              1701.4F/R 1701.48F ->         AluSg->
GAGACATTTTTAAATGAGGTTGCACGGCTATGGAAAACAGGGTGGggccaggcgcaatgg   1701480   14580
gtcacgcctgtcatcccagcactttgggaggccgaggctggcagatcacgaggtcagaag   1701540   14640
```

```
ttcgagaccagcctggccaacatggtgaaaccccatctttactaaaaacacaaaaattag   1701600   14700
         rs35128021 t
acaggcatggtggaacaggcctgtaatcccagctactcgggaggctgaggcaggagaatg   1701660   14760
gtgtgaagccgggaggcagaggttgcagtgagccgagaccgcaccactgccctccagcct   1701720   14820
                                     AT-rich (low complexity)->
gggtgatacagtgagacttcaactcaaaataaataaattaattaattaataaaaataaat   1701780   14880
                                               A-rich (UCSC: (TAAA)n)->
aaataaatatGACTGTTTTCCATAGTGGTGCAATCTCATTTGACTCTGTCTCaaaaaaaa   1701840   14940
                                     rs6588802 c   A1701.09R <-
aaataaataaatgaataagaatataaataaaACAATGCTTTCCTCCCTAGCCTGCCATGG   1701900   15000
           A1701.9R <-
TATGGGGTGTTTCTAAGACTGTGGTTCCCAGGAGGCTGAGCCCTGACCTTTTATTTCCGC   1701960   15060
A1702.0R <-
TCCTGCTCCAAGGTTCTCTTCGCCGCCTGCGAGCTGGGCGTGTTTGACCTTCTCGCCGAG   1702020   15120
GCCCCAGGGCCCCTGGACGTGGCGGCAGTGGCTGCAGGTGTGAGGGCCAGCGCCCATGGG   1702080   15180
ACAGAGCTCCTGCTGGACATCTGTGTGTCCCTGAAGCTGCTGAAAGTGGAGACGAGGGGA   1702140   15240
GGAAAAGGTGAGGACACGGGCGTTTTGAAGGAGGCATTTTACATAGACACCCTCTGCAGC   1702200   15300
  MER58B (DNA, MER_1 type)
rs28515673 a                 AluY->
Ccacgtgttctgtaaaaagtgaaacagtctccatcctggctcacacagtgaaaccccgtc   1702260   15360
   rs5948991 t
tctactaaaaaaaatacaaaaaattagccgggtgtggtggcgtgcacctgtcatcccagc   1702320   15420
tactcaggaggctgaggcaggagaatggcatgaacccgggaggcagaggttgcagtgagc   1702380   15480
                                                   MER58B (DNA, MER_1 type)
cgagatcgtgccactgcactccagtctgggcgacagagcgagactccatctgtatttgca   1702440   15540
ggtttggggg tcgtgtggtttttgtggcagctaccgacctctgccctggggacgtcaaag   1702500   15600
         AluJ/FLAM
ctgccagagtgtgagacccaggctggggg attccttgagcccaggagttcgagaccagcc   1702560   15660
         AluSq/x->
       rs6588804 g
tgggccacgtatttttaatagagacaggatttcatcacgttggccaggctggtctcgagc   1702620   15720
                                   AluSq/x->
tcctgacctcaggtgatccaccagcctgggccACTtatttttagtagagacaggatttcg   1702680   15780
     rs7880188 t
tcacgttggccaggccggtctcgagctcctgacctcaggtgatccacccgcctcggcctc   1702740   15840
                             rs7883695 t
cctaagcgctgagattacaggtgtgagccaccacgcccggccTCAATAAGTTATTTTAGA   1702800   15900
CATTCTGTGGAATGTAACGTACGTTTCCTCTCGTGGCTCATCTGCCTTTGTACTTGCTGC   1702860   15960
ACCCAAACTCCCGATGGGGATTTGTACAAACCGAACCCGACGGGGGTGCTAGCCGCCCCA   1702920   16040
GGTCACCTTTGCGCCTTCCACAGAAGGCTCCAGCTGTACAAGGCAAGAGGAAGACCCCAT   1702980   16100
CTCTAAAGAAGAGATGGGCTTGTAATCATGTTGAAATCACGATGTTGTCGAGCTGGGGAG   1703040   16160
                                   rs5989834 a
CGTCCGCCGGCAGGAACTCGGAATCTCACTGCTTTTTCCCTGTTTTTTTGTGTGTGTTTC   1703100   16220
```

```
AGCTTTCTATCGAAACACAGAGCTGTCCAGCGACTACCTGACCACGGTCAGCCCGACGTC   1703160   16280
ACAATGCAGCATGCTGAAGTACATGGGCAGGACCAGCTACCGGTGCTGGGGCCACCTGGC   1703220   16340
AGACGCCGTGAGGTGGGGGCTGCCCCCAGGCAGATGCTGGGAGGTGGTGGCTGCACCCAG   1703280   16400
GCAGATGCTGGGAGGTGGGGGCTGCCCCCAGGCAGATGCTGGGAGGTGGTGGCTGACCCC   1703340   16460
AGGCAGATGCTGGGAGGTGGGGGCTGACCCCAGGCAGATGCTGGGAGGTGGGGGCTGCAC   1703400   16520
CCAGGCAGATGCTGGGAGGTGGGGGCTGCCCCCAGGCAGATGCTGGGAGGTGGTGGCTGC   1703460   16580
ACCCAGGCAGATGCTGGGAGGTGGGGGCTGCCCCCAGGCAGATGCTGGGAGGTGGTGGCT   1703520   16640
GACCACAGGCAGATGCTGGGAGGTGGGGGCTGACCCCAGGCAGATGCTGGGAGGTGGTGG   1703580   16700
      c rs28541584
CTGACCACAGGCAGATGCTGGGAGGTGGGGGCTGACCCCAGGCAGATGCTGGGAGGTGGG   1703640   16760
      c rs28754686            t rs28645863
GGCTGCACCCAGGCAGATGCTGGGAGGTGGGGGCTGCACCCAGGCAGATGCTGGGAGGTG   1703700   16820
        c rs28600310
GGGGCTGCACCCAGGCAGATGCTGGGAGGTGGGGGCTGACCCCAGGCAGATGCTGGGAGG   1703760   16880
TGGGCGCTGCACCCAGGCAGATGCTGGGAGGTGGGGGCTGACCCCAGGCAGATGCTGGGA   1703820   16940
        A1703.8R  <-        A1703.88R <-
GGTGGGGGCTGCATCCAGGCAGTGACCAAGTCTCAGGACTACCGCCAGGGGCTGGTGAAG   1703880   17000
                                              AluY->
TGGCAGTGGAGGAGGTGAAAGGGAAGGTTTCTTTTTCTTTGTTTTTtcttttcttttctt   1703940   17060
```

Appendix 4.2

Primer combinations and PCR conditions used in the analysis of LDU profiles in a North European semen donor panel.

1. **Primers and PCR conditions used for generating genotyping amplicons over the LDU step region**
2.

Amplicon 1:

Primers	**PCR conditions**
Primary PCRs (1427 bp)- 1687.0F- 5' TGCATGCATGGGGGGTATAG 3′ 1688.40R- 5' GCTGTACTGGAGGCTCAGGC 3′	96°C for 1min 96°C for 20 sec 64°C for 30 sec } x 34 cycles 68°C for 3 min
Secondary PCRs (1358 bp)- 1687.08F - 5' AGGTGGTGATGTGTTGCCCT 3′ 1688.4R - 5' GCTCAGGCACCTGAAGCTTC 3′	96°C for 1min 96°C for 20 sec 64°C for 30 sec } x 34 cycles 68°C for 3 min

Amplicon 2:

Primers	**PCR conditions**
Primary PCRs (5764 bp)- 1688.4F- 5' GAAGCTTCAGGTGCCTGAGC 3′ 1694.1R- 5' TGTCCCACAGAGCGTATCCT 3′	96°C for 1min 96°C for 20 sec 64°C for 30 sec } x 32 cycles 68°C for 6 min
Secondary PCRs (5744 bp)- 1688.40F - 5' GCCTGAGCCTCCAGTACAGC 3′ 1694.16R- 5' AGAGCGTATCCTACCTGGGA 3′	96°C for 1min 96°C for 20 sec 64°C for 30 sec } x 32 cycles 68°C for 6 min

Amplicon 3:

Primers	**PCR conditions**
Primary PCRs (4027 bp)- 1693.7F- 5' CCTGCTCAATCCATAAGACG 3′ 1697.7R- 5' TGTGAACCAGCGACAGTCTT 3′	96°C for 1min 96°C for 20 sec 61°C for 30 sec } x 32 cycles 65°C for 5 min

Secondary PCRs (3934 bp)- 1693.8F – 5' CTACAGCTGATGAGTCAAGG 3' 1697.72R- 5' AGCAGTCAATAGCCCCAGCA 3'	96°C for 1min 96°C for 20 sec 61°C for 30 sec 65°C for 5 min x 32 cycles

Amplicon 4:

Primers	**PCR conditions**
Primary PCRs (5200 bp)- 1694.8F- 5' CACCCGTAGAGGAGCAACAC 3' 1700.0R- 5' TTCAGGCCACGTGTCTTCCA 3'	96°C for 1min 96°C for 20 sec 64°C for 30 sec 65°C for 6 min x 32 cycles
Secondary PCRs (3650 bp)- 1696.3F – 5' CCAGCATCATCGTGAAAGGG 3' 1700.01R- 5' AGTGGTAGGTGTAATCTGTG 3'	96°C for 1min 96°C for 20 sec 62°C for 30 sec 65°C for 6 min x 32 cycles

Amplicon 5:

Primers	**PCR conditions**
Primary PCRs (3474 bp)- 1698.5F- 5' GACCAAGGGTGAACTCAGCC 3' 1702.0R- 5' AACCTTGGAGCAGGAGCGGA 3'	96°C for 1min 96°C for 20 sec 62°C for 30 sec 65°C for 4 min x 32 cycles
Secondary PCRs (3429 bp)- 1698.5F- 5' GACCAAGGGTGAACTCAGCC 3' 1701.9R- 5' TGGGAACCACAGTCTTAGAA 3'	96°C for 1min 96°C for 20 sec 61°C for 30 sec 65°C for 4 min x 32 cycles

Amplicon 6:

Primers	**PCR conditions**
Primary PCRs (2421 bp)- 1701.48F- 5' GACCAAGGGTGAACTCAGCC 3' 1703.88R- 5' AACCTTGGAGCAGGAGCGGA 3'	96°C for 1min 96°C for 20 sec 64°C for 30 sec 68°C for 4 min x 34 cycles
Secondary PCRs (2401 bp)- 1701.48F- 5' GACCAAGGGTGAACTCAGCC 3' 1703.8R- 5' TGGCGGTAGTCCTGAGACTT 3'	96°C for 1min 96°C for 20 sec 64°C for 30 sec 68°C for 4 min x 34 cycles

2. Primers and PCR conditions used for generating amplicons for genotyping markers within repetitive DNA

To type rs7472200G/T and rs658801C/T:

Primary PCRs used: Amplicon 2

Primers	**PCR conditions**
Secondary PCRs- 1693.1F- 5' GTGCAGGCATTTAGATTGTC 3′ 1694.0R- 5' AGCCTCTGGAGCGCTTGCTT 3′	96°C for 1min 96°C for 20 sec 57.2°C for 30 sec 68°C for 1:30 min } x 32 cycles

To type rs28464926C/T, rs28649445C/G, rs28560383A/G and rs28550202C/T:

Primary PCRs used: Amplicon 2

Primers	**PCR conditions**
Secondary PCRs- 1689.9F- 5' CCCCCAAATCACCTTCTAAC 3′ 1690.3R- 5' TGTCTTACCTGGTGACAGGC 3′	96°C for 1min 96°C for 20 sec 56°C for 30 sec 65°C for 1 min } x 36 cycles

To type rs5949021C/G, rs4933062C/T, rs6644630C/T and rs6644632C/G:

Primary PCRs used: Amplicon 2

Primers	**PCR conditions**
Secondary PCRs- 1690.3F- 5' GCCTGTCACCAGGTAAGACA 3′ 1691.1R- 5' TGAGACTGCGTAATTTTGGG 3′	96°C for 1min 96°C for 20 sec 56°C for 30 sec 65°C for 1:30 min } x 36 cycles

To type rs4308913A/C:

Primary PCRs used: Amplicon 4

Primers	**PCR conditions**
Secondary PCRs- 1699.4F- 5' GCAACCTTAACCAAACAGGA 3′ 1700.01R- 5' AGTGGTAGGTGTAATCTGTG 3′	96°C for 1min 96°C for 20 sec 55.4°C for 30 sec 68°C for 1:10 min } x 32 cycles

To type 7580seqA/G:

Primary PCRs used: Amplicon 4

Primers	**PCR conditions**
Secondary PCRs- 1697.4F- 5' GCCACTGACATCCAGCTAAT 3' 1697.72R- 5' AGCAGTCAATAGCCCCAGCA 3'	96°C for 1min 96°C for 20 sec 64°C for 30 sec 66°C for 1 min $\}$ x 32 cycles

3. Primer used for re-sequencing selected regions between 1688 – 1702 kb

Note: All positions are as per NCBI36/hg18, all primer sequences are in the 5' to 3' direction.

Forward/ reverse primers are indicated using 'F/R' at the end of the primer name.

Primer name	Start position	Primer sequence
1688.40F	1688431	GCCTGAGCCTCCAGTACAGC
1691.7F	1691792	CTGTCTCCAGGTAAGGTGTC
1692.5F	1692555	CCCCTAGAAAATCACTCTGG
1692.5R	1692574	CCAGAGTGATTTTCTAGGGG
1693.1F	1693159	GTGCAGGCATTTAGATTGTC
1693.1R	1693159	GACAATCTAAATGCCTGCAC
1693.8F	1693807	CTACAGCTGATGAGTCAAGG
1694.16R	1694164	TGTCCCACAGAGCGTATCCT
1694.8R	1694915	GTGTTGCTCCTCTACGGGTG
1695.1F	1695180	ccaaggCCTCATTTCTTCTt
1695.1R	1695199	AAGAAGAAATGAGGCCTTGG
1696.3F	1696382	CCAGCATCATCGTGAAAGGG
1696.3R	1696401	CCCTTTCACGATGATGCTGG
1696.9R	1696947	CTGCGTATCAGTGGCTGGGG
1697.72R	1697741	AGCAGTCAATAGCCCCAGCA
1698.5F	1698560	GACCAAGGGTGAACTCAGCC
1699.4F	1699443	GCAACCTTAACCAAACAGGA
1699.4R	1699462	TCCTGTTTGGTTAAGGTTGC
1700.01R	1700032	AGTGGTAGGTGTAATCTGTG
1700.5F	1700547	AGGCATTACCTGATTACTGC
1700.5R	1700566	GCAGTAATCAGGTAATGCCT
1701.4F	1701494	TGAGGTTGCACGGCTATGGA
1701.4R	1701513	TCCATAGCCGTGCAACCTCA
1701.90R	1701970	AACACCCCATACCATGGCAG

Appendix 4.3.

Genotyping SNPs in the PAR1 LDU step region in a North European semen donor panel

1. Sequences of ASOs used for genotyping

All positions are as per NCBI36/hg18. All sequences are from 5' to 3'. Single-copy DNA sequences are in uppercase, while ASOs located within repeat sequences are in lowercase.

dbSNP entry	Position	ASO sequence (5' to 3')
rs28698568A/C	1687173	GACCGGC**C**AGCTCGGGCA GACCGGC**A**AGCTCGGGCA
rs28612605A/G	1687212	CCCAGGC**A**TCCCGTCTGG CCCAGGC**G**TCCCGTCTGG
rs28640198A/G	1687325	CGTCCTC**A**AGCCCCGTCC CGTCCTC**G**AGCCCCGTCC
rs28412948C/T	1687339	CCCGTCCACC**T**GCCCAGG CCCGTCCACC**C**GCCCAGG
rs28681628G/T	1687348	CCCAGGA**G**CTGGGGCATC CCCAGGA**T**CTGGGGCATC
rs28582979C/T	1687365	CCTGTGT**C**TCCCGCCTTG CCTGTGT**T**TCCCGCCTTG
rs5949017A/G	1687661	GGCACTC**A**GTCCGGAAGC GGCACTC**G**GTCCGGAAGC
rs5989848G/T	1688337	TGGTCGG**T**AGGGTCCCAG TGGTCGG**G**AGGGTCCCAG
rs5989849C/G	1688395	CGCTGCC**G**CGGTGAGACT CGCTGCC**C**CGGTGAGACT
rs28517389C/G	1689654	gccagac**g**ctgtctcaag gccagac**c**ctgtctcaag
rs28696100C/T	1689805	aacagcc**c**ggacgcactg aacagcc**t**ggacgcactg
rs7877401C/T	1689946	tccagct**c**tccctcccct tccagct**t**tccctcccct
rs28464926C/T	1690210	atgttgg**t**aggctggtct atgttgg**c**aggctggtct
rs28649445C/G	1690245	tgatcca**g**ctgtcttcgc tgatcca**c**ctgtcttcgc
rs28560383A/G	1690271	actgccc**a**gatgacacat actgccc**g**gatgacacat
rs28550202C/T	1690337	tcccctg**t**tgcacacgcc tcccctg**c**tgcacacgcc
rs5949021C/G	1690735	agatcga**g**ccactgcact agatcga**c**ccactgcact
rs4933062C/T	1690955	aggcggg**c**ggatcacgag aggcggg**t**ggatcacgag
rs6644630C/T	1690996	gctaaca**c**ggtgaaaccc gctaaca**t**ggtgaaaccc

dbSNP entry	Position	ASO sequence (5' to 3')
rs6644632C/G	1691072	agctact**g**gggaggctga agctact**c**gggaggctga
rs6644633A/G	1691375	aggtcag**g**agttcgagat aggtcag**a**agttcgagat
rs5948999G/A	1692170	cctcact**g**tgttgcccag cctcact**a**tgttgcccag
rs6644775C/G	1692602	cgcctgt**c**atcccagcaa cgcctgt**g**atcccagcaa
rs6644777A/G	1692823	AAATGGG**G**TAGAGACCTG AAATGGG**A**TAGAGACCTG
rs7472200G/T	1693347	gcacttt**g**ggaggccgag gcacttt**t**ggaggccgag
rs658801C/T	1693370	tggatca**c**cttaggtcag tggatca**t**cttaggtcag
rs4446909A/G	1693782	GATGGGT**G**AGCATATTGG GATGGGT**A**AGCATATTGG
rs5989681C/G	1693891	GGGAATC**G**CCACACGCTT GGGAATC**C**CCACACGCTT
3920SEQG/A	1693960	TAATACG**G**TTAATGCAAA TAATACG**A**TTAATGCAAA
rs6644635C/T	1693993	CAGCAGC**T**GTGAGCGGGT CAGCAGC**C**GTGAGCGGGT
4340SEQC/T	1694373	tacattt**c**acagagacat tacattt**t**acagagacat
rs5948839A/T	1694385	acatgag**t**catcaatcaa acatgag**a**catcaatcaa
rs7391804C/A	1694664	gcactcc**c**gcctggggga gcactcc**a**gcctggggga
rs28633458G/A	1694739	agcttaa**g**atgaaataca agcttaa**a**atgaaataca
rs5989853A/G	1695252	gccccat**a**tggtatgact gccccat**g**tggtatgact
5600seqC/T	1695632	agagtct**c**gctctgtcgc agagtct**t**gctctgtcgc
rs6644779T/C	1695916	ttaaaaa**t**tagccaggtg ttaaaaa**c**tagccaggtg
rs6644636C/T	1696705	ggtttct**c**tatgttggtc ggtttct**t**tatgttggtc
rs7061961C/T	1696865	cttctcc**t**tcctggtttg cttctcc**c**tcctggtttg
7580seqA/G	1697582	ctgcctc**a**gcctcccaaa ctgcctc**g**gcctcccaaa
9620seqG/T	1699645	ctggcat**g**gtggcgggca ctggcat**t**gtggcgggca
rs4308913A/C	1699929	aggttgc**c**gtgagccgag aggttgc**a**gtgagccgag

dbSNP entry	Position	ASO sequence (5' to 3')
rs6644637C/G	1700785	ggtttct**g**cacgttggtc
		ggtttct**g**cacgttggtc
rs7883695T/G	1702839	acgcccg**g**ccTCAATAAG acgcccg**t**ccTCAATAAG
rs28541584A/C	1703896	GCTGACC**A**CAGGCAGATG GCTGACC**C**CAGGCAGATG

2. Genotypes of North European semen donors

Genotypes of 92 North European semen donors are shown. Numbers (1-96) indicated local donor names. Samples 31 and 86, 50 and 95, 30 and 33, 19 and 49 are duplicates. Data for both samples from a pair of duplicates are shown below. Markers identified through re-sequencing are named according to approximate locations, followed by 'seq'. Donors who are heterozygous (H) for alleles at a given SNP are indicated in red.

SNP name	location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
rs28698568A/C	1687173	H	C	C	C	C	H	C	C	C	C	C	C	C	H	H	C	H	C	H	H	H	H	C	?	C	C	H	H	C	C	C	C
rs28612605A/G	1687212	H	H	H	G	A	G	G	G	A	A	G	G	A	G	H	H	G	G	H	H	H	G	H	?	H	G	G	G	H	G	H	G
rs28640198A/G	1687325	H	H	H	G	A	G	G	G	A	A	G	G	A	G	G	H	G	G	H	H	H	H	H	?	H	G	G	G	H	G	H	G
rs28412948C/T	1687339	H	H	H	C	T	C	C	C	T	T	C	C	T	C	H	H	C	C	H	H	H	H	H	?	H	C	C	C	H	C	H	C
rs28681628G/T	1687348	H	G	G	G	G	H	G	G	G	G	G	G	G	H	H	G	H	G	H	H	H	H	G	?	G	G	H	H	G	G	G	G
rs28582979C/T	1687365	H	H	H	T	C	T	T	T	C	C	T	T	C	T	H	H	T	T	H	H	H	H	H	?	H	T	T	T	H	T	H	T
rs5949017A/G	1687661	G	H	H	G	H	G	G	G	H	A	G	G	H	G	H	H	G	G	H	G	H	H	G	?	H	G	G	G	H	G	G	G
rs5989848G/T	1688337	G	G	T	T	G	G	G	T	G	G	H	T	G	H	G	H	H	T	G	G	G	G	G	?	H	T	G	G	H	T	H	H
rs5989849C/G	1688395	C	C	H	C	C	C	C	C	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	?	C	C	C	C	C	C	C	C
rs28517389C/G	1689654	G	G	G	H	H	G	H	H	H	G	C	G	G	G	G	G	H	G	G	G	G	G	C	C	G	H	G	H	H	H	C	H
rs28696100C/T	1689805	T	T	C	H	T	C	T	H	H	T	T	H	T	H	C	H	H	H	T	H	H	T	T	T	T	H	C	T	T	T	T	T
rs7877401C/T	1689946	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs28464926C/T	1690210	C	H	C	C	C	C	C	H	C	H	C	C	H	H	H	C	H	C	C	H	H	H	C	C	H	C	C	H	C	H	C	H
rs28649445C/G	1690245	C	H	C	C	C	C	C	C	C	H	C	C	C	H	C	C	H	C	C	C	C	C	C	C	C	C	C	H	C	H	C	H
rs28560383A/G	1690271	G	H	A	G	H	G	H	A	G	H	G	H	A	H	A	G	H	G	H	H	A	H	A	H	H	A	G	H	H	H	H	H
rs28550202C/T	1690337	C	H	C	C	C	C	C	H	C	H	C	C	H	H	H	C	H	C	C	H	H	H	C	C	H	C	C	H	C	H	C	H
rs5949021C/G	1690735	C	C	G	G	C	?	G	G	C	C	G	G	C	?	G	C	G	G	C	?	C	C	G	G	C	G	?	G	H	G	H	G
rs4933062C/T	1690955	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs6644630C/T	1690996	H	C	C	H	C	T	H	H	C	C	H	H	C	T	H	H	H	H	H	T	H	H	C	H	H	H	T	H	C	H	H	H
rs6644632C/G	1691072	H	G	G	C	G	C	C	C	G	G	H	H	G	C	H	H	H	H	H	C	H	H	G	C	H	C	C	C	G	C	H	C
rs6644633A/G	1691375	G	G	G	H	G	G	H	H	G	G	H	H	G	G	G	G	G	G	G	G	G	G	G	A	G	H	G	A	G	A	G	A
rs5948999G/A	1692170	H	H	H	A	H	G	H	H	A	H	H	A	A	A	G	H	A	A	G	H	H	G	A	A	H	A	G	H	H	H	G	H
rs6644775C/G	1692602	G	H	H	G	G	G	H	H	G	G	G	G	H	G	C	H	C	H	H	H	H	G	H	G	G	G	H	C	H	C	H	C
rs6644777A/G	1692823	H	H	H	H	A	A	H	H	A	A	H	H	A	A	G	H	G	H	H	H	H	H	H	G	A	H	H	G	H	H	H	G
rs7472200G/T	1693347	H	H	H	G	H	T	G	H	H	T	H	H	T	H	G	G	G	G	G	H	H	H	H	G	H	G	H	G	H	H	H	G
rs658801C/T	1693370	H	H	H	C	H	T	C	H	H	T	H	H	T	H	C	C	C	C	C	H	H	H	H	C	H	H	H	C	H	H	H	C
rs4446909A/G	1693782	A	H	H	H	H	A	G	A	H	H	A	H	A	G	G	G	G	G	G	H	H	H	H	G	H	H	H	G	G	H	H	G
rs5989681C/G	1693891	C	H	H	H	H	C	G	C	H	C	C	C	H	H	G	G	G	G	G	H	H	H	H	G	H	C	H	G	G	H	H	G
3920SEQG/A	1693960	G	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	H	G	G	G	H	G	G	G	G	H
rs6644635C/T	1693993	C	C	C	C	C	C	H	C	C	C	C	C	C	C	H	H	T	H	H	C	H	C	H	C	C	C	H	C	C	H	H	T
4340SEQC/T	1694373	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	H
rs5948839A/T	1694385	A	A	A	A	A	A	H	A	A	A	A	A	A	A	A	H	T	A	H	A	A	A	A	A	A	A	A	A	A	A	H	H
rs7391804C/A	1694664	A	A	A	A	A	A	A	A	A	A	H	A	H	A	H	A	H	H	H	A	A	H	H	A	A	A	A	H	H	A	H	A
rs28633458G/A	1694739	G	G	G	H	H	G	H	G	H	H	G	H	G	G	G	H	H	G	G	G	H	G	G	G	G	H	H	G	G	H	G	H
rs5989853A/G	1695252	H	A	A	A	A	H	A	H	A	A	H	H	H	A	G	A	H	A	A	H	A	A	A	A	A	H	H	H	G	H	A	H
5600seqC/T	1695632	C	C	C	C	C	C	C	H	C	C	C	C	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	H	C	C
rs6644779T/C	1695916	H	T	C	T	T	T	T	T	T	T	T	T	C	H	T	T	H	H	H	T	T	H	T	H	T	T	T	T	H	H	H	T
rs6644636C/T	1696705	C	H	H	C	C	C	C	C	C	C	H	H	C	C	C	C	C	C	C	C	H	C	H	C	C	H	C	C	H	H	C	C
rs7061961C/T	1696865	H	T	T	C	H	T	H	T	T	H	T	T	T	T	C	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	C
7580seqA/G	1697582	A	H	A	A	A	A	H	A	A	A	A	A	A	A	A	A	H	A	A	A	H	A	H	A	A	A	H	H	?	A	A	A
9620seqG/T	1699645	H	H	G	G	G	G	H	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	H	G	G	G	G	H	G	G	G	G
rs4308913A/C	1699929	C	C	C	C	H	C	H	C	C	H	H	H	?	C	A	C	C	C	C	H	H	C	C	C	C	H	C	C	C	C	C	A
rs6644637C/G	1700785	H	C	H	H	C	H	C	G	H	H	C	C	H	G	C	G	H	G	G	C	C	G	G	H	G	C	H	C	H	C	H	C
rs6588802C/T	1701946	H	C	H	H	C	H	C	T	H	T	C	C	H	T	C	T	H	T	T	C	C	T	C	H	H	C	H	C	H	C	H	H
rs6588802C/T	1701946	H	C	H	H	C	H	C	T	H	T	C	C	H	T	C	T	H	T	T	C	C	T	C	H	H	C	H	C	H	C	H	H
rs7883695T/G	1702839	G	H	G	H	G	G	G	G	G	G	H	H	G	G	G	G	G	G	G	G	G	G	H	G	G	H	G	G	G	G	G	G
rs28541584A/C	1703896	H	C	H	H	C	H	C	H	H	H	C	C	H	H	C	H	H	H	H	C	C	H	C	H	H	C	H	C	H	C	H	H

SNP name	location	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
rs28698568A/C	1687173	C	C	C	C	C	H	C	C	H	C	C	C	H	H	H	C	H	C	C	C	C	C	H	C	C	C	C	C	H	C	C	H
rs28612605A/G	1687212	G	H	H	G	H	G	A	H	H	H	G	H	G	H	G	H	H	H	G	H	H	H	H	H	H	H	G	H	G	A	H	G
rs28640198A/G	1687325	G	H	H	G	H	G	A	H	H	H	G	H	G	H	G	H	H	H	G	H	H	H	H	H	H	H	G	H	G	A	H	G
rs28412948C/T	1687339	C	H	H	C	H	C	T	H	H	H	C	H	C	H	C	H	H	H	C	H	H	H	H	H	H	H	C	H	C	T	H	C
rs28681628G/T	1687348	G	G	G	G	G	H	G	G	H	G	G	G	H	H	H	G	H	G	G	G	G	G	H	G	G	G	H	G	H	G	G	H
rs28582979C/T	1687365	T	H	H	T	H	T	C	H	H	H	T	H	T	H	T	H	H	H	T	H	H	H	H	H	H	H	T	H	T	C	H	T
rs5949017A/G	1687661	G	G	H	G	G	G	A	G	H	H	G	H	G	H	G	H	G	H	G	H	G	G	H	H	H	G	G	H	G	H	H	G
rs5989848G/T	1688337	T	G	H	H	H	H	G	G	G	H	H	H	H	G	H	H	G	G	H	G	H	H	H	H	H	H	H	H	G	G	T	H
rs5989849C/G	1688395	C	C	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	H	C	C	C	C	C	C	C	H	C
rs28517389C/G	1689654	H	G	H	H	H	H	G	G	G	G	C	H	H	H	G	G	H	H	G	H	G	C	G	H	H	G	H	H	G	G	H	G
rs28696100C/T	1689805	H	H	H	T	H	H	T	C	T	T	T	T	H	H	H	C	H	T	C	T	T	T	C	T	T	T	H	T	C	T	H	C
rs7877401C/T	1689946	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs28464926C/T	1690210	C	C	C	C	C	C	C	C	C	T	C	C	C	C	H	H	C	H	C	C	H	H	C	H	C	C	C	C	C	T	C	C
rs28649445C/G	1690245	C	C	C	C	C	C	C	C	C	H	C	C	C	C	H	C	C	H	C	C	H	C	C	H	C	C	C	C	C	G	C	C
rs28560383A/G	1690271	H	G	H	H	G	G	G	A	G	A	H	H	G	H	H	H	H	H	G	H	H	H	A	H	G	H	G	H	G	A	A	G
rs28550202C/T	1690337	C	C	C	C	C	C	C	C	C	T	C	C	C	C	H	H	C	H	C	C	H	H	C	H	C	C	C	C	C	T	C	C
rs5949021C/G	1690735	G	C	G	G	H	C	C	G	C	C	G	H	G	C	?	G	C	H	G	H	H	H	G	G	H	C	G	H	?	C	G	G
rs4933062C/T	1690955	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	H	T
rs6644630C/T	1690996	H	H	H	H	C	H	C	H	H	C	H	C	H	H	T	H	H	C	H	C	H	H	H	H	C	C	H	H	T	C	H	H
rs6644632C/G	1691072	H	H	H	C	G	H	G	H	C	G	H	G	H	H	C	H	H	G	H	G	H	H	H	C	G	G	C	H	C	G	H	C
rs6644633A/G	1691375	H	H	H	H	G	G	G	G	H	G	H	G	G	G	G	G	G	G	G	G	H	G	H	H	G	G	H	H	G	G	H	G
rs5948999G/A	1692170	A	H	H	H	G	H	H	H	H	H	A	H	G	H	G	G	H	H	H	H	G	G	H	H	G	H	G	G	A	G	G	H
rs6644775C/G	1692602	G	H	H	C	H	G	H	C	H	H	C	C	H	G	H	H	C	H	H	H	H	H	C	H	G	C	G	G	G	H	H	C
rs6644777A/G	1692823	H	G	H	G	G	A	H	G	H	H	G	G	G	H	A	H	G	H	G	H	H	H	G	G	A	G	A	H	H	A	G	G
rs7472200G/T	1693347	H	G	G	G	G	H	H	G	H	H	G	G	G	H	T	H	G	H	H	H	H	H	G	G	T	G	H	H	H	T	G	H
rs658801C/T	1693370	H	C	C	C	C	H	H	C	H	H	C	C	C	H	T	H	C	H	H	H	H	H	C	C	T	C	H	H	H	T	C	H
rs4446909A/G	1693782	A	G	G	G	G	H	H	G	H	H	G	G	G	H	A	H	G	H	H	H	H	H	G	G	A	G	H	H	H	A	G	H
rs5989681C/G	1693891	H	H	G	G	G	H	H	G	H	H	G	G	G	H	C	H	G	H	H	H	H	H	G	G	C	G	H	H	H	C	G	H
3920SEQG/A	1693960	G	G	H	G	H	G	G	G	G	H	G	G	G	G	G	G	H	G	G	G	H	G	H	G	G	H	G	G	G	G	G	G
rs6644635C/T	1693993	C	H	C	H	T	C	H	H	C	H	C	T	H	H	C	H	T	C	H	C	H	H	T	H	C	H	H	C	C	C	H	H
4340SEQC/T	1694373	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs5948839A/T	1694385	H	H	A	H	H	A	H	H	A	H	A	T	H	H	A	H	H	A	T	A	A	H	H	H	A	H	H	A	A	A	H	H
rs7391804C/A	1694664	H	H	H	A	H	H	A	H	A	H	A	A	A	H	A	A	H	H	H	A	H	H	H	A	H	H	A	A	A	A	H	A
rs28633458G/A	1694739	G	G	G	G	H	H	A	G	G	H	G	H	G	H	G	H	H	H	G	H	G	G	G	H	H	G	H	G	G	H	G	G
rs5989853A/G	1695252	A	A	H	A	H	A	H	H	A	A	A	A	A	A	H	H	A	H	A	A	A	A	A	A	A	A	A	A	A	H	H	A
5600seqC/T	1695632	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs6644779T/C	1695916	A	A	C	H	A	H	A	A	A	H	A	H	H	A	H	H	H	H	A	C	H	H	A	A	A	A	H	A	A	H	A	H
rs6644636C/T	1696705	C	C	C	C	C	C	C	H	H	C	C	H	C	H	H	C	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs7061961C/T	1696865	H	T	T	T	T	T	H	T	H	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	H	H	T	T	T	T	H
7580seqA/G	1697582	H	A	A	H	H	A	A	H	H	H	H	A	H	H	H	A	H	A	H	A	A	A	H	?	H	H	A	H	H	H	A	A
9620seqG/T	1699645	H	H	G	H	H	G	G	G	G	H	H	G	H	H	H	G	H	G	H	G	G	G	H	G	H	G	G	H	H	H	G	G
rs4308913A/C	1699929	H	C	C	C	C	C	H	H	C	H	C	C	C	C	C	C	?	C	C	C	C	C	C	C	H	C	H	C	C	?	C	?
rs6644637C/G	1700785	C	H	G	H	H	G	H	C	C	C	C	H	H	C	C	H	C	H	C	G	H	H	C	C	C	G	H	C	C	H	G	H
rs6588802C/T	1701946	C	H	T	H	H	T	H	C	C	H	C	H	H	C	C	H	C	H	C	T	H	H	C	C	H	T	H	C	C	H	T	H
rs6588802C/T	1701946	C	H	T	H	H	T	H	C	C	H	C	H	H	C	C	H	C	H	C	T	H	H	C	C	H	T	H	C	C	H	T	H
rs7883695T/G	1702839	G	G	G	G	G	G	G	G	H	G	G	G	G	H	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
rs28541584A/C	1703896	C	H	H	H	H	H	H	C	C	C	C	H	H	C	C	H	C	H	C	H	H	H	C	C	H	H	H	C	C	H	H	H

Appendix 4.4.

1. Primer combinations and PCR conditions used in Assay ‘B’

d30 PCRs:

Primers	**PCR conditions**
Primary PCRs: FAS6909A- 5' GAGTGCTAATGATGGGTA 3′ 1699.4R- 5' TCCTGTTTGGTTAAGGTTGC 3′ FAS6909G- 5' GAGTGCTAATGATGGGTG 3′ 1699.4R- 5' TCCTGTTTGGTTAAGGTTGC 3′	96°C for 1min 96°C for 20 sec 55°C for 30 sec x 2 cycles 65°C for 6:30 min 96°C for 20 sec 54°C for 30 sec x 26 cycles 65°C for 6:00 min
Secondary PCRs: FAS9681C- 5' CAGAAAGCCAGGGAATCG 3′ 1697.7R- 5' TGTGAACCAGCGACAGTCTT 3′ FAS9681G- 5' CAGAAAGCCAGGGAATCG 3′ 1697.7R- 5' TGTGAACCAGCGACAGTCTT 3′	96°C for 1min 96°C for 20 sec 66°C for 30 sec x 2 cycles 68°C for 4:30 min 96°C for 20 sec 64°C for 30 sec x 12 cycles 68°C for 4:30 min

d74 PCRs:

Primers	**PCR conditions**
Primary PCRs: FAS6909A- 5' GAGTGCTAATGATGGGTA 3′ 1700.01R- 5' AGTGGTAGGTGTAATCTGTG 3′ FAS6909G- 5' GAGTGCTAATGATGGGTG 3′ 1700.01R- 5' AGTGGTAGGTGTAATCTGTG 3′	96°C for 1min 96°C for 20 sec 55°C for 30 sec x 2 cycles 65°C for 8:00 min 96°C for 20 sec 54°C for 30 sec x 26 cycles 65°C for 7:30 min
Secondary PCRs: FAS9681C- 5' CAGAAAGCCAGGGAATCG 3′ 1700.01R- 5' AGTGGTAGGTGTAATCTGTG 3′ FAS9681G- 5' CAGAAAGCCAGGGAATCG 3′ 1700.01R- 5' AGTGGTAGGTGTAATCTGTG 3′	96°C for 1min 96°C for 20 sec 66°C for 30 sec x 2 cycles 68°C for 7:30 min 96°C for 20 sec 64°C for 30 sec x 10 cycles 68°C for 7:00 min

2. Primer conditions and PCR conditions used to generate tertiary products

- **d30 PCRs:**

Amplicon A: 287 bp

Primers	PCR conditions
1693.86F- 5' CACGCTTTGTGATCCCCTAA 3' 1694.1R- 5' TGTCCCACAGAGCGTATCCT 3'	96°C for 1min $\left.\begin{array}{l}96^\circ\text{C for 20 sec}\\ 56^\circ\text{C for 30 sec}\\ 65^\circ\text{C for 1:00 min}\end{array}\right\}$ x 32 cycles

Amplicon B: 751 bp

Primers	PCR conditions
1694.1F- 5' AGGATACGCTCTGTGGGACA 3' 1694.8R- 5' GTGTTGCTCCTCTACGGGTG 3'	96°C for 1min $\left.\begin{array}{l}96^\circ\text{C for 20 sec}\\ 56^\circ\text{C for 30 sec}\\ 65^\circ\text{C for 1:00 min}\end{array}\right\}$ x 32 cycles

Amplicon C: 1221 bp

Primers	PCR conditions
1695.1F- 5' CCAAGGCCTCATTTCTTCTT 3' 1696.3R- 5' CCCTTTCACGATGATGCTGG 3'	96°C for 1min $\left.\begin{array}{l}96^\circ\text{C for 20 sec}\\ 56^\circ\text{C for 30 sec}\\ 65^\circ\text{C for 1:30 min}\end{array}\right\}$ x 32 cycles

Amplicon D: 565 bp

Primers	PCR conditions
1696.3F- 5' CCAGCATCATCGTGAAAGGG 3' 1696.9R- 5' CTGCGTATCAGTGGCTGGGG 3'	96°C for 1min $\left.\begin{array}{l}96^\circ\text{C for 20 sec}\\ 56^\circ\text{C for 30 sec}\\ 65^\circ\text{C for 1:00 min}\end{array}\right\}$ x 32 cycles

- **d74 PCRs:**

Amplicon E: 1108 bp

Primers	PCR conditions
1693.8F- 5' CTACAGCTGATGAGTCAAGG 3′ 1694.8R- 5' GTGTTGCTCCTCTACGGGTG 3′	96°C for 1min 96°C for 20 sec 57°C for 30 sec 65°C for 2:00 min x 32 cycles

Amplicon F: 1767 bp

Primers	PCR conditions
1695.1F- 5' CCAAGGCCTCATTTCTTCTT 3′ 1696.9R- 5' CTGCGTATCAGTGGCTGGGG 3′	96°C for 1min 96°C for 20 sec 56°C for 30 sec 65°C for 1:10 min x 32 cycles

Amplicon G: 264 bp

Primers	PCR conditions
1697.4F- 5' GCCACTGACATCCAGCTAAT 3′ 1697.72R- 5' AGCAGTCAATAGCCCCAGCA 3′	96°C for 1min 96°C for 20 sec 55°C for 30 sec 65°C for 1:40 min x 32 cycles

Amplicon H: 568 bp

Primers	PCR conditions
1699.4F- 5' GCAACCTTAACCAAACAGGA 3′ 1700.01R- 5' CACAGATTACACCTACCACT 3′	96°C for 1min 96°C for 20 sec 56°C for 30 sec 65°C for 2:30 min x 32 cycles

Chapter 5 Appendices

Appendix 5.1

Determining location of markers used in low-resolution linkage analysis of CEPH pedigrees.

- SDF-1: Location of SDF-1 is determined from a primer sequence (0106) used by Schmitt *et al.* (1994) for analysis of the polymorphic marker.

- SDF-2: Location of SDF-2 is as stated in the UCSC Genome Browser.

- LH-1: Primer sequences used for the amplification of the LH-1 locus, as reported by Li and Hamer (1995), were first used to identify the approximate location for this marker. Co-ordinates for a 'GTTT' sequence (LH-1 is a 'GTTT' indel polymorphism) within this region were used as location of LH-1.

- LH-2: LH-2 was reportedly very close to the protein-coding gene *Interleukin 9 receptor (ILR9)*. The sequence in and around this gene was searched for (A)n sequences (LH-2 is an (A)n polymorphism). Co-ordinates for one such poly-A sequence were used to denote LH-2. This location is only an approximation, since its true physical location could not be ascertained.

- rs802488, rs963311 and rs1883051 were located using the UCSC Genome Browser.

- JXYQ: JXYQ reportedly lies in the boundary between PAR2 and sex-specific regions of X and Y. Therefore, co-ordinates for the first base within PAR2 is used in the figure.

Appendix 5.2

Annotated sequence, ChrX: 154640470-154656490

Note:

- Each line of sequence consists of 60bp, position of the first base of each line is indicated to the left of that line.
- Generally, repeat sequences are indicated in red. Alus are in green. Repeat type is indicated near the repeat start position.
- SNPs are indicated in **blue/bold.** ‘rs’ numbers are written above respective SNPs. SNPs identified through donor re-sequencing are indicated by ‘sequencing’ written above the SNP. The donor/ haplotype wherein the polymorphism was identified is also indicated below the SNP.
- Primer sequences are underlined, primer names are written above the sequence and direction (forward/ reverse) is indicated by arrows. If two primer sequences overlap, one of them is highlighted in yellow.

```
                                 DNA MER1->                    1.1F->
154640470  GTATCATGCTTTCTGCATGATACATCCAAGGATCATTTGGGACTTTTTGAAAATGCAGCT
154640530  TCTCTGGGTCTCACTCAAGATCTCTGAACAAAAAATCTCTGGAGGCAAGGTTTAGGAGTA
154640590  TGTAGCTTGATAAAGTTCCATATATAATTCTGATGTGCAGCCCTGCCTGAGAATCACACT
                                             1.2F->
154640650  CTAAAAAGAAACACGCACCTCCACCAAATTTACTAGTAATTCCAACAGAGTGGTCACATG
                                                          rs73249628
154640710  AGATGGGTGCCCATCAGCAGTGGATTGGATAAACAAGTATATATTTCCTCCACATCTCAG
                   LINE ->                                rs73249629
154640770  TAAATGGAACAACTTTTCTTCATCCACACAGTCCCTACAATGATTCAGGCCACAGTCATC
154640830  TCTGACTTGACTTATTGACAGATTCTATTGCAGTTTATGTGTTTATATAACTGTATTTCT
                                   rs9650962
154640890  CTTCAGTATTAGGCCATTTTGGCTAGCCCATCATTATTTAACCCCTCTGTTCTTCTATCC
                                    rs73249630 rs73249631  LINE ->
154640950  ATTTGTTTTAGCAGACTGTGAGTAATGAGAGGATGAAACGATTCATTTATTCGTTCAACC
154641010  AATATGTATTGATAACCTACTACATGGCATACCATGTCACATTCATTTCCGTATCTCACA
            rs6655070
154641070  TATCAAGTATATTGCCTTACCTTAAGTATGTACTCCATATATGTTTGTTTTATATTCAAA
                   SINE MIR ->
```

```
154641130  CTTACTGTGTCACAGATAAGGAAAATAGGGCCTAGAGATGAAGAATGACTTGCTAAAGCC
154641190  ACACAGCAGGTAAGTCCAGTTTTCTCTACATAGAACCACTTTGTCCCTCCCTCACACACT
154641250  TCCATTATTGGTAACATTGCTCCTAAACTATTTGTCAGCTTCTGTGGCAAAGAAATAATT
154641310  ATCTCCATCCCCATAGGCTTTTCAGTTAAGAATCTATGTTCCCTTGGCATTGCTTACAAT
154641370  GCTGGTGTCACCACTGCTGAATCAGGATGTCCACAGATCAGCTTTTTATCTGGCACTTCT
154641430  AAAGAGCTCCTGCTGACTTCAATAGAATATGGATAGAGCTGGCCTCCTATAGGAACTTTC
                          S18F ->
154641490  AGGCTCTCAAATGGATTTGGGATGTGTCCTGGGAGAAAGAGAATGAGGAAAGACTTTCAG
                              FAS5071C/G ->   rs6655071C/G
154641550  AATTTGTGTGGAAGCATGTGTTTTATAAAAACAAAACATGACAAAATGAGTTATAAGAGA
                                <- S17R
154641610  GAATTGATTGGGTTTATGGATGAAGTTCCTGATAATGATTACTTTTGACACTATATTGGA
154641670  CTATCACATAAAGTTGGTTCTTTATGTCTAATCTATTTCCACTTGCTACAATTTAGTAAA
            rs59817403C/T
154641730  AACCTAAATAAGTTTAGTACATTCTAGGAATATAAATAAGTCATGCAATGTGTAGCCTTC
                          FAS5757C/T ->   rs57915757C/T
154641790  TAGAGAAAATTTCTGCTTCTAATTTGCTCAAGCTGAGTGTAAGGGACCCTCTTAGAATGG
154641850  TTTAGAAACATGTTGTTTTGACATCTTTATATCTTTGTGCATGCCGTTTCTCTGCTCTCA
                         AluSx SINE ->
154641910  ATGGATCTTATTAATTCTTTTTTTTTTTTTTGACTCTTACACTGTTGCCCAGGCTGGAGT
154641970  GCAGTGGCACAATCTCACTGCAACCTCTGCCTCCCGAGTAGCTGGGATTACAAGCGCCCA
                                                        rs58248838A/G
154642030  CCATCACGCCCGGCTAATTTTTGTATTTTTAGTAGAGACTGGGTTTTGCCATGTTGGCCA
154642090  GGCTGGTCTCAAACTCCTGACCTCAAGTGATCCACCCACCTCGACCTCCCAAAATGCTGG
154642150  GATTACAGGCATAAACCATCACACCTGGCCAAATTTATATTTAAATATCACTTTTTCTTG
                              <- S19R
154642210  TAAGCATTTTATAATCTCCTCAATTACTGCATTAGGTATATTTCCTCTCTGCTCCTTTCC
154642270  AAAACACATGTCACCCTTTGTTACATTTGTCTATTGGCTAGCTTTCCTACTAAATTATGA
154642330  GCCCCTTGAAAACAGGGATTCATTTATTAATTATATATTAACTAATAACATATCGCTATA
                                              LINE -> rs28499570C/T
154642390  TATATATCTATATTAATAACATGTAACAAATTCATTTATTATTGTATCTTAATTTATTAC
154642450  TGTATCACCAGAGCCTGGTACAAAGCTGTCCCTCAATATATGTTTGCTGATTGTTGAATG
               <- S13R
154642510  AATCAAGGAGCTCCTAATTTAGTAAAGGAAATAGGCATGGAATCATGTAAATTATATTAC
154642570  CATATGATGACGGCAAAAATAGAAGTATGTTTACCAGGTGAATTTATTTACTAATTCCTT
                                         2.3F ->
154642630  TATTCCTCCATTCAACACACACAGAGGACTTCTGTCCCAAGTCTATTGGGAAACACAGAG
           rs58719765-/TAAG       2.4F ->                         LINE ->
154642690  AAGTAAGTAAGCCATTTAAATATAGTGCAGTACATGCTCTGACAACAGTTAAGTACATCC
154642750  TCTATGACTCAATTGAAAAAAATACTTCCTCCAGGAAGCCTTCCTTGACTTCTTCAGATA
154642810  GAGTTTGATTTCTCTTGTCTTTATCTCTCATCTGTATCCCTAAACTTTATGTGCTTTCTT
154642870  CAAATATAATGAATATTATGTTCTACATCTATTGTAATCACTGGTTTTAATGTGTTTGAC
154642930  TTTTCCACAAGCCAATGAGCCCCTTGAGGGCAGAGCCATTTTGATGCACACTGGAATAGA
              <- S20R                 MIRb SINE ->       rs28533516C/T
```

```
154642990   GTAGTAATCTCTCAGAAAATAATAGTAATGCCTACCATTTGTTGAGTGCCTATATTACTT
154643050   ATTACTTACATTATCTCAAGTCATTCTGTTTTCAACAAAGACTTTATTGTCCCTATTTTA
154643110   CAAATGAAGAAACTGAGGCTCAGAGGCTTAAGTAAGTTAACCAAGTATCACAAAACTAAT
154643170   TTATGGAACTGGTTGTAAGCCCAGGTCTGACTTCCAAGGCCTTAGCATAAGGATTGGCTT
154643230   AGCTTTGGAATTGTTTTCTATAGTTACTATTTATTTGAACCTTGAAGTACATTAAACTTA
154643290   GGGGGACTCAGTTGGCAGATTCTTAACCAGGTTTCAGGTGCTTAGATCCTTTTTCAGCAG

                   FAS0571C/G ->   rs5940571C/G
154643350   AGATAATTCCAGGTAAAACTGTAACAGCGACAAAGTTTGTGTCTTTCTTTGACTCTAAAG
            MER1B ->
154643410   CAGAGATCCCTAAACTTTTTGGCACCAGGGACCGGTTTTGTGGAAGACAATTTTTCCACC
154643470   ACGGTGGTGGTAGGAGGGGATGGTTTGGAGATGAAACTTTTCCGCCTCAGATTGTCAGGC
154643530   ATTAGTTAGATTCTCATAAGGAGCATGCAGCTTAGACCCCTCTCCTGTGCAGCTCACAAT
             S11 ->
154643590   AGGGTTCACGCTCCTATGAAAATCTAATGCCACTGCTGATTTGACAGGAGGTGGAGCTCA
                                                               <- S12R
154643650   GGTAGTCATGCTCCCTCAGGGGCCACTCACCTCTTGCTGTGTGGCCTGGTTCCTAACAGG
154643710   CCGCCGACCCCTACCGGTCCATGGTCTGCGGGCTGGGGACCCATGCTTTAAAGTACATGC
                                                        rs4893816C/G
154643770   CATGATTTGGTCCAAAATGTCAATCCTTATACATTTTCCTGGAATTTGTCTCTAATATTG
154643830   AATGCTTTTAAACTTCCTTCTTCATTTATTTTTTCCCATTGACTCCTTCCTTTCACTTGA
154643890   AAATCCACTGGGTTATCCTGTGTTGAAATCTTATAACTCTGCCATGATTTTTTAGTTTCC
154643950   TCTCAACATACTGAGATTTGACATTTTACCCCACCACCAAAACCATTTTCAATAAGTCAC
154644010   CAATACCCTATGGGGCTAAATTCATTGAATACTTTAAATTACCTATCCTAATTTCTCTTC
                     S22F ->
154644070   ACTATTAGCCCATTTTGACTACCCCATCTTTATTGAAACTCTCTCTGCTCTTCTGTGACA
                <- S21R
154644130   ATGTGTATCCTGATTCCTTTCTAACATCTGCTTACTCCTTCTCTTGTTTCTTCACTGATT
                                                LINE ->
154644190   CTTCTTATTTTAATTTCAATTTCATTTTAGATTCAAAAAAAACAGAATTTGTTCTTTTTC
154644250   ATGGCTGCATAGTATTCCATGGTGTATATGTGCCTTATTTTCTTTATCCAGTCTACCACT
154644310   GATGGGCACCTAGGTTGATTCCATGTCTTTGCTATTGTGAATAGTACTGTGATGAACAGG
154644370   CACATGCATGTGTCTTTTTGGTAGAACTATTTATTTTCCTTTGGTTATATACCCAATAGT
154644430   GGGATTGCTGGGTCGAATGGTAGTTCTATTTTAAGCTTTTTGAGAAATCTCCAAACTGCT
            rs5940572A/T
154644490   TTCCAAAGTAGTTGAACTAATTTACATTCTCAACAACAGTGCATAAGTGTTCCCTTTTCT
                                          rs59391494-/T (bad seq)
154644550   CTGCAGCCTTGCCAGCACCTATTTTTTTTTTTTTTTTACTTTTTAATAGTAGCCATTCTG
154644610   ACTGGTGTGAGATGGTATCTCATTGTGGTTTTGATTTGCATTTTTCTGATTAGTGATATG
                                                                64.46F ->
154644670   GAGCATTTTTTCATATGTTTGTTGGCTATATGATTCTTCTTTATCCTTTTGCTCTCTAAC
                         rs4617325G/T             64.47F ->
154644730   TGTTCCACATGGTAACACAGTCCTTTTCACTCTCCCTTTCCTCTACCTCAGCAATTTCAT
                       2.1F ->
```

```
154644790  CCATTTTCAAAACTCCAATGATCTCTTTCGTGTTAACAAGCCTTAAAATTCCTCAAGTTC
154644850  CTCTCCTGTCTTGTCACTTAAATTCTAGTACCATATCTCTAACTTTGTACCAAATATTCC
154644910  TACTTGAAAATGTCTCATCTTTACCTCAAATTCAACACTTCTAAAGAATAACTCAAAGAT
                             SINE ->
154644970  CTCATTATCTTCCCTATTTATGGTATAGTGAAAGGAATACTAGACCAAGTGTCAGGAGAC
154645030  TTGGATTCTACTCCTGGCTCTGCCACTAACTTGGTATTTCCTGGTTTTTCTCAACAAGAG
                                        2.2F/R ->
154645090  TTTATCTTATGAAAATCACAGTAAACAGTACTTTGAAGGGGTCATATATGAATCTTCAAG
154645150  GAGCTAATTCTAGTTGGAAAGATAAGACATATACATAAGTTACTACAGTACAAATTAAAA
                    rs4893815C/T ->
154645210  AGTGATAAATACCGTAAGAGAAGTATAACTACATTGTCATGAGAGTTCAGAGGTATTTCT
154645270  GGATGATGAGACAATGAATTATATTGTAGAGAAAATAATATTTGATGTAGTCTTTGGACT
154645330  ATTTGGAGGCAGAAACAATAATACCTGATTCATTTATGAGTGGAGGCTAACAGAGAGAAA
154645390  GGTTAAATAGATGCTAAAGCTTCAAACTTGTGTAACTAAAGCATTTGTGATGTCATCTCA
                                                       rs28669000G/T
154645450  AATAAAAGAGAAATCAGGATGAGCAGGTAGCTTGTGGTTTTGAATACAATTTTATAATTG
           LINE ->
154645510  AGGTTGGTCTCACACCAGTCCGAATGGCTATTATTAAGAAGTCAAAAATCAACAGATGTT
154645570  GGCAAGGCTGCAGATAAAAGAGTATGCTTATACACTGTTGGTGGGAATATAAATATAAAT
154645630  TAGTTCAACCACTGTAGAAAGCAGTTTGGAGGTTTCTCAAAAAAAACTTAAAACAGAACC
154645690  ACCATTCAACCCAGCAATCCTAATATTGGGTATATATCCAAAGGAAAATAAATCATTCTA
          rs73249635A/G
154645750  CCAGAAAGACACATGCACTCATATGTTCTTTGCAGCACTATTCACAATGGCAAAGACATG
154645810  GAATCAATCAACCTGGGTGCCCATCAGCAGTGGATTGGATAAAGAAAATATGGTACACCA
154645870  TGGAATACTATGCAGCCACAAAAAAGAACAAAATCATGTTCTTCGCTGCAACATGGATGC
                                                     rs61499113-/A
154645930  AGCTGGAGGCCATTTTCCTAAGTGAAATAATACAGGAACAACAAAAAAAAACCCCTCATG
                                    <- 3.2R
154645990  TTCTCACTTATAAGTGGGAGCTTAACATTGGGTACTCAGGGACATAAAGATGGCAATTGG
154646050  GTAGCCTGCTTACTACCTGGGTGACAGGATTAATCATACCCCAGCCAAAGTTCTTTACAA
154646110  CTCTAATATCATATATTTATTCCTTTAAACTGACCTCCCCCTTCCAACTTCTCAGTATGC
                                                  <- S7
154646170  AATGGTATCATATTCAACTGAAAACATCTCATCCTTACTTCCCATTCTCACTGTCAATTC
                                                                <- 3.1R
154646230  CCTAACCAGGCCCTTATCATCTTCTGTTTAGATTACTCCAAAATACCCCAATTTGGCTGA
154646290  CCTGACTCTAGTATCTCCCCACTCCTATATATCATGCCCAGTGTATCTTTCTTAAATGTG
                              rs34104985G/T
154646350  GCTTATTTTCATATTTTATTCCTTTGTTCTAAAATTTATTTCTAATTCCCAGAGCATTAA
154646410  GTCTAAATTGCCTGTCATTCGAGTTATTCTCCAGCACTATATATTCCAATCCTCTCTTCA
154646470  ATATACCTCTTACATAAGACAGTTTGAATGTTTCACTCTTTGCAGACTCCATGCACATTG
154646530  CTACCAATCCAGATTTCCATGTGAATTCCTTCATGTGAATTATATATGGTTATTCATTTT
154646590  CTGCCCAGATCTCAATTCCTCTTTGAATCCTTACCTATCTAATTTATTTCAACATTTGTA
                                             rs28666767G/T
```

```
154646650   ACTCCTTCATCACTTTCACCCATGCCCTAGTCTCTTTTTCAGTATGCTCTGTTAGATATA
154646710   CTCTCCCTTCTTTCAATATGTTTTTGAGTTAACTGTCATATCAGAAATGATCTGAACAAC
154646770   CGGTTACTACAAAGCATGGAAAGGTGTATTTTATTCATTACAAAATCATAACTCAGAGGT
154646830   TTCTAAAACAGACTTAGTAAAGAAGATATTTCCAAGGCAGGGTGACTTTTCTGGTGTCCT
154646890   TCAAGGTTACTTATTTATTTATTGAAAATCAATGGCTTTCAGAAATAGATGTGTTGATCT
                                                         SINE ->
154646950   GTGCTGAGACAATATATCCAAATTCAACTGGAATTGAATTAATGTCAGCAAGGTGCCTTA
             rs28595824A/G
154647010   AAAAGAGAAAGAAGGCATTGGATTGGGAAACAGGAGACCTGTGTTCACTACTAGGCCCAA
154647070   TGCTGCCACTTACTAGCTGTATGATCTTGGACAACTCAATGAATCTGTCAAAATATGTTT
                    <- S9
154647130   TCTGAAGAGTACACCATTAACTAAGGAGCTAAGAAAACAAATTTTCTGAAGTGCACTGCT
154647190   ACAACAATCTAACAAGTTTGTAAAAGAAGTTGTCTGGCAATACATAGTGAGTTTTGTAAT
154647250   GTCTCAACATCAAATTTCTGTATATAATTGTACTATAGTGATAGCCTATCAAAAATCACG
                                                             <- S10
154647310   GAACTTCGAGTTGTCTTTTAGAAACAAAACAGCTTAAAGAACAAATAAAGATATTCATGT
                               LINE ->
154647370   AGATTTCCAAAAATTGGTTATCCATTCATTCAATAAGTATTAATTTAGTGATCAGTATGC
            rs6642319C/G
154647430   TAGGCCTTGGGGATGTGGCAGTGAAGACAAATATAATAAATATTATATCTGCCCTCAAAG
154647490   AGATCACTGTCTAAATTGGGAGACAGACAATTATAATAAAGTGTGATAAGTTCTGCAATG
154647550   ATGGTAAGCACAAAGGGGGCCTCTGGAGCACAAAGAAGAGGACAAATCTGTTCTAAACTG
154647610   GTCTTTTGAAGAGAGAGGTGATTAGGCAAGGCTTTTCCTGCAATAAGTGACTGCTAGGCC
                                LTR ->
154647670   TATAGGATATTCAGGAATATTATCTTAGTCTTATTGGGCTACTGTAACAAAATACCTTTG
154647730   ACTAGATAACTTATAAACAACAGAAATGTATTTCTTACAGTTGTAAAGGCTGGGAAGTCC
154647790   AAGATTAAGGTGTTACCAGTGTCTGGGGGAGTCCGACTTCCTCATAGACAGCACCTTCTC
154647850   ACTGTGTCCTCACATGATGGAAGTGCCAAGGCAGGTCTTCGGGGCCTCTTTTATAAGAGA
154647910   ACTAATCCCATTCATGAGGGCTCCACATACATTATTTAGTCACCTCCCAAAGGCCCCACC
154647970   TCCTAATATTATCATATTGGTAGATAGGTTTCAACACAGGAATTTTGGAGAGACACAAAT
                                                    S22F ->
154648030   ATTCAGACCATAGCATTTACCAAATAATTTTTAAAAGTTAGAACAAAAATCCCTGAAGGG
             <- S23R         SINE ->                         rs55997085 +/-
154648090   ATTTGCAGATGGTTGGATACTGTGGTATAGTGGTTAAAAGTACAAGCTCTGGAGTGAGAC
                                             rs68008496C/T
154648150   AGCCTGAGTCAGCCTGAGTTAAAATCCTAGATCTGCAAACTGCCAACTGTGTAACCTTGG
154648210   ACAAGTTACTTAAGGTCTTTGGACCTTGGTTTCTCATTTTAAAAATCAGTATAATTCATT
        rs67804731A/G
154648270   AATATGCCTCATAGGTTTGTTGTGAGAATTAAATAGGTTAAAACATGTAAAATTCATAGA
                             LINE ->
154648330   ACAGCACCTGGCAAAGAGTAAACATGTCACCCTGCCCACTTGAATGTCAATTCCAGGAGC
154648390   GCAGGAAATCTAGCTGTTGAATTACTAGCCTTTAGTGAGTGGCCCATAGCAGGCATGGTT
                                                           rs17149475A/G
```

```
154648450  GAATGAATGAATGATCATCAATTTAACTTTTTTTGTCATGGGTATCATCCTAAGAGGTAA
                                                      64.85F ->
154648510  GAAGGCTGAATAATCCCAATTGTTTTCATTGTTGCCTTAATATAGGCCCACAGAGTAAAG
154648570  GAGGGAAAGGAATTAGGCAACCCAGCTATTCAACTTTAAAGCTAGACTTTAAAACATAAA
                                      64.86F ->
154648630  AGGGCTTCAAATTGAAATCATCTCTTGCCTCTCCTATGGGTTACATTAGAGCAAGAACAA
                          SINE ->
154648690  TTTCTTAACATTCATAGCTACTATTCATGGGGTGCTTGCTATATGCCAGAGAGGTTAATT
154648750  GAATTGCCTCAAGGGAGGTAGCTCCCAGGTTGAGGAGCAAAAATTTGAACCCTGATCTAT
154648810  TTGACCTGAATCTAGTTAGTACAGTGCTTGACACATAGTTAATGTGCAATAAATATTTGT
                        <- S8         seq1 ->
154648870  TGAATAAGTGTAGATGAACCTACAGTGAATCACAAAAGAGAATTTACAGGAGATTTCCTG
                                   <- S24              Simple repeat ->
154648930  ACTAGAAATTCCAAATGGGAAAAGAGAGCTGGATTTGTTTAGAAATTACACACACGCACA
154648990  TACACAGGCATGCACGCATGCACACACACATATACCTACATACATGTCATCACAACTGTT
154649050  TGGAGCACCTTAAACACATACATACACACAGGAGCATGCAGACCCACACAGCCGTACTTT


                                                            SINE ->
154649110  TTTAAATCTCATTATTTCTTTTGTCTTCTTAACACATCTGAGAGAGGGTAAATGACAGAG
154649170  TGTCAGAGGTAGAAAGGACGTTAGCAATTACCTAATCAGATCCCCTCATGTTAATGTAAA
154649230  AATTGACACTCTAGAGGGGAAGAAAATTGCTAAGGTCACACAAGGAGTCAGAATAAGATT
154649290  AAGAGTCTCAGTTTCCTGACTCACACAGTTTGTTTTACTCTGAACTGTCACCATTTACCA
154649350  TCTCAATTTTATAGATGAGGTAACCAAAGCATCCTAAGAATGCTGATGGCCAAGCTCCTG
154649410  CCACTTGTCCATTGAAATAGGACTTGAAATCAGATCTCCTTTATTTTAAATATAAAGCTG
                                                               S1 ->
154649470  CTGTCATCCAAGTTCTTCTGTCTGGATGCCCAGAAGAAAGATCCTTTCCTATTGAATTAA
154649530  ATAGTGGACTGAGGGTAATTTACAGAAAGCTTTGTCCTTGGCTTTAAGAGAAAATGTCCA
                                   rs17149476C/G            SINE ->
154649590  AGGCTGCTTTATTTTCAGTCCCAGGGAGGCTGCATGAAGGCAGGAATCCACAACACCTAG
                                    rs17149478C/T
154649650  CACAAAGGAAGGTACTCAATAAATATTTGTTGTTGAATGCATGCCATGCTAAAGGGAAAG


               sequencing rs61115275A/C
154649710  TAGGTTGGGTAATGGTTTCCCAGTGGCAATGGATGGGCTGAAAAAGTGGGAACACATGAG
                      G (d274GT)
                                           sequencing sequencing
154649770  GAAGAGGTGGGAGGAACCCTTAAACAAAGGAGAAGCTCTTTCTCTTCATCTTATTGAAGC
                                           (d274GT) GC (R12CG, R68AA)
                                  64.98F ->                sequencing
154649830  TGCAATCACTCGGCACAGGCTGAGTAGATTGGTAGAAATGTGATTGGCAAAGATATTTAA
                                                               G (R29AA)
          rs28535456C/G  FAS0753C/G ->  rs56750753C/G
154649890  ATGAGACCTCCTTAAAATTGTTACTCTCCACCCCACCCCCACGTTCCTTTAAATTTTGTT
```

```
154649950  TCAGTCATTACCCAGGGACCGGATACAGAGATTCCTATGGGGAGTTTTAATGGGTCGATT
                <- 65.00R                4.1F ->             rs306882A/C
154650010  CAATTTCATCTGGATGACGCCACTTCGCCAAGCATTAAGAGCTACAGCTCCGGGAAAGCC

                                sequencing      sequencing
154650070  AACGACACGCGGGGGGAGGGGGGAGAGAAAGAAATTATAACGAGGAGCAATAAATCCCTT
                            (R29)-  G(R12,R24,R26,  AG(d5,R29,R68,R75,WI20)
                                      R29

154650130  CTTTCCCATCCTCCCCTATAACCCATTCACAAGACCTCCAGAAATCATCTACATTCACAA
                                            main Myers motif SNP
                                                rs700442C/T
154650190  AACGGCCAGCTAGCTAAGCTGAACACACTGCCAACTCACTCCCTGCCCCCACTGCAGATT
                                                 ||||||  || |
                                                CCTCCCTNNCCAC
          sequencing                                   rs306883A/C(G in R20)
154650250  ATA-TACCAACACACCCTGAGCCATATAAATAATCACACACTGGCGGTATCTATTGCTCCG
              A (d252GT)
                                     sequencing     4.2F ->      STR ->
154650310  TTGAGATACCCTGTGTCTACAACTGCAGTTTCTGCTGCTTCGGTTCATCTGTCAGATTGG
                                         C (R26)

154650370  AGGAGAGGGAGACCAAGGTGGAGGCGGAGGCGGAGGCGAAAGAGGAGGGGGAGGAGGTAA
154650430  AGGAGGAAGAAGGGGAGGAGGGAAAGGGGAGGGCAAGAGGAGGGGAAGGAAAATACTGGA
154650490  GGGGGAGGGGGAAGAGAAACAGGAGGAGGAGGAGAAGGGGGAGGAGAAAGAGGCGGAGGA
                               rs306884A/G (bad seq)
154650550  GGAGGAGAAAGAAGAGGAGGAGGAGAAAGAGGAGGAGGAGGAGAGAGAGGAGGAGGAGGA
                                                     sequencing
                                                      S6 <-
154650610  GAGAGAGGAGGAGGAGGAGAAAGAGGAGGAGGTGAACAACTTACCCTGCTGAGCTTTCTT
                                                        A (d244GT)
154650670  TGGGAAATACGTCCATCAAGATTTAGATCTGCCTGTAAAATCTATACAAAGTATATGCCA
                                          STR ->
154650730  CTACAGGTTTGACTCGCCCCCTCCCCCGTTTTTTTGTTTTGTTTTGTTTTGTTTTGTTTT
154650790  GTTTTGTGTTTTCTCTGCTGTGTCAAAGAACAAGACAGAACTATCTCTGTTTCTGGCTCC
154650850  ACTGCCTGCCAGTGAAGGAGTTTTCATTCAGACTTTCCGAAGAGAGGTGGAGAAACCTAA
                                              rs9657862C/G
154650910  AGACTGAGGAGAAGAGATCCTTTGAGCCAGATGGGGCATTAGTTCTTCTGCTTTTCTCAG
154650970  CATGGATAAACCATTTCCTCAAGGTAAGCCATAGAAATTAGCTCTTTAAAAACCCAGAAT
                         5.2R <-
154651030  TCTTTCTATGCAATGCACAGATTGCCATTCATTCCAGCCATCCTGTGCTGTCTCTGTGTG
       rs1736451C/T rs1736452C/T
```

```
154651090    CGTGTGTGTGCCGCGCGCTTGGGCGCGCGCGCGCAAGCGCATGGATAAAATAAAATGAAA
                 AT rich ->                                S14 ->
154651150    ACCCTTTAAATTATATTTAAATAAATCGGCTCTCGCAGAAAACTATCCACATTGGAAATG
154651210    TGTAAATCCAAATAGGACTTCAAAATAATATTTTTCTAGGCGAATGTCAATTTAATTTCT
                                                          <- 5.1R
154651270    AGCCTGTTAACCTTTAAATGCATTTTGGTACTTGCCTAAACCCCTCAAAACGCAGCCTAA
154651330    CCACCCAAGCTGAGGACAGAGAGAGAGCCCGTGCTAAGGCCAGTCTGGCTGCCCAGCGGG
154651390    CATCCTGTCAGCATGGGGCTAGCCTCTGCTATTGGCTCACAAGTCTGGGCTGTGCATCGT
                                                              rs306885G/T
154651450    TCAGATTTCATCACATCACCTTGCCCTGGGTAGACATGGCGCAAGCTGACACGGCAGAAC
154651510    CCCCCACTTACATTGAGTCAGAGCAGTGATTTAGAAATCTCTGTATCCGTGCCTGGGTAC
              rs3870343C/G                 <- S5
154651570    GTGTGCACATGTCTCTGTGTGGCATATGGAGATTCTGTCAAAAGTGAGAACCTGTGAGCA
             rs306886G/T
154651630    CGTCTAGTTAGCCAATAGGCATTTGATAAGACCTAGGGCAATGGTTTACTGCACACTCAC
154651690    TTCCATGAAGCTAATTATAATTATCATTCTTCACCTCTGCATGCTGCAAACCAAGTTGGA
154651750    TGAACAGAAAACAGACCTTGAACTCTTAAGCAGAAAATCTATTTTGAAATGGAAACTATT
154651810    GTTGCTTTTCTGGTAATATTAAAACCTCACTGTATCATGATGTGTTATCACATGAATTTG
154651870    GATATACCAGGTGTTAAATCATGTTCCCAAAAAGTCGGCTGCATTCATAAATACAAATAA
154651930    GGCACAGGCTAGCCGTTTATCCACAATGAAATATTGGTCTTGCTCTCCTTGCCTCTAAAC
154651990    AAACATATACAAAGACATTACTATAACATTCTATATATTTGTCTCTTTGTGTTCTGTTTT
154652050    TAGGTCTCTCTAGCTGTGAAACAGTCTAGATAGTGATCTGAGTGACTTAGATTTGAAGAT
                          65.21F
154652110    ATTTATTAAGAGTTGAGGAGTTGAGTAGTGAGGAAGTAAAGAACCTAAAGGAAAGGACAC
                                                                 SINE ->
154652170    TAATTCTTCCTTAAAGACTTTTGAAACAGAGAAATGCAACCAACAAACCCAACATAAAAT
154652230    TTAGAGCATCTGGACAAGAAAGGCCTTTAAAAACCCTCCTTTTATTTTGCAGATGAAGAA
                                             rs6642320A/G
154652290    ACTGAGGACCAGAACGGAGGTCATTTGCCTACAGCCACGCAGCAAAGTAGAGCTACCATT
                                          rs34648235-/A and rs71278022-/A
154652350    AGGACTAGACCTTAAATTTCTTGTGTCATATAAAAGGATTTGAGTTTTTAAAAAAAAAAT
             S3 ->
154652410    CCTACTTCTTAGATGAACAACCAGCTTCTTCTTACAATTTTCCTCAGGATTAGATGTGAA
154652470    TTCATTTGGTATTACAAATTTTCCCTGTATCATTCATCACTAAAAGACCTATATCTTTTA
                                              <- S4
154652530    ACATTCAAGGTCTTAAAACTGCAAGACTCCCTGGTTTGAAATAAACTGCTATGAGTCTTA
154652590    ATGAAACCATTAAAGTAATGAATAGGACTAGGGTAGCAAAGTGACAAGTTAAGAGTTAAA
                                                                  <- 7.1R
154652650    TGTGGTGAGAACCTGGCAGCTCCAACAAGATGTATGCGGTAGGAGGGAGGCCCAGTCAGT
154652710    GGCAGAGAGGTTAGTAATGTAAACTATATACAACTTGGTATTTGTAAGGGGGGAGTGCTT
154652770    TGATAGAGATGCAAATGCTTTTTGCACAGAGAGGGATTTTTAGAGAGCAGACTTTGGAAC
154652830    AGTAGAACTAAGAAAACATATAAAGCCAAAGCAGAAATTCCCCCTATGTTTGCTTTCAGT
              rs17653343A/T   <- RAS3343A/T
```

```
154652890   TATTCTGAGTAGGAGTGTATATAACAGCAGAGCATAGAAAAAAGCAACACACCTAAGAGC
154652950   CGTTACACAGTAGGGTACAAAACAGCATTGAACCCGTGAGACTGGGAGAAAACAAGAGCC
154653010   AGGATCTCCTCTTCCTCACTCCACTCTTCTTTGAAATCAACGTTGAGTCATACTGTATCG
154653070   GAAAGGATTTACCATAGAAATCAGAACTGTCTTGCGGTGTTAGCATAGAAGGGCTTTAGA
154653130   TAGGAAAGTTTTCAACTTTATTCTCAACATGGCAGATTTTAGGTTCCAGGGAGAAAATAC
                                                   S15 ->
154653190   ATTTTCTTTCATTAACAAACATTAGCATTTCTTGAATGGGCTTAGCTTTCTAGTCCCTAA
154653250   ATTGGCATAAGAAAAGAAATTTTTTTCAGGTTGAACATGGCAAATTCAGGGCTACAGAAA
154653310   AACTGTTTAGAAAGCTATGAGCACCTGGCCAGAGAGACTGTGATGGGAGTGGGTTTACAA
                                 rs73638004A/G
154653370   TGCATCATTCATTACTGCAGAGGCACTACGTGACAAAAATCACTAGATCCCATCAATGTG
                                                                   STR ->
154653430   CAGCCAAGACCTGTAATCACCTTTGCTCTTATTTAATAATATGTGATGTGTACATTCTCT
154653490   CTCTCTCTCTCTCTCTCTCTCTCTCATTTGATAGACTATGGAATTGGCCAACTGCACAGTTT
154653550   CTTAACTGTTGCTGCTCACAATGAATGGCTCTTTAGTGGGTGGATTTCTGAATCATAAAC
154653610   ATTCACGTAGTGGTACTTTTTCATCTAGCCTATGTGCTACAGCAGTGTCTTGCCTGTTCC
               rs9645286G/A   <- RAS5286A/G
154653670   TTTCTCAACATATTTTTCCCTCTTACCTTCCTGTACAGTATCTATTTCAGTTCATCCTGT
              <- S2
154653730   TAGCCAGATCAGATTCAGCATGTCCAGAATGGCAAACAATTTATTGTCATTCAATCCTTA
154653790   AATCCTACTCAACACTCTGATTTTATTCAGGTCCTACTGTCCACTTACTGCAGGTCTCCC
154653850   ACATTAGATCCCAGAGTCCATTTGAATAAAAGATTCTACAGAGCAGGAAGAAACCTTCTA
154653910   AGTTCATCCTGTCAGTTCCCTTACTTGTAGGCAAGACAGTGCATCATTATTTCAGACTGA
154653970   AATAATGCCCACAGCTTCTTTTAATCATCCAACCAAGCATCTCAATTTATCAGTCTTGCA
154654030   GAAATCGAATCAATATTGAAATACAGCTTTTCAACCTAAACACCATGTACCCTTCTGCTC
154654090   CCCAGACAGGATCCCCTTTGAACATAATTCTCATTATTGTTTCTGAGTCATTTATATTAT
154654150   GACCTAGATTCCTCCCAGATTAGCTTTCAACCTTCTGGGCATGTCCCTGAGGTTCCTAGT
154654210   GACATCCCTAGAATCCTGTCTTTGGAATCTCACCTGTTTATCTCCTGTAACTTTGAGACC
154654270   TTGTATAACTTCCTTTAAATTATTGCTTCCTGGAATTTTCAATATGTTTTGAATCTAGTT
154654330   TCCTTTGGGAAATCAATTAGTCTAAACTGCATTCAGCATGTATATTATTGTTCTTACAGT
                          rs306887C/T  <- RAS6887C/T
154654390   ACAAAGGACTTGCAGGGCTCCACATATTCTTTAGCCGATCAATGGGACATAAATCTATTT
                                  <- S16              rs28630004C/T
154654450   CTCATAAATTTGTCAAAAGACTCCTCCCAGTCCCAAGTTCCTACATCTACATACCCTCCA
154654510   AGCCAAATGAACAATGTGTCCTTTTTTATTTGGGTGTCTACATGAAAGCTGCCAGTTCCT
              <- get seq2
154654570   TTCCAATGCTGAGACTATGAATCTATAAAGATACAGCAAAGGTAGAGGTGGCAGGTTGCA
                                                         rs700441C/T
154654630   TGAGCCTTTGAGTGCAAGTAAATAAAGGAAACATACATTTTGCTTTTGAATGCCCCAATA
154654690   AGTAAGGGCACATTGCCCACCTTACCCTCAAGGACTTGAAGATGTAGAAGATACGGATGA
                                                                    SINE ->
154654750   AAAGAAATATTTTATAAAATTGTGTAGTTATAGAAAGAAATACAGACCATAATGGCTCAA
154654810   GGTCCTTCAAGATCACCTTGTTCACATCACATATCCTTCAGTTTCACATAGATAGTGAAA
```

```
154654870   CTGAGGCCCAAAGAAAGCAAGTGATTTGACAAAAGTGACATAAATATTTGAGTCTGAGAG
154654930   TTTCCATAATATTTATTTGCATTCTCTTATGCTGCATTTATTTAACACCAAAACAACTAG
154654990   TTAAAAAGTGAAATTTTTGTCAGACTGGTTCTTTGATGTGATATGCCAAAAAGGTTATTT
154655050   GTAACCAAATTAAATCTTTTCAAGTAATGAGCAGATGTAAATTCAGCTTATCAGGTTTGC
154655110   ATGAGGATATAGCCTCAAAAGTCAACTGAGTCTGGTCATATTATTAGCCACTGAGTCCTT
154655170   TATGAACTGCATCCTTGTTTTTCCTGTTATTGAGCTTTGGGCCCATTTCAGTGTCATGTT
154655230   TGCTTGCTTATGGAAACGTAATCAAGCACCATCCTCTCTCACTTGTTCATTTCTATTACT
154655290   CATTCAAGCTATCCATCACATACAGTAGTGTGAGTAATTGATCTATTTTGAGAAGGCCTG
154655350   AAATCTGATGGATCTCAAATATATACTTTATTCATGCTAGTGTTTTGCTACGTCCTAATG
154655410   CCTCCAGGATGTTTCCTGGATTTGTGTTTAGGTTTCTTCATGTACACCTTTTTGACTCTT
                                                           rs35535695A/C
154655470   CAGTCTTCTAGGTGGTGCAATACAGCAGTCTGCTATTTGTACAGTATTTTCCCCATGGTC
154655530   ATCATTATCAGTTCTAGGCTTGGTAAAATCCTTCAGGGAATGACCAACCACCTCTTCCTG
154655590   ACTGTAGTACAATGGAAATGCAAGCATATGCTTTCTAGACATTCAACTTAGCTTAAATAT
154655650   GGTATAAGAATATCACTTTCCTTTAAGAAAGTCAGTTTATTGTGTAAGGTAGTGATAGAG
154655710   AAAAGAAGTCCAAGCAGATTCTGCTTGGTAAGTCTCATCAATATGGAGGTGACTTAGGAT
154655770   AGTAATAGGTATAATTACAGATGATTTTTGAAATAATTCAGGACTTGTCTTCTATAGTTC
154655830   CTTGATCCCCAGTATTACTCCCACGTGTACAAAAGTTTGATTCTGGAGCCTCCATTCCAT
                                           <- 6.2R
154655890   TTAAAAAGAATCACTTCCTTAGTAAGTTAAAGCAGCCAGTAAGGTGAAGAGAACTGCAAG
154655950   ATGATTAGTCTTGGTTTCCTTATACTCTTGGGAAGCCTTTTTTTTTCCTACTACACTGCA
                                                  <- 6.1R
154656010   GACAGCTTAATTCCTACCTTCTGTGTGTTTGCTGTATCTCTGCCTTTATTGCATTCTCCT
154656070   TGCTTGCCTGTATTCACTGAATCAAGTAATGTCCTTTCTTATCTGTTTATTTCTGCTTAT
                                             Simple TA ->
154656130   CACAAATTCACACTGAAAAGATAATTTTGATTGGATATATATATATATATATATATATAT
154656190   ATATATATGAGCAACTTCTCATGGGAGACTTTTGAGGCTTTCTCTTTTCCTTCAGCATTG
154656250   CTTCAGAAAATAGCCTTTGCAGTTAATATTAGGTTCCATTTGTATTAGAATTCTGATGCC
154656310   CAAGGCTTGGATGTCTCTGCTAGCAAATACCTGAATGAGCTAGCTGCTCCTGTGAAACTA
154656370   GGCATTTGATGGAATACAAAAAGAATCAACTATCTGATATTTGTTCCCAAACTTCTTTAT
154656430   AGTCTGGAGGGGATGATGAGATTTGCCTATAAAGAAGTAAGTGAGGTTAATTACAGAATA
154656490   A
```

Appendix 5.3.

a. Primer sequences and PCR conditions used for amplifying region of high historical activity in PAR2 recombinant CEPH pedigrees.

Two rounds of fully nested PCR were used to generate all amplicons. Forward primers are listed first, followed by reverse primers. Note: All primers used were optimised first using annealing and extension temperature titrations.

Amplicon 1:

Primers	PCR conditions

Primary PCRs-
`1.1F - 5'TGCAGCTTCTCTGGGTCTC 3′`
`2.2R - 5'TGACCCCTTCAAAGTACTGT 3′`
	96°C for 1min
	96°C for 20 sec
	56°C for 30 sec } x 32 cycles
	65°C for 5:30 min

Secondary PCRs-
`1.2F - 5' AGTAATTCCAACAGAGTGGTC 3′`
`2.1R - 5' CACGAAAGAGATCATTGGAG 3′`
	96°C for 1min
	96°C for 20 sec
	58°C for 30 sec } x 32 cycles
	65°C for 5:30 min

Amplicon 2:

Primers	PCR conditions

Primary PCRs-
`64.46F - 5' ACTGTTCCACATGGTAACAC 3′`
`S24 -    5' AGCTCTCTTTTCCCATTTGG 3′`
	96°C for 1min
	96°C for 20 sec
	56°C for 30 sec } x 32 cycles
	65°C for 5:30 min

Secondary PCRs-
`64.47F - 5' TCCCTTTCCTCTACCTCAGC 3′`
`S8 -     5' GTGATTCACTGTAGGTTCAT 3′`
	96°C for 1min
	96°C for 20 sec
	54°C for 30 sec } x 32 cycles
	65°C for 5 min

Amplicon 3:

Primers	PCR conditions

Primary PCRs-
`64.85F - 5' TAGGCCCACAGAGTAAAGGA 3′`
`7.1R -   5' TACTAACCTCTCTGCCACTG 3′`
	96°C for 1min
	96°C for 20 sec
	56°C for 30 sec } x 32 cycles
	65°C for 5:30 min

Secondary PCRs- 64.86F - 5' GCCTCTCCTATGGGTTACAT 3′ S4 - 5' CTCATAGCAGTTTATTTCAAAC 3′	96°C for 1min 96°C for 20 sec 58°C for 30 sec 65°C for 5:30 min x 32 cycles

Amplicon 4:

Primers	PCR conditions
Primary PCRs- 65.21F - 5' TGAGGAGTTGAGTAGTGAGG 3′ 6.1R - 5' GGAGAATGCAATAAAGGCAGA 3′	96°C for 1min 96°C for 20 sec 56°C for 30 sec 65°C for 5:30 min x 32 cycles
Secondary PCRs- S3 - 5' CCTACTTCTTAGATGAACAACC 3′ 6.2R - 5' CTTCACCTTACTGGCTGCT 3′	96°C for 1min 96°C for 20 sec 54°C for 30 sec 65°C for 5 min x 32 cycles

b. Primer sequences and PCR conditions used for determining linkage phase of semen donors

Primary PCRs:

Primers	PCR conditions
Primary PCRs- Forward primer: FAS0571C/G - 5' cCCAGGTAAAACTGTAACAGC 3′ 5' cCCAGGTAAAACTGTAACAGG 3′ Reverse primer: RAS5286A/G - 5' gggGGTAAGAGGGAAAAATAT 3′ 5' gggGGTAAGAGGGAAAAATAC 3′	96°C for 1min 96°C for 20 sec 56°C for 30 sec 62°C for 11:30 min x 28 cycles

Secondary PCRs:

Primers	PCR conditions
Primary PCRs- Forward primer: 2.5F - 5' TCCATCCCCATAGGCTTTTC 3′ Reverse primer: RAS3343A/T - 5' ggGCTGTTATATACACTCCTA 3′ 5' ggGCTGTTATATACACTCCTT 3′	96°C for 1min 96°C for 20 sec 56°C for 30 sec 62°C for 11 min x 14 cycles

c. Primer sequences and PCR conditions used in PAR2 sperm crossover assay (1)

Primary PCRs:

Primers	PCR conditions
Forward primer: 2.5F – 5' TCCATCCCCATAGGCTTTTC 3' Reverse primer: RAS5286A/G – 5' gggGGTAAGAGGGAAAAATAT 3' 5' gggGGTAAGAGGGAAAAATAC 3'	96°C for 1min 96°C for 20 sec 60°C (occasionally 58°C) for 30 sec 62°C for 13:30 min x 28 cycles

Secondary PCRs:

Primers	PCR conditions
Forward primer: FAS0571C/G – 5' cCCAGGTAAAACTGTAACAGC 3' 5' cCCAGGTAAAACTGTAACAGG 3' Reverse primer: RAS3343A/T – 5' ggGCTGTTATATACACTCCTA 3' 5' ggGCTGTTATATACACTCCTT 3'	96°C for 1min 96°C for 20 sec 60°C (occasionally 58°C) for 30 sec 62°C for 11 min x 28 cycles

Tertiary PCRs:

Primers	PCR conditions
Forward primer: S11 – 5' GGGTTCACGCTCCTATGA 3' Reverse primer: 7.1R – 5' TACTAACCTCTCTGCCACTG 3'	96°C for 1min 96°C for 20 sec 56°C for 30 sec 65°C for 10:30 min x 30 cycles

Appendix 5.4.

Sequences of ASOs used for SNP typing

SNP name	Position (ChrX)	SNP allele	ASO sequence (5' to 3')
rs73249628C/T	154640762	rs73249628C	tcctccacatctcagtaa
		rs73249628T	tcctccatatctcagtaa
rs73249629A/C	154640823	rs73249629A	aggccacagtcatctctg
		rs73249629C	aggccaccgtcatctctg
rs9650962C/T	154640919	rs9650962T	ataatgatgggctagcca
		rs9650962C	ataatgacgggctagcca
rs73249630A/G	154640984	rs73249630G	agaggatgaaacgattca
		rs73249630A	agaggataaaacgattca
rs73249631G/T	154640989	rs73249631G	atgaaacgattcatttat
		rs73249631T	atgaaactattcatttat
rs6655070A/G	154641075	rs6655070A	cacatatcaagtatattg
		rs6655070G	cacatatcaggtatattg
rs6655071G/C	154641591	rs6655071C	aaaacatcacaaaatgag
		rs6655071G	aaaacatgacaaaatgag
rs59817403C/T	154641734	rs59817403C	gtaaaaacccaaataagt
		rs59817403T	gtaaaaacctaaataagt
rs57915757C/T	154641827	rs57915757C	caagctgagcgtaaggga
		rs57915757T	caagctgagtgtaaggga
rs58248838A/G	154642082	rs58248838A	ttttgccatattggccag
		rs58248838G	ttttgccatgttggccag
rs28499570C/T	154642436	rs28499570C	attattgtaccttaattt
		rs28499570T	attattgtatcttaattt
rs58719765+/-	154642700	rs58719765+	agaagtaagtaagccatt
		rs58719765-	cagagaagtaagccattt
rs28533516C/T	154643038	rs28533516C	tgttgagtgcctatatta
		rs28533516T	tgttgagtgtctatatta
rs5940571G/C	154643377	rs5940571C	ctgtaacagcgacaaagt
		rs5940571G	ctgtaacagggacaaagt
rs4893816C/G	154643821	rs4893816C	ctggaatttctctctaat
		rs4893816G	ctggaatttgtctctaat
rs35976541+/-	154644141	rs35976541-	tatcctgattcctttcta

SNP name	Position (ChrX)	SNP allele	ASO sequence (5' to 3')
		rs35976541A	tatcctgaattcctttct
rs36064135+/-	154644158	rs36064135-	aacatctgcttactcctt
		rs36064135G	aacatctggcttactcct
rs5940572A/T	154644494	rs5940572A	ctgctttccaaagtagtt
		rs5940572T	ctgctttcctaagtagtt
rs4617325G/T	154644750	rs4617325T	aacacagtccttttcact
		rs4617325G	aacacaggccttttcact
rs4893815C/T	154645226	rs4893815C	taaataccgcaagagaag
		rs4893815T	taaataccgtaagagaag
rs28669000G/T	154645501	rs28669000G	aatacaattgtataattg
		rs28669000T	aatacaatttataattg
rs73249635A/G	154645758	rs73249635A	cagaaagacacatgcact
		rs73249635G	cagaaaggcacatgcact
rs34104985A/G	154646376	rs34104985G	attcctttggtctaaaat
		rs34104985T	attcctttgttctaaaat
rs28666767G/T	154646691	rs28666767G	tctttttcagtatgctct
		rs28666767T	tctttttcattatgctct
rs28595824A/G	154647019	rs28595824A	aagagaaagaaggcattg
		rs28595824G	aagagaaggaaggcattg
rs6642319C/G	154647432	rs6642319C	tatgctacgccttgggga
		rs6642319G	tatgctaggccttgggga
rs55997085+/-	154648157	rs55997085-	ggagtgagtcagcctgag
		rs55997085+	tgagacagcctgagtcag
rs68008496C/T	154648189	rs68008496C	ctgcaaactgccaactgt
		rs68008496T	ctgcaaattgccaactgt
rs67804731A/G	154648273	rs67804731A	cattaatatgcctcatag
		rs67804731G	cattaatgtgcctcatag
rs17149475A/G	154648508	rs17149475A	ctaagaggtaagaaggct
		rs17149475G	ctaagaggtgagaaggct
rs17149476C/G	154649621	rs17149476C	agggaggctccatgaagg
		rs17149476G	agggaggctgcatgaagg
rs17149478C/T	154649682	rs17149478C	tatttgttgctgaatgca
		rs17149478T	tatttgttgttgaatgca
rs61115275A/C	154649730	rs61115275A	atggtttccaagtggcaa
		rs61115275C	atggtttcccagtggcaa
rs28535456C/G	154649896	rs28535456C	aatgagacctccttaaaa

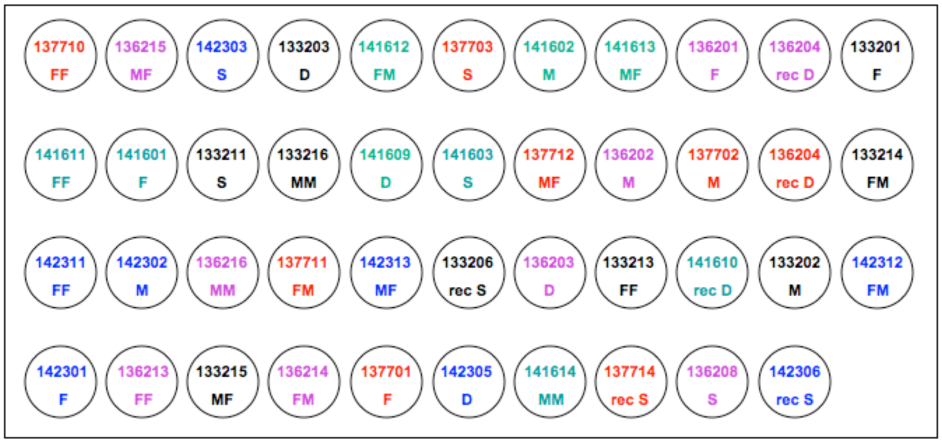
SNP name	Position (ChrX)	SNP allele	ASO sequence (5' to 3')
		rs28535456G	aatgagagctccttaaaa
rs56750753C/G	154649921	rs56750753C	actctccacccaccccc
		rs56750753G	actctccacgccacccccc
rs306882A/C	154650069	rs306882A	cgggaaagcaaacgacac
		rs306882C	cgggaaagccaacgacac
rs700442C/T	154650233	rs700442C	ctcactccccgcccccac
		rs700442T	ctcactccctgcccccac
rs306883A/C	154650301	rs306883A	ggtatctattgctccgtt
		rs306883C	ggtatctcttgctccgtt
rs9657862C/G	154650951	rs9657862C	agaagaactaatgcccca
		rs9657862G	agaagaagtaatgcccca
rs34059672+/-	154651089	rs34059672-	tctgtgtgcgtgtgtgtg
		rs34059672+	tctgtgtgcgcgtgtgtg
rs1736452C/T	154651094	rs1736452C	gtgtgcgtgcgtgtgccg
		rs1736452T	gtgtgcgtgtgtgtgccg
rs71190340C/G	154651100	rs71190340C	gtgtgcccgcgcgcttgg
		rs71190340G	gtgtgcgcgcgcgcttgg
rs306885G/T	154651506	rs306885G	gacacggcagaacccccc
		rs306885T	gacacggcataacccccc
rs3870343C/G	154651570	rs3870343C	cctgggtacctgtgcaca
		rs3870343G	cctgggtacgtgtgcaca
rs306886G/T	154651636	rs306886G	gcacgtctagttagccaa
		rs306886T	gcacgtctatttagccaa
rs6642320A/G	154652328	rs6642320A	tacagccacacagcaaag
		rs6642320G	tacagccacgcagcaaag
rs17653343A/T	154652899	rs17653343A	tattctgagaaggagtgt
		rs17653343T	tattctgagtaggagtgt
rs73638004A/G	154653399	rs73638004G	gcactacgtgacaaaaat
		rs73638004A	gcactacatgacaaaaat
rs9645286A/G	154653679	rs9645286G	tttctcaacatatttttc
		rs9645286A	tttctcaacgtatttttc
rs306887C/T	154654410	rs306887C	agggctccacatattctt
		rs306887T	agggctctacatattctt
rs28630004C/T	154654497	rs28630004T	ctacatctacataccctc
		rs28630004C	ctacatccacataccctc
rs700441C/T	154654680	rs700441C	ttttgaatgccccaatat

SNP name	Position (ChrX)	SNP allele	ASO sequence (5' to 3')
		rs700441T	ttttgaacgccccaatat
rs35535695A/C	154655521	rs35535695C	tattttccccatggtcat
		rs35535695A	tattttcaccatggtcat

Appendix 5.5.

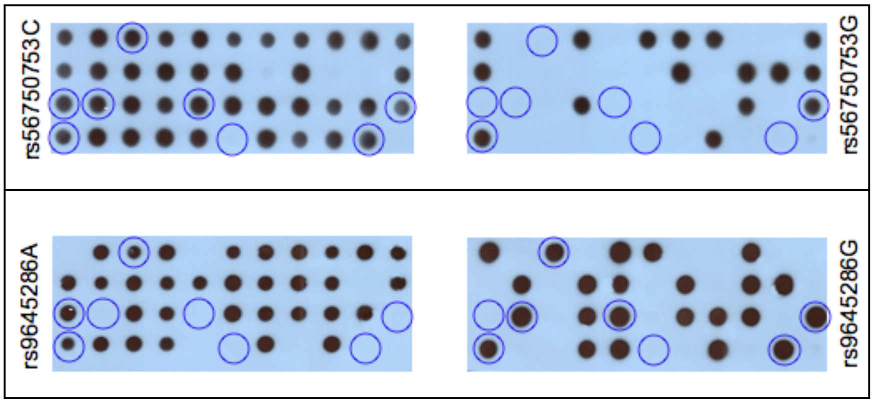
Mapping recombination intervals in CEPH families 1362, 1423, 1377, 1332 and 1416.

PCR products from selected individuals in PAR2-recombinant CEPH pedigrees were dot blotted in the format shown below. Genotyping of markers used to illustrate mapping of COs in family 1423 in Figure 5.4 is also shown.



137710 FF	136215 MF	142303 S	133203 D	141612 FM	137703 S	141602 M	141613 MF	136201 F	136204 rec D	133201 F
141611 FF	141601 F	133211 S	133216 MM	141609 D	141603 S	137712 MF	136202 M	137702 M	136204 rec D	133214 FM
142311 FF	142302 M	136216 MM	137711 FM	142313 MF	133206 rec S	136203 D	133213 FF	141610 rec D	133202 M	142312 FM
142301 F	136213 FF	133215 MF	136214 FM	137701 F	142305 D	141614 MM	137714 rec S	136208 S	142306 rec S	

Format of dot blots generated for purposes of mapping COs in PAR2-recombinant CEPH pedigrees. Order of samples was randomised on purpose. Individuals from CEPH reference family 1377 are shown in red, 1362 in purple, 1423 in blue, 1332 in black and 1416 in green. Individuals are numbered as in the CEPH database. Note: MF = mother's father, MM = mother's mother, FF = father's father, FM = father's mother, M = mother, F = father, S = son, RS = recombinant son, D = daughter, RD = recombinant daughter.



Individuals from CEPH pedigree 1423 are highlighted using blue circles. For SNP rs56750753, individual 142303 (first row, third) is positive for the 'C' allele, but not the 'G' allele. He is thus CC homozygous. Typing of SNP rs9645286 is also shown. Sequences of ASOs used are listed in Appendix 5.4.

In total, 18 SNPs were typed over nearly 15kb (chrX: 154,640, 762 and 154,655,521, as per NCBI36/hg18). Genotypes for all individuals studied are listed in the following tables. It is also indicated whether a marker is uninformative or otherwise, for a particular pedigree. If a marker is informative, it is classed as either a 'recombinant' or a 'non-recombinant'. 'Recombinant' implies that the genotype data for that marker is consistent with an exchange event between the Xq/Yq regions of X and Y-chromosomes. Similarly, 'non-recombinant' indicates a lack of detectable recombination in a particular individual within the pedigree. Where possible, a 'recombination interval', that is, the smallest interval to which a recombination event could be mapped in the present study, is also indicated.

Note: Locations of markers are indicated and are as per NCBI36/hg18. 'Amp.' refers to the name of the amplicon used for genotyping a given marker.

Pedigree 1362:

Recombination interval: 154648189-154650233 (2,044bp).

db SNP entry	**Position**	**Amp.**	**FF**	**FM**	**F**	**MF**	**MM**	**M**	**D**	**S**	**RD**	**Status**
rs73249628C/T	154640762	1	H	H	H	C	C	C	H	C	H	non-recombinant
rs5940571G/C	154643377	1	H	H	H	C	H	C	H	C	H	non-recombinant
rs5940572A/T	154644494	1	H	H	H	A	H	A	H	A	H	non-recombinant
rs28666767G/T	154646691	2	G	G	G	G	G	G	G	G	G	*uninformative*
rs55997085+/-	154648157	2	H	H	H	H	H	+	H	+	H	non-recombinant
rs68008496C/T	154648189	2	H	H	H	H	H	C	H	C	H	non-recombinant
rs67804731A/G	154648273	2	A	A	A	A	H	A	A	A	A	*uninformative*
rs17149475A/G	154648508	2	A	A	A	A	A	A	A	A	A	*uninformative*
rs56750753C/G	154649921	3	C	C	C	C	C	C	C	C	C	*uninformative*
rs306882A/C	154650069	3	H	A	A	C	C	C	H	H	H	*uninformative*
rs700442C/T	154650233	3	H	C	H	H	H	T	H	T	T	recombinant
rs71190340C/G	154651100	3	C	G	H	H	C	H	C	H	H	recombinant
rs306885G/T	154651506	3	G	T	H	T	G	H	G	H	H	recombinant
rs3870343G/C	154651570	3	G	H	H	G	G	G	H	G	G	recombinant
rs6642320G/A	154652328	3	G	H	H	G	G	G	H	G	G	recombinant
rs9645286A/G	154653679	1	A	H	H	A	A	A	A	A	H	recombinant
rs35535695A/C	154655521	1	C	C	C	C	C	C	C	C	C	*uninformative*

Pedigree 1423:

Recombination interval: 154649921- 154651570 (1,649bp).

db SNP entry	**Position**	**Amp.**	**FF**	**FM**	**F**	**MF**	**MM**	**M**	**S**	**D**	**RS**	**Status**
rs73249628C/T	154640762	1	C	C	C	H	/	H	H	/	C	*uninformative*
rs5940571G/C	154643377	1	C	C	C	G	/	G	H	/	H	*uninformative*

db SNP entry	Position	Amp.	FF	FM	F	MF	MM	M	S	D	RS	Status
rs5940572A/T	154644494	1	A	A	A	T	/	T	H	/	H	*uninformative*
rs28666767G/T	154646691	2	G	G	G	G	/	G	G	/	G	*uninformative*
rs55997085+/-	154648157	2	+	+	+	-	/	-	H	/	H	*uninformative*
rs68008496C/T	154648189	2	C	C	C	H	/	H	C	/	H	*uninformative*
rs67804731A/G	154648273	2	A	A	A	G	/	G	H	/	H	*uninformative*
rs17149475A/G	154648508	2	A	A	A	A	/	H	H	/	A	*uninformative*
rs56750753C/G	154649921	3	C	H	H	C	/	C	C	/	C	non-recombinant
rs306882A/C	154650069	3	H	H	C	H	/	A	H	/	H	*uninformative*
rs700442C/T	154650233	3	H	H	C	C	/	C	C	/	C	*uninformative*
rs71190340C/G	154651100	3	H	G	G	H	/	H	G	/	H	*uninformative*
rs306885G/T	154651506	3	H	T	T	T	/	T	T	/	T	*uninformative*
rs3870343G/C	154651570	3	G	C	H	C	/	C	H	/	C	recombinant
rs6642320G/A	154652328	3	G	A	H	A	/	A	H	/	A	recombinant
rs9645286A/G	154653679	1	A	G	H	G	/	G	H	/	G	recombinant
rs35535695A/C	154655521	1	C	C	C	C	/	C	C	/	C	*uninformative*

Pedigree 1377:

Recombination interval: 154648189-?

db SNP entry	Position	Amp.	FF	FM	F	MF	MM	M	S	RS	Status
rs73249628C/T	154640762	1	C	C	C	H	C	C	C	C	*uninformative*
rs5940571G/C	154643377	1	C	H	C	H	C	C	C	C	*uninformative*
rs5940572A/T	154644494	1	A	H	A	H	A	A	A	A	*uninformative*
rs28666767G/T	154646691	2	G	G	G	G	G	G	G	G	*uninformative*
rs55997085+/-	154648157	2	H	+	H	H	+	+	H	H	non-recombinant
rs68008496C/T	154648189	2	H	C	H	C	C	C	H	H	non-recombinant
rs67804731A/G	154648273	2	A	A	A	H	A	A	A	A	*uninformative*
rs17149475A/G	154648508	2	A	A	A	A	A	A	A	A	*uninformative*
rs56750753C/G	154649921	3	H	H	C	G	G	G	H	H	*uninformative*
rs306882A/C	154650069	3	C	H	H	H	H	H	C	H	*uninformative*
rs700442C/T	154650233	3	H	H	T	H	T	T	T	T	*uninformative*
rs71190340C/G	154651100	3	G	G	G	H	G	H	H	G	*uninformative*
rs306885G/T	154651506	3	H	H	T	G	T	H	H	T	*uninformative*
rs3870343G/C	154651570	3	C	H	C	H	C	H	H	C	*uninformative*
rs6642320G/A	154652328	3	A	H	A	H	A	H	H	A	*uninformative*
rs9645286A/G	154653679	1	G	H	G	H	G	H	H	G	*uninformative*
rs35535695A/C	154655521	1	C	C	C	C	C	C	C	C	*uninformative*

Pedigree 1332:

Recombination interval: 154650233-154651100 (867bp)

db SNP entry	**Position**	**Amp.**	**FF**	**FM**	**F**	**MF**	**MM**	**M**	**S**	**RS**	**D**	**Status**
rs73249628C/T	154640762	1	C	C	C	C	C	C	C	C	C	*uninformative*
rs5940571G/C	154643377	1	C	C	C	C	C	C	C	C	C	*uninformative*
rs5940572A/T	154644494	1	A	A	A	A	A	A	A	A	A	*uninformative*
rs28666767G/T	154646691	2	G	G	G	G	G	G	G	G	G	*uninformative*
rs55997085+/-	154648157	2	H	+	H	+	+	+	H	H	+	non-recombinant
rs68008496C/T	154648189	2	H	C	H	C	C	C	H	H	C	non-recombinant
rs67804731A/G	154648273	2	A	A	A	A	A	A	A	A	A	*uninformative*
rs17149475A/G	154648508	2	A	A	A	A	A	A	A	A	A	*uninformative*
rs56750753C/G	154649921	3	C	H	H	C	C	C	C	C	H	non-recombinant
rs306882A/C	154650069	3	C	H	H	C	C	C	C	C	H	non-recombinant
rs700442C/T	154650233	3	C	H	H	H	H	C	C	C	H	non-recombinant
rs71190340C/G	154651100	3	H	G	H	C	H	C	C	H	H	recombinant
rs306885G/T	154651506	3	H	G	G	G	H	G	G	G	G	*uninformative*
rs3870343G/C	154651570	3	H	G	G	G	H	G	G	G	G	*uninformative*
rs6642320G/A	154652328	3	G	G	G	G	H	G	G	G	G	*uninformative*
rs9645286A/G	154653679	1	H	A	A	A	H	A	A	A	A	*uninformative*
rs35535695A/C	154655521	1	C	C	C	C	C	C	C	C	C	*uninformative*

Pedigree 1416:

Recombination interval: 154653679-?

db SNP entry	**Position**	**Amp.**	**FF**	**FM**	**F**	**MF**	**MM**	**M**	**D**	**S**	**RD**	**Status**
rs73249628C/T	154640762	1	C	C	C	C	C	C	C	C	C	*uninformative*
rs5940571G/C	154643377	1	C	C	C	C	C	C	C	C	C	*uninformative*
rs5940572A/T	154644494	1	A	A	A	A	A	A	A	A	A	*uninformative*
rs28666767G/T	154646691	2	G	G	G	G	G	G	G	G	G	*uninformative*
rs55997085+/-	154648157	2	H	+	H	H	H	H	H	-	+	non-recombinant
rs68008496C/T	154648189	2	H	C	H	H	C	C	C	H	C	non-recombinant
rs67804731A/G	154648273	2	A	A	A	A	H	H	A	H	H	*uninformative*
rs17149475A/G	154648508	2	A	A	A	A	A	A	A	A	A	*uninformative*
rs56750753C/G	154649921	3	H	C	C	H	C	H	C	C	H	*uninformative*
rs306882A/C	154650069	3	C	H	C	H	C	H	C	H	C	*uninformative*
rs700442C/T	154650233	3	H	H	C	H	C	H	C	H	C	*uninformative*
rs71190340C/G	154651100	3	C	G	H	H	H	G	G	H	G	non-recombinant
rs306885G/T	154651506	3	H	T	T	T	T	T	T	T	T	*uninformative*
rs3870343G/C	154651570	3	G	C	H	G	G	G	H	G	H	non-recombinant
rs6642320G/A	154652328	3	G	A	H	G	G	G	H	G	H	non-recombinant
rs9645286A/G	154653679	1	A	G	H	A	A	A	H	A	H	non-recombinant
rs35535695A/C	154655521	1	C	C	C	C	C	C	C	C	C	*uninformative*

Appendix 5.6.

a. Genotype data for sperm donor panels

Genotypes for the North European semen donor panel are shown first. Samples 31 and 86, 50 and 95, 30 and 33, 19 and 49 are duplicates. Data from both samples from a pair of duplicates is shown, to highlight consistency of data. Genotypes for the second panel of donors are shown thereafter. Donor 134, 146, 150, 211 and 232 are North European, whilst others are of African origin. For purposes of clarity, this panel is referred to as the 'African panel'. Note: Orange highlight indicates donor names and yellow indicates position on plate/dot blot. All locations are as per NCBI36/hg18. Sequences of ASOs used are listed in appendix 5.2.

b. Minor allele frequencies (MAF) and LDU values for informative markers.

Program used to calculate these values is available online (http://143.210.151.197/cgi-bin/LDMAP/ldumapping.pl). LDU values are plotted against distance in Chapter 5, Figure 5.6.

5.6a

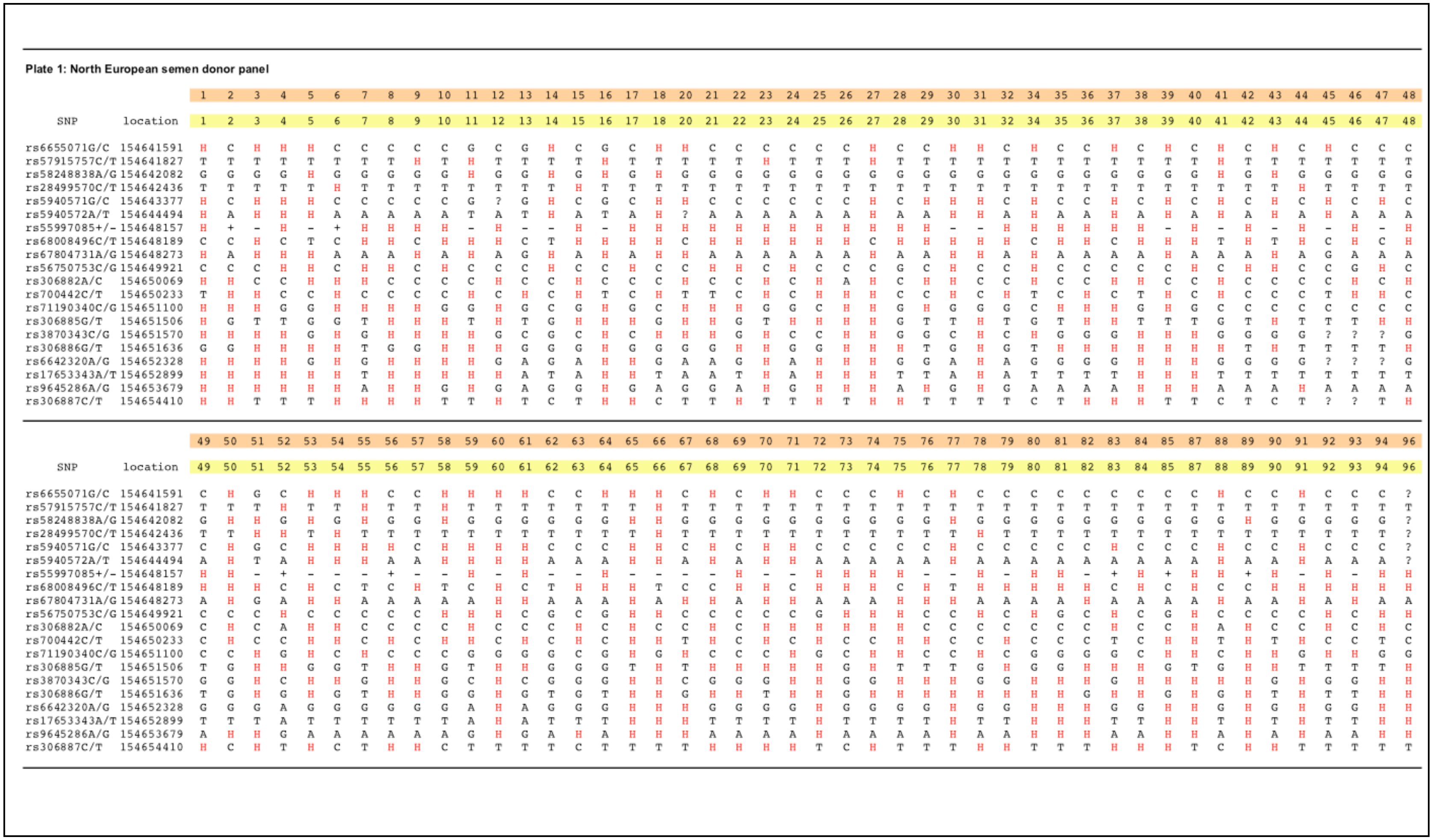


Plate 1: North European semen donor panel

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	20	21	22	23	24	25	26	27	28	29	30	31	32	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
SNP	location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	20	21	22	23	24	25	26	27	28	29	30	31	32	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
rs6655071G/C	154641591	H	C	H	H	H	C	C	C	C	C	G	C	G	H	C	G	C	H	H	C	C	C	C	C	C	H	C	C	H	H	C	H	C	C	H	C	H	C	H	C	H	C	H	C	C	C
rs57915757C/T	154641827	T	T	T	T	T	T	T	T	H	T	H	T	T	T	T	H	T	T	T	T	T	H	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T
rs58248838A/G	154642082	G	G	G	G	H	G	G	G	G	G	H	G	G	H	G	H	G	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	H	G	G	G	G	G
rs28499570C/T	154642436	T	T	T	T	T	H	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T
rs5940571G/C	154643377	H	C	H	H	H	C	C	C	C	C	G	?	G	H	C	G	C	H	H	C	C	C	C	C	C	H	C	H	H	H	C	H	C	C	H	C	H	C	H	C	H	C	H	C	H	C
rs5940572A/T	154644494	H	A	H	H	H	A	A	A	A	A	T	A	T	H	A	T	A	H	?	A	A	A	A	A	A	H	A	A	H	H	A	H	A	A	H	A	H	A	H	A	H	A	H	A	A	A
rs55997085+/-	154648157	H	+	-	H	-	+	H	H	H	H	-	H	-	-	H	-	H	H	H	H	H	H	H	H	H	H	H	H	-	-	H	H	H	H	H	H	-	H	-	H	-	H	-	H	-	H
rs68008496C/T	154648189	C	C	H	C	T	C	H	H	C	H	H	H	C	T	H	H	H	H	C	H	H	H	H	H	H	C	H	H	H	H	H	C	H	H	C	H	H	H	T	H	T	H	C	H	C	H
rs67804731A/G	154648273	H	A	H	H	H	A	A	A	H	A	H	A	G	H	A	H	A	H	H	A	A	A	A	A	A	H	A	A	H	H	A	H	A	A	A	A	H	A	A	A	H	A	G	A	A	A
rs56750753C/G	154649921	C	C	C	H	H	C	H	H	C	H	C	C	C	H	C	C	H	C	C	H	H	C	H	C	C	C	G	C	H	C	C	H	C	C	C	C	C	H	C	H	H	C	C	G	H	C
rs306882A/C	154650069	H	H	C	C	H	H	H	C	C	C	C	H	C	C	H	C	C	C	H	C	C	H	C	H	A	H	C	H	H	C	C	H	C	C	H	H	C	C	H	C	C	C	C	H	C	H
rs700442C/T	154650233	T	H	H	C	C	H	C	C	C	C	H	C	H	C	H	T	C	H	T	T	C	H	C	H	H	H	C	C	H	C	H	T	C	H	C	T	H	C	H	C	C	C	T	H	H	C
rs71190340C/G	154651100	H	H	H	G	G	H	H	H	H	G	G	H	G	C	G	H	G	C	H	H	H	G	G	C	H	H	G	H	G	G	G	C	H	H	H	G	H	G	C	C	C	C	C	C	C	C
rs306885G/T	154651506	H	G	T	T	G	G	T	H	H	H	T	H	T	G	H	H	H	G	H	H	G	T	H	H	H	H	G	T	T	H	T	G	T	H	H	T	T	T	G	T	H	T	T	T	H	H
rs3870343C/G	154651570	H	H	H	H	G	H	G	H	H	H	H	G	C	G	C	H	C	H	H	H	G	H	C	C	H	H	G	G	C	H	C	H	G	G	G	H	H	H	G	G	G	G	?	?	?	G
rs306886G/T	154651636	G	G	H	H	H	H	T	G	G	H	H	H	G	G	G	H	G	G	G	G	H	H	G	G	H	H	H	T	G	H	G	T	H	H	H	H	H	H	H	T	H	T	T	T	T	H
rs6642320A/G	154652328	H	H	H	H	G	H	G	H	H	H	H	G	A	G	A	H	H	G	A	A	G	H	A	H	H	H	G	G	A	H	A	G	G	G	G	H	H	H	G	G	G	G	?	?	?	G
rs17653343A/T	154652899	H	H	H	H	H	H	T	H	H	H	H	H	A	T	A	H	H	T	A	A	T	H	A	H	H	H	T	T	A	H	A	T	T	T	T	H	H	H	T	T	T	T	T	T	T	T
rs9645286A/G	154653679	H	H	H	H	H	H	A	H	H	G	H	G	A	G	G	H	G	A	G	G	A	H	G	H	H	H	A	H	G	H	G	A	A	A	A	H	H	H	A	A	A	H	A	A	A	A
rs306887C/T	154654410	H	H	T	T	T	H	H	H	H	T	T	H	T	C	T	H	H	C	T	T	H	T	T	H	T	H	H	T	T	T	T	C	T	H	H	H	T	T	C	T	C	T	?	?	T	H

		49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	87	88	89	90	91	92	93	94	96
SNP	location	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	87	88	89	90	91	92	93	94	96
rs6655071G/C	154641591	C	H	G	C	H	H	H	C	C	H	H	H	H	C	C	H	H	H	C	H	C	H	H	C	C	C	H	C	H	C	C	C	C	C	C	C	C	C	H	C	C	H	C	C	C	?
rs57915757C/T	154641827	T	T	T	H	T	T	H	T	T	H	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs58248838A/G	154642082	G	H	H	G	H	G	H	G	G	H	G	G	G	G	G	G	H	H	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G	G	?
rs28499570C/T	154642436	T	T	H	H	T	H	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	?
rs5940571G/C	154643377	C	H	G	C	H	H	H	H	C	H	H	H	H	C	C	C	H	H	C	H	C	H	H	C	C	C	C	C	H	C	C	C	C	C	H	C	C	C	H	C	C	H	C	C	C	?
rs5940572A/T	154644494	A	H	T	A	H	H	H	A	A	H	H	H	H	A	A	A	H	H	A	H	A	H	H	A	A	A	A	A	H	A	A	A	A	A	A	A	A	A	H	A	A	H	A	A	A	?
rs55997085+/-	154648157	H	H	-	+	-	-	-	+	-	-	H	-	H	-	-	H	-	-	-	-	H	-	-	H	H	H	H	-	-	H	-	H	H	-	+	H	+	H	H	+	H	-	H	-	H	H
rs68008496C/T	154648189	H	H	H	C	H	C	T	C	H	T	C	H	C	T	H	H	H	T	C	C	H	H	C	H	H	H	C	H	T	H	H	H	H	H	C	H	C	H	C	C	H	H	H	H	H	H
rs67804731A/G	154648273	A	H	G	A	H	H	A	A	A	A	A	H	H	A	A	A	H	A	H	H	A	H	H	A	A	A	H	H	H	A	A	A	A	H	A	A	A	A	H	A	A	H	A	H	A	A
rs56750753C/G	154649921	C	C	C	H	C	C	C	C	C	H	H	H	C	G	C	G	H	H	C	C	C	C	C	G	H	H	H	C	C	H	C	H	G	C	H	C	G	H	C	C	C	C	H	C	H	H
rs306882A/C	154650069	C	H	C	A	H	H	C	C	C	C	H	C	C	C	H	C	H	C	C	H	C	H	H	H	H	H	H	C	C	C	C	C	C	C	H	C	C	H	A	H	C	C	H	C	H	C
rs700442C/T	154650233	C	H	C	C	H	H	C	H	C	H	H	C	H	C	H	C	H	H	T	H	C	H	C	H	C	C	H	H	C	C	H	C	C	C	T	C	H	H	T	H	T	H	C	C	T	C
rs71190340C/G	154651100	C	C	H	G	H	C	H	C	C	C	G	G	G	G	C	G	H	G	H	C	C	C	H	G	C	H	H	C	C	H	C	G	G	G	G	C	H	H	C	H	H	G	H	H	G	G
rs306885G/T	154651506	T	G	H	H	G	G	T	H	H	G	T	H	H	G	G	G	T	H	T	H	H	H	H	H	G	H	T	T	T	G	H	G	G	H	H	H	G	T	G	H	H	T	T	T	T	H
rs3870343C/G	154651570	G	G	H	C	H	H	G	H	H	G	C	H	C	G	G	G	H	H	C	G	G	G	H	H	G	G	H	H	H	G	G	H	H	H	G	H	H	H	H	H	G	H	G	G	H	H
rs306886G/T	154651636	T	G	H	G	H	G	T	H	H	G	G	H	G	T	G	T	H	H	G	H	H	T	H	H	G	H	H	H	H	H	H	H	H	G	H	H	G	H	G	H	T	H	T	T	H	H
rs6642320A/G	154652328	G	G	G	A	G	G	G	G	G	G	A	H	A	G	G	G	H	H	H	G	G	G	G	H	G	G	G	G	H	G	G	H	H	H	G	G	H	H	G	H	G	H	G	G	H	H
rs17653343A/T	154652899	T	T	T	A	T	T	T	T	T	T	A	H	A	T	T	T	H	H	H	T	T	T	T	H	T	T	T	T	H	T	T	H	H	H	T	T	H	H	T	H	T	H	T	T	H	H
rs9645286A/G	154653679	A	H	H	G	A	A	A	A	A	A	G	H	G	A	H	A	H	H	H	A	A	A	A	H	A	A	A	A	H	A	A	H	H	H	A	A	H	H	A	H	A	H	A	A	H	H
rs306887C/T	154654410	H	C	H	T	H	C	T	H	H	C	T	T	T	T	C	T	T	T	T	H	H	H	H	T	C	H	T	T	T	H	H	T	T	T	H	H	H	T	C	H	H	T	T	T	T	T

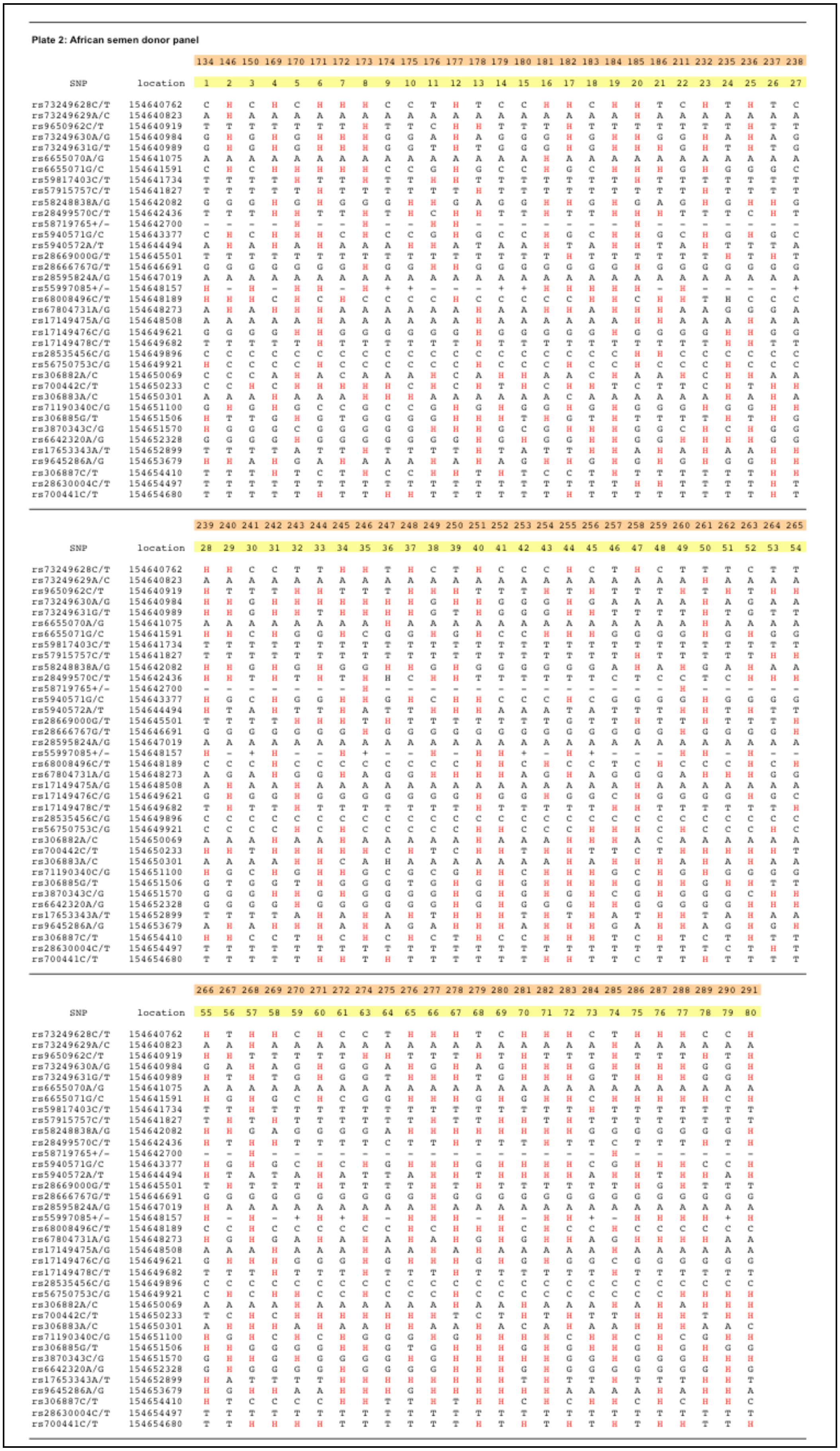


Plate 2: African semen donor panel

		134	146	150	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	211	232	235	236	237	238
SNP	location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
rs73249628C/T	154640762	C	H	C	H	C	H	H	H	C	C	T	H	T	C	C	H	H	C	H	H	T	C	H	T	H	T	C
rs73249629A/C	154640823	A	H	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	A
rs9650962C/T	154640919	T	T	T	T	T	T	T	H	T	T	C	H	H	T	T	T	H	T	T	T	T	T	T	T	H	T	T
rs73249630A/G	154640984	G	H	G	H	G	H	H	H	G	G	A	H	A	G	G	G	H	G	H	H	G	G	H	A	H	A	G
rs73249631G/T	154640989	G	H	G	H	G	H	H	H	G	G	T	H	T	G	G	G	H	G	H	H	H	G	H	T	H	T	G
rs6655070A/G	154641075	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	A	A	A	A	A
rs6655071G/C	154641591	C	H	C	H	H	H	H	H	C	C	G	H	G	C	C	H	G	C	H	H	H	G	H	G	G	G	C
rs59817403C/T	154641734	T	T	T	T	H	T	T	H	T	T	H	H	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T
rs57915757C/T	154641827	T	T	T	T	T	H	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	H	T	T	T	T
rs58248838A/G	154642082	G	G	G	H	G	H	G	G	G	H	H	G	A	G	G	H	H	G	H	G	A	G	H	G	H	H	G
rs28499570C/T	154642436	T	T	T	H	H	T	T	H	T	H	C	H	H	T	T	H	T	T	H	H	H	T	T	T	C	H	T
rs58719765+/-	154642700	-	-	-	-	H	-	-	H	-	-	H	H	-	-	-	-	-	-	-	H	-	-	-	-	-	-	-
rs5940571G/C	154643377	C	H	C	H	H	H	C	H	C	C	G	H	G	C	C	H	G	C	H	H	G	C	H	G	H	G	C
rs5940572A/T	154644494	A	H	A	H	A	H	A	A	A	H	H	A	T	A	A	H	T	A	H	H	T	A	H	T	T	T	A
rs28669000G/T	154645501	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	H	T	H	T
rs28666767G/T	154646691	G	G	G	G	G	G	G	H	G	G	H	H	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G
rs28595824A/G	154647019	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs55997085+/-	154648157	H	-	H	-	H	H	-	H	+	+	-	-	-	+	+	H	H	H	H	H	-	H	-	-	-	-	+
rs68008496C/T	154648189	H	H	H	C	H	C	H	C	C	C	C	H	C	C	C	C	C	H	H	C	H	H	T	H	C	C	C
rs67804731A/G	154648273	A	H	A	H	H	H	A	A	A	A	A	A	H	A	A	H	H	A	H	H	H	A	A	G	G	G	A
rs17149475A/G	154648508	A	A	A	A	A	H	A	A	A	A	A	A	H	A	A	A	A	A	A	H	H	A	A	A	H	A	A
rs17149476C/G	154649621	G	G	G	G	H	H	G	G	G	G	G	G	H	G	G	G	G	G	H	G	G	G	G	H	H	G	G
rs17149478C/T	154649682	T	T	T	T	T	H	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	H	H	T	T
rs28535456C/G	154649896	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	H	H	C	C	C	C	C	C
rs56750753C/G	154649921	H	C	C	C	C	H	C	C	C	C	C	C	H	C	C	C	H	C	C	H	C	C	C	H	C	C	C
rs306882A/C	154650069	C	C	C	A	H	A	C	A	A	A	H	C	A	H	H	A	A	C	H	A	A	H	C	H	H	A	A
rs700442C/T	154650233	C	C	H	C	H	H	H	H	H	C	H	C	H	T	H	C	H	H	T	C	T	T	C	H	T	H	H
rs306883A/C	154650301	A	A	A	H	A	A	A	H	H	H	A	A	A	A	A	A	C	A	A	A	A	A	A	H	A	H	A
rs71190340C/G	154651100	G	H	G	H	G	C	C	G	C	C	G	H	G	H	G	G	H	G	H	G	G	G	H	G	G	H	H
rs306885G/T	154651506	H	T	T	G	H	G	T	G	G	G	G	H	H	H	T	H	G	T	H	T	T	T	T	H	T	H	G
rs3870343C/G	154651570	H	G	G	G	C	G	G	G	G	G	H	H	H	G	C	G	H	H	H	G	G	C	H	C	H	G	G
rs6642320A/G	154652328	G	G	G	G	H	G	G	G	G	G	G	G	H	G	H	G	G	H	H	G	G	H	H	H	H	G	G
rs17653343A/T	154652899	T	T	T	T	A	T	T	H	T	T	T	T	H	T	A	T	T	H	H	A	H	A	H	A	A	H	H
rs9645286A/G	154653679	H	H	A	H	G	A	H	A	A	A	H	A	H	A	G	H	H	G	H	G	H	G	H	G	G	H	H
rs306887C/T	154654410	T	T	T	H	T	C	T	H	C	C	H	H	T	H	T	C	C	T	H	T	T	T	T	T	T	H	H
rs28630004C/T	154654497	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	H	T	T	T	T	H	T
rs700441C/T	154654680	T	T	T	T	T	H	T	T	H	H	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	H	T

		239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265
SNP	location	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
rs73249628C/T	154640762	H	H	C	C	T	T	H	H	T	H	C	T	H	C	C	C	H	C	T	H	C	T	T	T	C	T	T
rs73249629A/C	154640823	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	H	A	A	A	A
rs9650962C/T	154640919	H	T	T	T	H	H	T	T	T	H	H	H	T	T	T	H	T	H	T	T	T	H	T	H	T	H	H
rs73249630A/G	154640984	H	H	G	H	H	H	H	H	H	H	G	H	H	G	G	G	H	G	A	A	A	A	H	A	G	A	A
rs73249631G/T	154640989	H	H	G	H	H	T	H	H	H	H	G	T	H	G	G	G	H	H	T	T	T	T	H	T	G	T	T
rs6655070A/G	154641075	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	A	A	A	A	A	A	A	H	A	A	A	A
rs6655071G/C	154641591	H	H	C	H	G	G	H	C	G	G	H	G	H	C	C	H	H	H	G	G	G	G	H	G	H	G	G
rs59817403C/T	154641734	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs57915757C/T	154641827	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	H	H
rs58248838A/G	154642082	H	H	G	H	G	H	G	G	H	H	G	H	G	G	G	G	G	G	A	H	A	H	G	A	H	A	A
rs28499570C/T	154642436	H	H	T	H	T	H	T	H	H	C	H	H	T	T	T	T	T	T	C	T	C	C	T	C	H	H	H
rs58719765+/-	154642700	-	-	-	-	-	-	-	H	-	-	-	-	-	-	-	-	-	-	-	-	-	H	-	-	-	-	-
rs5940571G/C	154643377	H	G	C	H	G	G	H	H	G	H	C	H	H	C	C	C	H	C	G	G	G	G	H	G	G	G	G
rs5940572A/T	154644494	H	T	A	H	T	T	H	A	T	T	H	H	H	A	A	A	T	A	T	T	T	H	H	T	H	T	T
rs28669000G/T	154645501	T	T	T	T	H	H	H	T	H	T	T	T	T	T	T	T	G	T	T	H	T	T	H	T	T	T	H
rs28666767G/T	154646691	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G	H
rs28595824A/G	154647019	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs55997085+/-	154648157	H	-	+	H	-	-	H	+	-	-	H	-	H	H	+	-	H	+	-	-	-	H	H	-	H	-	-
rs68008496C/T	154648189	C	C	C	H	C	C	C	C	C	C	C	C	H	H	C	H	C	C	T	C	H	C	C	C	H	C	H
rs67804731A/G	154648273	A	G	A	H	G	G	H	A	G	G	H	H	H	H	A	G	H	A	G	G	G	A	H	H	H	G	G
rs17149475A/G	154648508	A	H	A	A	H	A	A	A	A	A	A	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	A
rs17149476C/G	154649621	G	H	G	G	H	G	G	G	G	G	G	G	H	G	G	H	G	G	C	H	G	G	G	G	H	G	C
rs17149478C/T	154649682	T	H	T	T	H	T	T	T	T	T	T	T	H	T	T	T	T	T	H	H	T	T	T	T	T	T	H
rs28535456C/G	154649896	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs56750753C/G	154649921	C	C	C	C	H	C	H	C	C	C	C	C	H	H	C	C	C	H	H	H	C	H	C	C	C	H	C
rs306882A/C	154650069	A	A	A	H	A	A	H	A	A	A	A	A	H	A	A	A	H	H	H	A	C	A	A	A	A	A	A
rs700442C/T	154650233	H	H	T	H	H	H	H	H	C	H	T	C	H	H	T	H	H	T	T	C	T	H	H	H	H	H	T
rs306883A/C	154650301	A	A	A	A	H	H	C	A	H	A	A	A	A	A	A	A	H	A	H	H	H	A	H	A	H	A	A
rs71190340C/G	154651100	H	G	C	H	G	H	H	G	C	G	C	G	H	H	C	H	H	H	G	C	H	G	H	G	G	G	G
rs306885G/T	154651506	G	T	G	G	T	H	G	G	G	T	G	H	G	H	G	H	H	H	H	G	H	H	G	H	H	T	T
rs3870343C/G	154651570	G	G	G	H	H	G	H	G	G	G	G	H	G	H	G	H	G	H	C	G	H	G	G	G	C	H	H
rs6642320A/G	154652328	G	G	G	G	H	G	G	G	G	G	G	H	G	H	G	G	G	H	H	G	G	G	G	G	H	H	H
rs17653343A/T	154652899	T	T	T	T	A	H	A	H	A	H	T	H	H	H	T	H	T	H	A	T	H	H	T	A	H	A	A
rs9645286A/G	154653679	A	H	A	H	H	H	A	H	A	G	A	H	H	H	A	H	H	H	G	A	H	H	A	G	H	G	H
rs306887C/T	154654410	H	H	C	C	T	H	C	H	C	H	C	T	H	C	C	H	H	H	T	C	H	T	C	T	H	T	T
rs28630004C/T	154654497	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	C	T	H	T
rs700441C/T	154654680	T	T	T	T	T	H	H	T	H	T	T	T	T	T	T	H	H	T	T	C	T	T	H	T	T	T	T

		266	267	268	269	270	271	272	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291
SNP	location	55	56	57	58	59	60	61	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
rs73249628C/T	154640762	H	T	H	H	C	H	C	C	T	H	H	H	T	C	H	H	H	C	T	H	H	H	C	C	H
rs73249629A/C	154640823	A	A	H	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A
rs9650962C/T	154640919	H	H	T	T	T	T	T	H	H	T	T	T	H	T	H	T	T	T	H	T	T	T	H	T	H
rs73249630A/G	154640984	G	A	H	A	G	H	G	G	A	H	G	H	A	G	H	H	H	G	H	H	H	H	G	G	H
rs73249631G/T	154640989	H	T	H	T	G	H	G	G	T	H	H	H	T	G	H	H	H	G	T	H	H	H	G	G	H
rs6655070A/G	154641075	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs6655071G/C	154641591	H	G	H	G	C	H	C	G	G	H	H	H	G	H	G	H	H	C	H	H	H	H	H	C	H
rs59817403C/T	154641734	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T
rs57915757C/T	154641827	T	H	T	H	T	T	T	T	T	T	H	T	T	H	H	T	H	T	T	T	T	T	T	T	T
rs58248838A/G	154642082	H	H	G	A	G	G	G	G	A	H	H	H	H	H	H	H	H	G	G	G	G	G	G	G	H
rs28499570C/T	154642436	H	T	H	H	T	T	T	T	C	T	T	H	T	T	T	H	T	T	C	T	T	T	H	T	H
rs58719765+/-	154642700	-	-	H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	H	-	-	-	-	-	-
rs5940571G/C	154643377	H	G	H	G	C	H	C	H	G	H	H	H	G	H	H	H	H	C	G	H	H	H	C	C	H
rs5940572A/T	154644494	H	T	A	T	A	H	A	T	T	A	H	H	T	H	H	H	H	A	H	H	T	H	H	A	H
rs28669000G/T	154645501	T	H	T	T	T	H	T	T	T	T	H	T	H	T	T	T	T	T	T	H	G	H	T	T	T
rs28666767G/T	154646691	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G
rs28595824A/G	154647019	H	A	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs55997085+/-	154648157	H	-	H	-	+	H	+	H	-	H	H	H	-	H	-	H	H	+	-	H	H	H	H	+	H
rs68008496C/T	154648189	C	C	H	C	C	C	C	C	C	H	C	H	H	H	C	H	C	C	H	C	C	C	C	C	C
rs67804731A/G	154648273	H	G	H	G	A	H	A	H	A	H	A	H	G	H	G	H	H	A	G	H	H	H	H	A	A
rs17149475A/G	154648508	A	A	A	H	A	A	A	H	A	A	H	A	H	A	A	A	A	A	H	A	A	A	A	A	A
rs17149476C/G	154649621	G	H	H	H	G	G	G	H	G	H	H	H	G	H	G	H	G	G	C	G	G	G	G	G	G
rs17149478C/T	154649682	T	T	T	H	T	T	T	H	T	T	T	H	T	T	T	T	T	T	H	T	T	T	T	T	T
rs28535456C/G	154649896	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs56750753C/G	154649921	C	H	C	H	H	C	C	H	C	C	C	H	C	C	C	C	C	C	C	C	C	H	H	H	H
rs306882A/C	154650069	A	A	A	A	H	A	A	A	A	A	A	H	A	A	H	A	A	A	H	A	H	A	H	H	H
rs700442C/T	154650233	T	C	H	C	H	H	H	H	H	H	H	T	C	T	H	T	H	T	T	H	H	H	T	H	H
rs306883A/C	154650301	A	H	H	H	A	H	A	A	H	H	A	A	H	A	C	A	H	A	A	H	H	H	A	A	C
rs71190340C/G	154651100	H	G	H	C	H	C	H	G	G	G	H	G	H	H	H	H	C	H	H	C	H	C	G	H	H
rs306885G/T	154651506	H	H	G	G	G	G	H	H	G	T	G	H	H	H	G	H	G	H	H	G	H	G	H	G	G
rs3870343C/G	154651570	G	H	H	G	H	G	G	G	G	H	G	H	H	H	H	H	G	G	H	G	G	G	H	H	H
rs6642320A/G	154652328	G	H	G	G	G	G	H	G	G	G	G	H	H	H	G	H	G	G	G	G	G	G	G	H	G
rs17653343A/T	154652899	H	A	T	T	T	T	H	H	H	H	H	H	H	H	T	H	T	T	H	T	T	T	H	H	T
rs9645286A/G	154653679	H	G	H	H	A	A	H	H	H	G	H	H	H	H	H	H	A	A	A	A	H	A	H	H	A
rs306887C/T	154654410	H	T	C	C	C	C	H	H	T	T	H	T	H	H	C	H	C	H	H	C	H	C	H	H	C
rs28630004C/T	154654497	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs700441C/T	154654680	T	T	H	H	H	H	T	T	T	T	T	T	H	T	H	T	H	T	H	T	H	H	T	T	H

Table 5.6b

African dataset

dbSNP entry	Location (kb)	MAF	LD units
rs73249628	154641	0.456	0.000
rs73249629	154641	0.032	0.000
rs9650962	154641	0.177	0.000
rs73249630	154641	0.418	0.233
rs73249631	154641	0.462	0.233
rs6655070	154641	0.019	0.233
rs6655071	154642	0.443	0.233
rs59817403	154642	0.044	0.233
rs57915757	154642	0.076	0.233
rs58248838	154642	0.310	0.233
rs28499570	154642	0.297	0.233
rs58719765	154643	0.057	0.233
rs5940571	154643	0.487	0.254
rs5940572	154644	0.500	0.277
rs28669000	154646	0.127	0.379
rs28666767	154647	0.051	0.489
rs28595824	154647	0.013	0.489
rs55997085	154648	0.392	0.489
rs68008496	154648	0.184	0.708
rs67804731	154648	0.449	0.708
rs17149475	154649	0.082	1.252
rs17149476	154650	0.171	1.393
rs17149478	154650	0.089	1.393
rs28535456	154650	0.013	1.474
rs56750753	154650	0.152	1.501
rs306882	154650	0.241	1.574
rs700442	154650	0.468	1.654
rs306883	154650	0.215	1.654
rs71190340	154651	0.392	2.573
rs306885	154652	0.405	2.573
rs3870343	154652	0.278	2.573
rs6642320	154652	0.152	2.573
rs17653343	154653	0.380	2.573
rs9645286	154654	0.437	2.573
rs306887	154654	0.487	2.573
rs28630004	154654	0.038	2.573
rs700441	154655	0.152	2.573

North European dataset

dbSNP entry	Location (kb)	MAF	LD units
rs6655071	154642	0.231	0.000
rs57915757	154642	0.054	0.000
rs58248838	154642	0.088	0.000
rs28499570	154642	0.044	0.000
rs5940571	154643	0.244	0.000
rs5940572	154644	0.217	0.000
rs55997085	154648	0.353	0.045
rs68008496	154648	0.413	0.062
rs67804731	154648	0.212	0.062
rs56750753	154650	0.250	0.442
rs306882	154650	0.234	0.864
rs700442	154650	0.337	1.302
rs71190340	154651	0.489	3.302
rs306885	154652	0.457	3.338
rs3870343	154652	0.376	3.395
rs306886	154652	0.440	3.396
rs6642320	154652	0.303	3.396
rs17653343	154653	0.304	3.396
rs9645286	154654	0.348	3.396
rs306887	154654	0.311	3.396

Chapter 6 Appendices

Appendix 6.1

Primer combinations and PCR conditions used in ‘assay 2’.

14 men were studied using ‘assay 1’ described in Chapter 5, using primers and conditions listed in Appendix 5.3(c). Others were assayed using ‘assay 2’. All, but one of these men were assayed using the primers and reaction conditions listed in table 6.1a. One man, d240, was assayed using different primers, as listed in table 6.1b.

Table 6.1a. Assays conditions for ‘assay 2’

Primary PCRs:

Primers	PCR conditions
Forward primer: 2.5F - 5' TCCATCCCCATAGGCTTTTC 3' Reverse primer: RAS5286A/G 5' gggggTAAGAGGGAAAAATAT 3' 5' gggggTAAGAGGGAAAAATAC 3'	96°C for 1min $\left.\begin{array}{l}\text{96°C for 20 sec}\\ \text{60°C for 30 sec}\\ \text{62°C for 13:30 min}\end{array}\right\}$ x 28 cycles

Secondary PCRs:

Primers	PCR conditions
Forward primer: 2.6F - 5' CTATGTTCCCTTGGCATTGC 3' Reverse primer: RAS3343A/T 5' gggCTGTTATATACACTCCTA 3' 5' gggCTGTTATATACACTCCTT 3'	96°C for 1min $\left.\begin{array}{l}\text{96°C for 20 sec}\\ \text{60°C for 30 sec}\\ \text{62°C for 12:30 min}\end{array}\right\}$ x 28 cycles

Tertiary PCRs:

Primers	PCR conditions
Forward primer: S11 - 5' GGGTTCACGCTCCTATGA 3' Reverse primer: 7.1R - 5' TACTAACCTCTCTGCCACTG 3'	96°C for 1min $\left.\begin{array}{l}\text{96°C for 20 sec}\\ \text{56°C for 30 sec}\\ \text{65°C for 10:30 min}\end{array}\right\}$ x 30 cycles

Table 6.1b. Assay conditions used for d240

Primary PCRs:

Primers	PCR conditions

Forward primer:
`64.46F -       5'ACTGTTCCACATGGTAACAC 3'`
Reverse primer:
`RAs6887C/T     5' gggTCGGCTAAAGAATATGTG 3'`
`               5' gggTCGGCTAAAGAATATGTA 3'`
	96°C for 1min
	96°C for 20 sec
	60°C for 30 sec
	62°C for 11:30 min
	x 28 cycles

Secondary PCRs:

Primers	PCR conditions

Forward primer:
`64.47F -       5' TCCCTTTCCTCTACCTCAGC 3'`
Reverse primer:
`RAS5286A/G     5' gggggTAAGAGGGAAAAATAT 3'`
`               5' gggggTAAGAGGGAAAAATAC 3'`
	96°C for 1min
	96°C for 20 sec
	60°C for 30 sec
	62°C for 10:30 min
	x 28 cycles

Tertiary PCRs:

Primers	PCR conditions

Forward primer:
`2.2F -   5' ACAGTACTTTGAAGGGGTC 3'`
Reverse primer:
`7.1R -   5' TACTAACCTCTCTGCCACTG 3'`
	96°C for 1min
	96°C for 20 sec
	56°C for 30 sec
	65°C for 9 min
	x 30 cycles

Appendix 6.2.

SNP haplotypes of men assayed at hot spot *SPRY3*

man	origin	marker[a]	rs5940572A/T	rs28669000G/T	rs28666767G/T	rs28595824A/G	rs55997085+/-	rs68008496C/T	rs67804731A/G	rs17149475A/G	rs17149476C/G	rs17149478C/T	s9.72A/G	s9.81C/G	s649.7C/T	s649.8A/G	rs56750753C/G	rs306882A/C	s92+A/-A/+G	s11GA/AG	rs700442C/T	s0.25+/-	rs306883A/C	s0.31C/T	s0.66A/C	rs71190340C/G	rs306885G/T	rs3870343C/G	rs306886G/T	rs6642320A/G	rs17653343A/T	rs9645286A/G
		location[b]	154644494	154645501	154646691	154647019	154648157	154648189	154648273	154648508	154649621	154649682	154649721	154649811	154649812	154649882	154649921	154650069	154650095	154650111	154650233	154650253	154650301	154650340	154650655	154651100	154651506	154651570	154651636	154652328	154652899	154653679
d1	Eu		A	T	T	A	+	C	A	G	G	T	A	C	T	A	C	C	+A	GA	T	-	A	T	C	C	G	G	G	G	A	A
			T	T	T	A	-	C	G	G	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d3	Eu		A	T	T	A	-	T	A	G	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	C	T	G	T	G	A	A
			T	T	T	A	-	C	G	G	G	T	A	C	T	A	C	C	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d66	Eu		T	T	T	A	-	T	A	G	G	T	A	C	T	A	C	C	+A	GA	T	-	A	T	C	G	T	G	T	G	A	A
			A	T	T	A	-	T	A	G	G	T	A	C	T	A	G	C	+A	GA	C	-	A	T	C	G	G	C	G	A	T	G
d278	Af		A	T	G	A	+	C	A	A	G	T	A	C	C	A	C	A	+A	AG	T	-	A	T	C	G	T	G	T	G	A	A
			T	T	G	A	-	T	G	A	C	C	A	C	T	A	G	C	+A	GA	T	-	A	T	C	G	G	C	G	A	T	G
d25	Eu		A	T	T	A	-	T	A	G	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	C	G	C	G	G	A	A
			A	T	T	A	+	C	A	G	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	C	T	C	G	A	T	G
d274	Af		T	T	G	A	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	G	G	G	G	G	A	A
			T	T	G	A	-	C	G	G	C	C	G	G	T	A	G	A	+A	GA	C	-	A	T	C	G	T	G	G	G	T	G
d39	Eu		A	T	T	A	-	T	G	G	G	T	A	C	T	A	C	C	+A	GA	T	-	A	T	C	C	T	G	T	A	A	A
			T	T	T	A	-	C	A	G	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	G	T	C	G	G	T	G
d27	Eu		T	T	T	A	-	C	G	G	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	C	G	G	T	G	A	A
			A	T	T	A	+	C	A	G	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d250	Af		A	T	G	A	-	C	A	A	G	T	A	C	T	A	C	A	+A	GA	C	-	A	T	C	G	G	G	G	G	A	A
			T	T	G	A	-	C	G	A	G	T	A	C	T	A	C	A	+G	GA	C	-	A	T	C	G	T	C	G	A	T	G
d5	Eu		A	T	G	A	-	T	A	G	G	T	A	C	T	A	G	C	+A	GA	C	-	A	T	C	G	G	G	T	G	A	A
			T	T	G	A	-	T	G	G	G	T	A	C	T	A	C	A	+A	AG	C	-	A	T	C	G	G	G	G	G	T	G
d252	Af		A	T	G	A	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	C	G	G	G	G	A	A
			A	T	G	A	-	T	G	A	G	T	A	C	T	A	G	A	+A	AG	C	+	A	T	C	G	T	C	G	A	T	G
d91	Eu		A	T	T	A	-	T	A	G	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	G	T	G	T	G	A	A
			T	T	T	A	-	C	G	G	G	T	A	C	T	A	C	C	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d280	Af		A	T	G	A	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	C	G	G	G	G	A	A
			T	T	G	A	-	T	G	A	C	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d232	Eu		A	T	G	A	-	T	A	A	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	C	T	G	T	G	A	A
			T	T	G	A	-	T	A	A	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	G	T	C	G	A	T	G
d277	Af		A	T	G	G	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	C	G	G	G	G	A	A
			T	G	T	A	-	C	A	G	C	T	A	C	T	A	C	A	+A	GA	C	-	A	T	C	G	G	G	G	G	T	G

	marker[a]	rs5940572A/T	rs28669000G/T	rs28666767G/T	rs28595824A/G	rs55997085+/-	rs68008496C/T	rs67804731A/G	rs17149475A/G	rs17149476C/G	rs17149478C/T	s9.72A/G	s9.81C/G	s649.7C/T	s649.8A/G	rs56750753C/G	rs306882A/C	s92+A/-A/+G	s11GA/AG	rs700442C/T	s0.25+/-	rs306883A/C	s0.31C/T	s0.66A/C	rs71190340C/G	rs306885G/T	rs3870343C/G	rs306886G/T	rs6642320A/G	rs17653343A/T	rs9645286A/G
	location[b]	154644494	154645501	154646691	154647019	154648157	154648189	154648273	154648508	154649621	154649682	154649721	154649811	154649812	154649882	154649921	154650069	154650095	154650111	154650233	154650253	154650301	154650340	154650655	154651100	154651506	154651570	154651636	154652328	154652899	154653679
man	origin																														
d244	Af	T	G	G	A	-	C	G	A	G	T	G	C	T	A	C	A	+A	AG	C	-	C	T	C	C	G	G	G	G	A	A
		T	T	G	A	-	C	G	A	G	T	A	C	T	A	C	A	+G	GA	T	-	A	T	A	G	T	G	G	G	T	G
d184	Af	A	T	G	A	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	C	G	G	?	G	A	A
		T	T	G	A	-	T	G	A	C	T	A	C	T	A	C	C	+A	AG	T	-	A	T	C	G	T	C	?	A	T	G
d251	Af	A	T	G	A	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	C	G	G	G	G	A	A
		T	T	G	A	-	T	G	A	C	C	A	C	T	A	G	C	+G	GA	C	-	A	C	C	G	G	G	G	G	T	G
d246	Af	A	T	G	A	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	C	-	A	T	C	G	G	G	G	G	A	A
		A	T	T	A	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	T	-	G	T	C	G	G	G	G	G	T	G
d178	Af	T	T	G	A	-	C	A	A	G	T	A	C	T	A	C	A	+A	GA	C	-	A	T	C	G	G	G	G	G	A	A
		T	T	G	A	-	C	G	G	C	C	A	C	T	A	G	A	+A	AG	T	-	A	T	C	G	T	C	G	A	T	G
d23	Eu	A	T	T	A	-	T	A	G	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	G	T	G	T	G	A	A
		A	T	T	A	+	C	A	G	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d87	Eu	A	T	T	A	-	T	A	G	G	T	A	C	T	A	G	C	+A	GA	C	-	A	T	C	C	T	G	T	G	A	A
		A	T	T	A	+	C	A	G	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d282	Af	A	T	G	A	+	C	A	A	G	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	C	G	G	G	G	A	A
		T	T	G	A	-	T	G	A	C	T	A	C	T	A	C	A	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d72	Eu	A	T	T	A	-	T	A	G	G	T	A	C	T	A	G	C	+A	GA	C	-	A	T	C	G	G	G	T	G	A	A
		A	T	T	A	+	C	A	G	G	T	A	C	T	A	G	A	+A	GA	T	-	A	T	C	G	T	C	G	A	T	G
d60	Eu	A	T	T	A	-	T	A	G	G	T	A	C	T	A	G	C	+A	GA	C	-	A	T	C	G	G	G	T	G	A	A
		T	T	T	A	-	C	G	G	G	T	A	C	T	A	C	C	+A	GA	C	-	A	T	C	G	T	C	T	A	T	G
d240	Af	T	T	G	A	-	C	G	G	C	C	A	C	C	A	C	A	+A	GA	T	-	A	T	C	G	T	G	T	G	T	A
		T	T	G	A	-	C	G	A	G	T	A	C	T	A	C	A	+G	GA	C	-	A	T	C	G	T	G	T	G	T	G
d279	Af	T	G	G	A	-	?	G	A	G	T	A	C	T	A	C	A	+A	AG	C	-	C	T	C	C	G	G	G	G	A	A
		T	T	G	A	-	?	G	G	G	T	A	C	T	A	C	A	+A	AG	C	-	A	T	C	G	T	C	G	A	T	G
d254	Af	A	T	G	A	-	T	G	A	C	T	A	C	T	G	C	A	+G	GA	T	-	A	T	C	C	G	G	G	G	A	A
		A	T	G	A	-	C	G	A	G	T	A	C	T	A	C	A	-A	AG	C	-	A	T	C	G	T	C	G	G	T	G

[a] markers prefixed with "s" were found through direct sequencing of separated haplotypes

[b] locations relative to the NCBI36/hg18 build of chr X

Appendix 6.3.

Transmission frequencies of hot spot alleles in men assayed in both orientations at *SPRY3*

Note: Positions of all markers are listed in Appendix 6.2.

- **d3**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	221	0.500	1.000
	G	99	0	0.500	
rs306886	G	1	221	0.505	0.309
	T	98	0	0.495	
rs3870343	C	1	221	0.505	0.309
	G	98	0	0.495	
rs71190340	G	3	218	0.508	0.378
	C	96	3	0.492	
rs700442	T	26	75	0.301	5.50E-11
	C	73	146	0.699	
rs67804731	G	96	5	0.496	0.706
	A	3	216	0.504	
rs68008496	C	99	5	0.511	0.329
	T	0	216	0.489	
rs5940572	T	99	0	0.500	1.000
	A	0	221	0.500	

- **d25**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	46	0.500	1.000
	G	26	0	0.500	
rs306885	T	0	43	0.467	0.549
	G	26	3	0.533	
rs700442	T	11	19	0.418	0.223
	C	15	27	0.582	
rs306882	A	17	8	0.414	0.147
	C	9	38	0.586	
rs68008496	C	25	0	0.481	0.361
	T	1	46	0.519	
rs55997085	+	26	0	0.500	1.000
	-	0	46	0.500	

- **d23**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	3	0.500	1.000
	G	6	0	0.500	
rs306886	G	0	3	0.500	1.000
	T	6	0	0.500	
rs3870343	C	0	3	0.500	1.000
	G	6	0	0.500	
rs700442	T	5	2	0.750	0.226
	C	1	1	0.250	
rs306882	A	6	2	0.833	0.083
	C	0	1	0.167	
rs67804731	C	6	0	0.500	1.000
	T	0	3	0.500	
rs55997085	+	6	0	0.500	1.000
	-	0	3	0.500	

- **d27**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	160	0.500	1.000
	G	133	0	0.500	
rs306886	G	0	160	0.500	1.000
	T	133	0	0.500	
rs3870343	C	0	160	0.500	1.000
	G	133	0	0.500	
rs306885	T	0	159	0.497	1.000
	G	133	1	0.503	
rs71190340	G	1	159	0.501	1.000
	C	132	1	0.499	
rs700442	T	22	79	0.330	9.73E-10
	C	111	81	0.670	
rs306882	A	58	53	0.384	8.88E-05
	C	75	107	0.616	
rs67804731	A	130	8	0.514	0.355
	G	3	152	0.486	
rs55997085	+	131	5	0.508	0.462
	-	2	155	0.492	
rs5940572	A	133	0	0.500	1.000
	T	0	160	0.500	

- **d39**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	G	0	6	0.500	1.000
	A	13	0	0.500	
rs306886	G	0	6	0.500	1.000
	T	13	0	0.500	
rs71190340	G	0	6	0.500	1.000
	C	13	0	0.500	

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs700442	C	7	3	0.519	1.000
	T	6	3	0.481	
rs67804731	A	13	0	0.500	1.000
	G	0	6	0.500	

- **d66**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	323	0.500	1.000
	G	245	0	0.500	
rs306886	G	1	323	0.502	0.431
	T	244	0	0.498	
rs3870343	C	1	322	0.500	1.000
	G	244	1	0.500	
rs306885	G	3	321	0.503	0.656
	T	242	2	0.497	
rs700442	C	139	225	0.632	2.95E-10
	T	106	98	0.368	
rs56750753	G	214	74	0.551	0.002
	C	31	249	0.449	
rs5940572	A	245	0	0.500	1.000
	T	0	323	0.500	

- **d72**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	6	0.500	1.000
	G	3	0	0.500	
rs306886	G	0	6	0.500	1.000
	T	3	0	0.500	
rs3870343	C	0	6	0.500	1.000
	G	3	0	0.500	
rs306885	G	0	6	0.500	1.000
	T	3	0	0.500	
rs700442	C	2	3	0.583	1.000
	T	1	3	0.417	
rs306882	C	2	2	0.500	1.000
	A	1	4	0.500	
rs68008496	C	3	0	0.500	1.000
	T	0	6	0.500	
rs55997085	+	3	0	0.500	1.000
	-	0	6	0.500	

- **d87**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	9	0.500	1.000
	G	9	0	0.500	
rs306886	G	0	9	0.500	1.000

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
	T	9	0	0.500	
rs3870343	C	0	9	0.500	1.000
	G	9	0	0.500	
rs71190340	G	0	9	0.500	1.000
	C	9	0	0.500	
rs700442	T	6	5	0.611	0.637
	C	3	4	0.389	
rs306882	A	6	1	0.389	0.577
	C	3	8	0.611	
rs56750753	C	7	0	0.389	0.471
	G	2	9	0.611	
rs68008496	C	9	0	0.500	1.000
	T	0	9	0.500	
rs55997085	+	9	0	0.500	1.000
	-	0	9	0.500	

- **d91**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	193	0.500	1.000
	G	168	0	0.500	
rs306886	G	0	193	0.500	1.000
	T	168	0	0.500	
rs3870343	C	0	193	0.500	1.000
	G	168	0	0.500	
rs700442	T	49	83	0.361	1.01E-07
	C	119	110	0.639	
rs67804731	G	167	15	0.536	0.001
	A	1	178	0.464	
rs68008496	C	167	1	0.500	1.000
	T	1	192	0.500	
rs5940572	T	168	0	0.500	1.000
	A	0	193	0.500	

- **d178**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs6642320	A	0	5	0.500	1.000
	G	9	0	0.500	
rs3870343	C	0	5	0.500	1.000
	G	9	0	0.500	
rs306885	T	0	5	0.500	1.000
	G	9	0	0.500	
rs700442	T	4	1	0.322	0.301
	C	5	4	0.678	
S11	AG	4	0	0.222	0.086
	GA	5	5	0.778	
rs56750753	G	4	0	0.222	0.086
	C	5	5	0.778	
rs17149478	C	5	0	0.278	0.221
	T	4	5	0.722	
rs17149476	C	5	0	0.278	0.221

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
	G	4	5	0.722	
rs17149475	G	8	0	0.444	1.000
	A	1	5	0.556	
rs67804731	G	9	0	0.500	1.000
	A	0	5	0.500	

- **d244**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs306885	T	0	51	0.500	1.000
	G	18	0	0.500	
SS0.66	A	2	43	0.477	1.000
	C	16	8	0.523	
rs306883	A	5	36	0.492	1.000
	C	13	15	0.508	
rs70442	T	5	35	0.482	1.000
	C	13	16	0.518	
S11	GA	7	34	0.528	1.000
	AG	11	17	0.472	
S92	G	7	34	0.528	0.78
	A	11	17	0.472	
rs28669000	T	18	0	0.500	1.000
	G	0	51	0.500	

- **d274**

Marker	Allele	Orientation A COs	Orientation B COs	Transmission frequencies	Fisher exact *P*, 2-sided
rs306885	T	0	47	0.500	1.000
	G	20	0	0.500	
rs70442	C	16	32	0.740	0.0004
	T	4	15	0.260	
rs56750753	G	19	4	0.518	1.000
	C	1	43	0.482	
S9.81	G	20	2	0.521	1.000
	A	0	45	0.479	
S9.72	G	20	1	0.511	1.000
	C	0	46	0.489	
rs17149478	C	20	1	0.511	1.000
	T	0	46	0.489	
rs17149476	C	20	1	0.511	1.000
	G	0	46	0.489	
rs17149475	G	20	0	0.500	1.000
	A	0	47	0.500	
rs67804731	G	20	0	0.500	1.000
	A	0	47	0.500	
rs55997085	-	20	0	0.500	1.000
	+	0	47	0.500	

Appendix 6.4.

Various repeat types seen in the *PRDM9* ZnF coding domain

Different repeat types observed either by Baudat *et al.* (2010) or by Berg *et al.* (2010) are shown. Variant bases are coloured, rest are in black. Repeats M and N were reported by Baudat *et al.* (2010), but not seen in the semen donor panels. Amino acids sequences encoded by various repeats are indicated in grey and positioned above respective codons. Sequences/amino acids specifying DNA-binding residues are highlighted in yellow. -1, 2, 3 and 6 refer to positions, relative to the alpha helix of a ZnF.

							-1		2	3			6															
	C	G	Q	G	F	S	V	K	S	D	V	I	T	H	Q	R	T	H	T	G	E	K	L	Y	V	C	R	E
A	TGT	GGA	CAA	GGT	TTC	AGT	GTT	AAA	TCA	GAT	GTT	ATT	ACA	CAC	CAA	AGG	ACA	CAT	ACA	GGG	GAG	AAG	CTC	TAC	GTC	TGC	AGG	GAG
	C	G	R	G	F	S	W	K	S	H	L	L	I	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
B	TGT	GGG	CGG	GGC	TTT	AGC	TGG	AAG	TCA	CAC	CTC	CTC	ATT	CAC	CAG	AGG	ATA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	S	W	Q	S	V	L	L	T	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
C	TGT	GGG	CGG	GGC	TTT	AGC	TGG	CAG	TCA	GTC	CTC	CTC	ACT	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	S	R	Q	S	V	L	L	T	H	Q	R	R	H	T	G	E	K	P	Y	V	C	R	E
D	TGT	GGG	CGG	GGC	TTT	AGC	CGG	CAG	TCA	GTC	CTC	CTC	ACT	CAC	CAG	AGG	AGA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	S	W	Q	S	V	L	L	S	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
E	TGT	GGG	CGG	GGC	TTT	AGC	TGG	CAG	TCA	GTC	CTC	CTC	AGT	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	S	N	K	S	H	L	L	R	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
F	TGT	GGG	CGG	GGC	TTT	AGC	AAT	AAG	TCA	CAC	CTC	CTC	AGA	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	R	D	K	S	H	L	L	R	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
G	TGT	GGG	CGG	GGC	TTT	CGC	GAT	AAG	TCA	CAC	CTC	CTC	AGA	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	R	D	K	S	N	L	L	S	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
H	TGT	GGG	CGG	GGC	TTT	AGA	GAT	AAG	TCA	AAC	CTC	CTC	AGT	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	R	N	K	S	H	L	L	R	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
I	TGT	GGG	CGG	GGC	TTT	CGC	AAT	AAG	TCA	CAC	CTC	CTC	AGA	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAC	GTC	TGC	AGG	GAG
	C	G	R	G	F	S	D	R	S	S	L	C	Y	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
J	TGT	GGG	CGG	GGC	TTT	AGC	GAT	AGG	TCA	AGC	CTC	TGC	TAT	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAC	GTC	TGC	AGG	GAG
	C	G	R	G	F	R	D	K	S	H	L	L	S	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
K	TGT	GGG	CGG	GGC	TTT	AGA	GAT	AAG	TCA	CAC	CTC	CTC	AGT	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	S	W	Q	S	V	L	L	R	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
L	TGT	GGG	CGG	GGC	TTT	AGC	TGG	CAG	TCA	GTC	CTC	CTC	AGA	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG
	C	G	R	G	F	R	D	K	S	H	L	L	R	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
M	TGT	GGG	CGG	GGC	TTT	AGA	GAT	AAG	TCA	CAC	CTC	CTC	AGA	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAC	GTC	TGC	AGG	GAG
	C	G	R	G	F	S	R	Q	S	V	L	L	S	H	Q	R	T	H	T	G	E	K	P	Y	V	C	R	E
N	TGT	GGG	CGG	GGC	TTT	AGC	CGG	CAG	TCA	GTC	CTC	CTC	AGT	CAC	CAG	AGG	ACA	CAC	ACA	GGG	GAG	AAG	CCC	TAT	GTC	TGC	AGG	GAG

```
     C   G   R   G   F   R   D   K   S   N   L   L   S   H   Q   R   T   H   T   G   D   K   P   Y   V   C   R   E
O   TGT GGG CGG GGC TTT AGA GAT AAG TCA AAC CTC CTC AGT CAC CAG AGG ACA CAC ACA GGG GAC AAG CCC TAT GTC TGC AGG GAG

     C   G   R   G   F   R   D   E   S   N   L   L   S   H   Q   R   T   H   T   G   E   K   P   Y   V   C   R   E
P   TGT GGG CGG GGC TTT AGA GAT GAG TCA AAC CTC CTC AGT CAC CAG AGG ACA CAC ACA GGG GAG AAG CCC TAT GTC TGC AGG GAG

     C   G   R   G   F   R   N   K   S   H   L   L   R   H   Q   R   T   H   T   G   E   K   P   Y   V   C   R   E
Q   TGT GGG CGG GGC TTT CGC AAT AAG TCA CAC CTC CTC AGA CAC CAG AGG ACA CAC ACA GGG GAG AAG CCC TAT GTC TGC AGG GAG

     C   G   R   G   F   S   R   Q   S   V   L   L   T   H   Q   R   T   H   T   G   E   K   P   Y   V   C   R   E
R   TGT GGG CGG GGC TTT AGC CGG CAG TCA GTC CTC CTC ACT CAC CAG AGG ACA CAC ACA GGG GAG AAG CCC TAT GTC TGC AGG GAG

     C   R   R   G   F   R   W   Q   S   V   L   L   T   H   Q   R   T   H   T   G   E   K   P   Y   V   C   R   E
S   TGT AGG CGG GGC TTT AGC TGG CAG TCA GTC CTC CTC ACT CAC CAG AGG ACA CAC ACA GGG GAG AAG CCC TAT GTC TGC AGG GAG

     C   G   R   G   F   R   D   K   S   H   L   L   R   H   Q   R   T   H   T   G   E   K   P   Y   V   C   R   E
T   TGT GGG CGG GGC TTT CGC GAT AAG TCA CAC CTC CTC AGA CAC CAG AGG ACA CAC ACA GGG GAG AAG CCC TAT GTT TGC AGG GAG
                            -1       2   3           6
```

Chapter 7 Appendices

Appendix 7.1.

Analysis of NCOs at the *SPRY3* hot spot

NCO data for 14 men assayed at the *SPRY3* hot spot are shown. Parental haplotypes were amplified either using ASPs RAS5286A and RAs3343A (haplotype ‘AA’) or RAS5286G and RAS3343T (haplotype ‘GT’). All positions indicate chrX coordinates, according to NCBI36/hg18.

1. d23

a. Haplotype amplified- ‘AA’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	T → C	154648189	0	6430	0.000	0.000	0.057
rs306882	C → A	154650069	0	6430	0.000	0.000	0.057
rs700442	C → T	154650233	1	6460	0.015	0.004	0.086
rs3870343	G → C	154651570	0	6640	0.000	0.000	0.056
rs306886	T → G	154651636	0	6640	0.000	0.000	0.056
rs6642320	G → A	154652328	1	6640	0.015	0.004	0.084

b. Haplotype amplified- ‘GT’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	C → T	154648189	1	6530	0.015	0.004	0.085
rs306882	A → C	154650069	9	6600	0.140	0.073	0.259
rs700442	T → C	154650233	12	6600	0.188	0.105	0.317
rs3870343	C → G	154651570	0	6640	0.000	0.000	0.056
rs306886	G → T	154651636	0	6640	0.000	0.000	0.056
rs6642320	A → G	154652328	0	6640	0.000	0.000	0.056

2. d25

a. Haplotype amplified- ‘AA’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	T → C	154648189	1	8000	0.013	0.000	0.071
rs306882	C → A	154650069	4	8180	0.049	0.013	0.125
rs700442	C → T	154650233	4	8320	0.048	0.013	0.123
rs306885	G → T	154651506	0	8570	0.000	0.000	0.045
rs6642320	G → A	154652328	0	8570	0.000	0.000	0.045

b. Haplotype amplified- ‘GT’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	C → T	154648189	0	10120	0.000	0.000	0.038
rs306882	A → C	154650069	27	10520	0.258	0.170	0.373
rs700442	T → C	154650233	71	10830	0.657	0.513	0.825
rs306885	T → G	154651506	3	11700	0.026	0.005	0.075
rs6642320	A → G	154652328	1	11770	0.009	0.000	0.048

3. d27

a. Haplotype amplified- ‘AA’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs55997085	- → +	154648157	0	12410	0.000	0.000	0.031
rs67804731	G → A	154648273	2	12410	0.016	0.002	0.058
rs306882	C → A	154650069	6	12560	0.048	0.017	0.104
rs700442	C → T	154650233	5	12590	0.040	0.013	0.093
rs71190340	C → G	154651100	2	12800	0.016	0.001	0.057
rs306885	G → T	154651506	0	12800	0.000	0.000	0.030
rs3870343	G → C	154651570	0	12800	0.000	0.000	0.030
rs306886	T → G	154651636	0	12800	0.000	0.000	0.030
rs6642320	G → A	154652328	0	12800	0.000	0.000	0.030

b. Haplotype amplified- ‘GT’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs55997085	+ → -	154648157	0	12100	0.000	0.000	0.032
rs67804731	A → G	154648273	3	12130	0.025	0.005	0.072
rs306882	A → C	154650069	45	12210	0.371	0.271	0.494
rs700442	T → C	154650233	52	12370	0.421	0.314	0.549
rs71190340	G → C	154651100	4	12710	0.032	0.008	0.080
rs306885	T → G	154651506	2	12740	0.016	0.001	0.057
rs3870343	C → G	154651570	1	12740	0.008	0.000	0.044
rs6642320	A → G	154652328	0	12740	0.000	0.000	0.063

4. d39

a. Haplotype amplified- ‘AA’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs700442	T → C	154650233	21	5730	0.370	0.241	0.560
rs71190340	C → G	154651100	1	5950	0.017	0.004	0.097
rs306886	T → G	154651636	0	5950	0.000	0.000	0.062
rs6642320	A → G	154652328	2	5950	0.034	0.010	0.121

b. Haplotype amplified- ‘GT’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs700442	C → T	154650233	3	5860	0.052	0.019	0.150
rs71190340	G → C	154651100	1	5950	0.017	0.004	0.094
rs306886	G → T	154651636	0	5950	0.000	0.000	0.062

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs6642320	G → A	154652328	0	5950	0.000	0.000	0.062

5. d66

a. Haplotype amplified- 'AA'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs56750753	C → G	154649921	13	10640	0.124	0.066	0.210
rs700442	T → C	154650233	42	11070	0.380	0.274	0.511
rs306885	T → G	154651506	2	12110	0.017	0.002	0.060
rs3870343	C → G	154651570	0	12140	0.000	0.000	0.032
rs306886	T → G	154651636	0	12160	0.000	0.000	0.032
rs6642320	G → A	154652328	1	12190	0.008	0.000	0.046

b. Haplotype amplified- 'GT'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs56750753	G → C	154649921	2	10030	0.020	0.002	0.072
rs700442	C → T	154650233	16	11200	0.146	0.084	0.234
rs306885	G → T	154651506	0	12190	0.000	0.000	0.032
rs3870343	G → C	154651570	1	12190	0.008	0.000	0.046
rs6642320	A → G	154652328	1	12190	0.008	0.000	0.046

6. d72

a. Haplotype amplified- 'AA'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	T → C	154648189	0	9050	0.000	0.000	0.041
rs306882	C → A	154650069	0	9070	0.000	0.000	0.041
rs700442	C → T	154650233	1	9070	0.011	0.003	0.061
rs306885	G → T	154651506	0	9120	0.000	0.000	0.041
rs3870343	G → C	154651570	0	9120	0.000	0.000	0.041
rs306886	T → G	154651636	0	9120	0.000	0.000	0.041
rs6642320	G → A	154652328	0	9120	0.000	0.000	0.041

b. Haplotype amplified- 'GT'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	C → T	154648189	0	9040	0.000	0.000	0.041
rs306882	A → C	154650069	8	9080	0.089	0.045	0.174
rs700442	T → C	154650233	18	9120	0.202	0.126	0.312
rs306885	T → G	154651506	0	9200	0.000	0.000	0.040
rs3870343	C → G	154651570	0	9200	0.000	0.000	0.040
rs306886	G → T	154651636	0	9200	0.000	0.000	0.040
rs6642320	A → G	154652328	0	9200	0.000	0.000	0.040

7. d87

a. Haplotype amplified- 'AA'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	T → C	154648189	0	10600	0.000	0.000	0.036
rs56750753	G → C	154649921	0	10650	0.000	0.000	0.036
rs306882	C → A	154650069	3	10690	0.028	0.005	0.082
rs700442	C → T	154650233	0	10730	0.000	0.000	0.036
rs71190340	C → G	154651100	2	10900	0.018	0.002	0.067
rs3870343	G → C	154651570	0	10900	0.000	0.000	0.035
rs306886	T → G	154651636	0	10900	0.000	0.000	0.035
rs6642320	G → A	154652328	0	10900	0.000	0.000	0.035

b. Haplotype amplified- 'GT'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	C → T	154648189	0	17620	0.000	0.000	0.022
rs56750753	C → G	154649921	8	17660	0.046	0.020	0.089
rs306882	A → C	154650069	53	17720	0.297	0.222	0.387
rs700442	T → C	154650233	85	17870	0.473	0.378	0.584
rs71190340	G → C	154651100	9	18020	0.050	0.023	0.095
rs3870343	C → G	154651570	0	18020	0.000	0.000	0.021
rs306886	G → T	154651636	0	18020	0.000	0.000	0.021
rs6642320	A → G	154652328	2	18020	0.011	0.001	0.040

8. d178

a. Haplotype amplified- 'AA'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs17149475	A → G	154648508	1	7100	0.014	0.003	0.078
rs17149476	G → C	154649621	0	7190	0.000	0.000	0.051
rs17149478	T → C	154649682	0	7190	0.000	0.000	0.051
rs56750753	C → G	154649921	0	7220	0.000	0.000	0.051
s11	GA → AG	154650111	1	7220	0.014	0.003	0.077
rs700442	C → T	154650233	3	7220	0.042	0.015	0.121
rs306885	G → T	154651506	0	7380	0.000	0.000	0.050
rs3870343	G → C	154651570	1	7380	0.014	0.003	0.076
rs6642320	G → A	154652328	0	7380	0.000	0.000	0.050

b. Haplotype amplified- 'GT'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs17149475	G → A	154648508	0	5220	0.000	0.001	0.071
rs17149476	C → G	154649621	0	5220	0.000	0.001	0.071
rs17149478	C → T	154649682	0	5220	0.000	0.001	0.071
rs56750753	G → C	154649921	0	5220	0.000	0.001	0.071
s11	AG → GA	154650111	1	5220	0.019	0.005	0.107
rs700442	T → C	154650233	4	5250	0.076	0.031	0.195
rs306885	T → G	154651506	0	5360	0.000	0.001	0.069
rs3870343	C → G	154651570	0	5360	0.000	0.001	0.069
rs6642320	A → G	154652328	1	5360	0.019	0.005	0.104

9. d240

a. Haplotype amplified- 'TA'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs17149476	C → G	154649621	1	8260	0.012	0.003	0.067
rs17149478	C → T	154649682	1	8260	0.012	0.003	0.067
s649.7	C → T	154649812	2	8260	0.024	0.008	0.087
s92	+A → +G	154650095	1	8260	0.012	0.003	0.067
rs700442	T → C	154650233	0	8260	0.000	0.000	0.045

b. Haplotype amplified- 'CG'

Note: No NCOs were detected on this haplotype.

10. d244

a. Haplotype amplified- 'AA'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
s92	+A → +G	154650095	2	5170	0.039	0.012	0.140
s11	AG → GA	154650111	0	5170	0.000	0.001	0.071
s0.66	C → A	154650655	0	5310	0.000	0.001	0.069
rs306883	C → A	154650301	1	5220	0.019	0.005	0.107
rs700442	C → T	154650233	2	5220	0.039	0.012	0.138
rs306885	G → T	154651506	0	5370	0.000	0.001	0.069

b. Haplotype amplified- 'GT'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
s92	+G → +A	154650095	0	11720	0.000	0.000	0.032
s11	GA → AG	154650111	0	11720	0.000	0.000	0.032
rs700442	A → C	154650233	5	11760	0.043	0.019	0.099
rs306883	A → C	154650301	3	11790	0.026	0.009	0.074
s0.66	T → C	154650655	0	12040	0.000	0.000	0.031
rs306885	T → G	154651506	2	12320	0.016	0.005	0.059

11. d252

a. Haplotype amplified- 'AA'

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	C → T	154648189	2	13100	0.015	0.005	0.055
rs67804731	A → G	154648273	5	13100	0.038	0.017	0.089
rs56750753	C → G	154649921	0	13100	0.000	0.000	0.028
s11	GA → AG	154650111	0	13100	0.000	0.000	0.028
rs700442	T → C	154650233	6	13100	0.046	0.022	0.100
s0.25	- → +	154650253	0	13100	0.000	0.000	0.028
rs71190340	C → G	154651100	0	13100	0.000	0.000	0.028
rs306885	G → T	154651506	0	13100	0.000	0.000	0.028
rs6642320	G → A	154652328	2	13100	0.015	0.005	0.055

b. Haplotype amplified- ‘GT’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs68008496	T → C	154648189	2	13230	0.015	0.005	0.055
rs67804731	G → A	154648273	1	13230	0.008	0.002	0.042
rs56750753	G → C	154649921	0	13300	0.000	0.000	0.028
s11	AG → GA	154650111	2	13300	0.015	0.005	0.054
rs700442	C → T	154650233	2	13300	0.015	0.005	0.054
s0.25	+ → -	154650253	1	13300	0.008	0.002	0.042
rs71190340	G → C	154651100	0	13300	0.000	0.000	0.028
rs306885	T → G	154651506	0	13300	0.000	0.000	0.028
rs6642320	A → G	154652328	4	13300	0.031	0.012	0.077

12. d254

a. Haplotype amplified- ‘AA’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs17149476	C → G	154649621	0	7090	0.000	0.000	0.052
s649.8	G → A	154649882	0	7090	0.000	0.000	0.052
s92	+G → -A	154650092	1	7090	0.014	0.003	0.079
s11	GA → AG	154650111	1	7090	0.014	0.003	0.079
rs700442	T → C	154650233	0	7090	0.000	0.000	0.052
rs71190340	C → G	154651100	0	7090	0.000	0.000	0.052
rs306885	G → T	154651506	0	7090	0.000	0.000	0.052
rs3870343	G → C	154651570	0	7090	0.000	0.000	0.052

b. Haplotype amplified- ‘GT’

Marker	Type of events	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs17149476	G → C	154649621	0	7090	0.000	0.000	0.052
s649.8	A → G	154649882	1	7090	0.014	0.003	0.079
s92	-A → +G	154650092	1	7090	0.014	0.003	0.079
s11	AG → GA	154650111	1	7090	0.014	0.003	0.079
rs700442	C → T	154650233	1	7090	0.014	0.003	0.079
rs71190340	G → C	154651100	0	7090	0.000	0.000	0.052
rs306885	T → G	154651506	0	7090	0.000	0.000	0.052
rs3870343	C → G	154651570	0	7090	0.000	0.000	0.052

13. d274

a. Haplotype amplified- ‘AA’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs67804731	A → G	154648273	0	6890	0.000	0.000	0.054
rs17149475	A → G	154648508	0	6890	0.000	0.000	0.054
rs17149476	G → C	154649621	0	6890	0.000	0.000	0.054
rs17149478	T → C	154649682	1	6890	0.015	0.004	0.081
s9.72	A → G	154649721	1	6890	0.015	0.004	0.081
s9.81	C → G	154649811	0	6890	0.000	0.000	0.054
rs56750753	C → G	154649921	2	6930	0.029	0.009	0.104
rs700442	T → C	154650233	9	7050	0.131	0.068	0.242
rs306885	G → T	154651506	0	7670	0.000	0.000	0.048

b. Haplotype amplified- ‘GT’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs67804731	G → A	154648273	0	13450	0.000	0.000	0.027
rs17149475	G → A	154648508	1	13450	0.007	0.002	0.041
rs17149476	C → G	154649621	0	13490	0.000	0.000	0.027
rs17149478	C → T	154649682	2	13490	0.015	0.005	0.054
s9.72	G → A	154649721	0	13490	0.000	0.000	0.027
s9.81	G → C	154649811	0	13530	0.000	0.000	0.027
rs56750753	G → C	154649921	1	13610	0.007	0.002	0.041
rs700442	C → T	154650233	3	14720	0.020	0.007	0.060
rs306885	T → G	154651506	5	15330	0.033	0.014	0.076

14. d279

a. Haplotype amplified- ‘AA’

Note: No NCOs were detected on this haplotype.

b. Haplotype amplified- ‘GT’

Marker	Type of event	Position	NCOs	Molecules screened	Frequency of NCOs	Lower 95% C.I.	Upper 95% C.I.
rs17149475	C → A	154648508	0	4950	0.000	0.001	0.074
rs306883	A → C	154650301	1	4950	0.020	0.005	0.113
rs71190340	G → C	154651100	0	5000	0.000	0.001	0.074
rs306885	T → G	154651506	0	5000	0.000	0.001	0.074
rs3870343	C → G	154651570	0	5000	0.000	0.001	0.074
rs6642320	A → G	154652328	0	5000	0.000	0.001	0.074

Chapter 8 Appendices

Appendix 8.1

Annotated sequence, ChrX: 154626392-154639892

Note:

- Each line of the sequence consists of 60bp, coordinates to the left of each line indicate ChrX, NCBI36/hg18 coordinates for the first base of the line, whilst coordinates to the right represent local positions assigned to the same.
- Generally, repeat sequences are indicated in red. Repeat types are indicated near the repeat start position.
- SNPs in single-copy DNA are indicated in **blue/bold**, whereas those within repeat sequences are in **purple**. 'rs' numbers are written above respective SNPs. rs-numbers in bold indicate markers that have been genotyped as part of the HapMap project. SNPs identified through donor re-sequencing are highlighted in purple and donors wherein these SNPs were identified are also indicated. All SNPs except ones indicated were typed using ASO hybridisation to dot blots.
- Primer sequences are underlined, primer names are written above the sequence. Primers with 'F' at the end of their names are forward primers, whilst those with 'R' are reverse primers. Where two primer sequences overlap, sequence for one primer is indicated in bold for clarity. Primer names including FAS and RAS indicate forward and reverse allele-specific primers respectively.

```
          SINE (MIR) ->                                 rs3912967C/T
154626392 TTCTTAGTAACACTGTGAGGGTAAAATATACATATAAAAATGCCTTAGAAATGTAAAGAC 1
     rs3912966C/T
154626452 CTGAGAGTTTTTGAGCAAACCAGAAAGTTATCGAGAACCTTTCCTTTTGTGCAATTCAGT 61
154626512 ATATTTTGGGGGGTAAATTTTAAACACCAAATAATCCTCAACAAACAAATGTACAAATAG 121
      rs3912965C/T
154626572 TTCCTGAAATGAGGCAAATATACTCCTCCTGGAACATAAAACAATTCAAATTTCCAACAA 181
                                                    LINE (L1) ->
154626632 ATTTATCAAATATGTAAATAAGTGTTGATAACTTGAAAAACAAAAAAGGAATAAAATCAT 241
154626692 GATACATCAACATGAATGAACCTTGAAAACATTATGCTAAGAAGCTAGACATAAAAGGAC 301
```

```
154626752 AATTTTAAATAATTCCATTTATATATTTATATTAAGAGATAAATGTGTTTTCATATCCCA 361
                           26.8F ->
154626812 ATTTTCTCCTCATCTATTCATAAAAGGCAAGCTGCCCAAAGGAAGCATGCAACACTGCAA 421
          26.87F ->
154626872 GGAATGCAATTTAGTGTGGCTGGGCAACTTTGGCTGGAGGGGGTACAGAAGGGAGAAAGA 481
154626932 AGATGGAAGACAGTCTAAGACTCAAGGCATCTTGCGACTCTGAGACTTCAGAGGTGAGGC 541
154626992 CACCTATGAACTGGGTGGATGTTTGGGAAAACATGTAATTCAAGGCAGTTCTGAAACTCT 601
                         rs34496779-/G
154627052 GTGGGTTTTGTGTGGATTCCGAGGTAATCCTACAGTAGAAAGTCCAGCCAAATGGTGAGA 661
154627112 ATAAGCAAGAGTCTTGGAGAAAAATGAACAGAATGAAGGCAATGTCCTACCATAATGCAT 721
                                          FAS0568->
                                         rs5940568G/T
154627172 GCTCTTGCAAGGACAGACAGACACCATGATATCCTTAGGATCTCCTGAGGCAGTTTGTGT 781
154627232 AAAAGGAGGCAACTTGCCACCAAAGTGGGAAAGAAGTGGTTAAACAGCAACAGTAGGGCT 841
154627292 CCTGCTTCCTGGAGCTGTTGGGTGACATTTCCCTGCAACAATCCTAGAATCTTAAAGTGT 901
154627352 GGTGTCGGTTTCACTTCGGATTCCAAACGGGAAGGGACTTTTCCTGGCAAACGCAGTTTA 961
154627412 AAGTGTTTCTTTCAGTGTCATGCTGTACAGAAAAATACTAATCCCACAGAGAAAATTAGT 1021
154627472 CACAAAAACAGAAGGAAAAAAATACAACATGAGTGCAGACCCAGGACACTAAAGTCAGAG 1081
154627532 AGACAGAACCACAGCCTCGCACAACCATTGCTCACATTGTGTAAAGCATTACTGGAAGGA 1141
154627592 AACAAAGTCTTCAATAGAAACCTGCTGCACCACTCTTAGCAAACTCCTTTTGTTTGACCT 1201
154627652 GGGGCCACTTTACCTGTGGATTGTTTTGTAGCAGATTTTCCAGGTAATGCTTAGATTTTT 1261
154627712 CATGCTCCCCCAGTTCATTGGATTGGGCTTCAGGGGTTCTTTCTGGTGTGTGTCTGATTC 1321
154627772 AGATTGTATTTTAGGAATACTTAGAATTGAGGAAGAATAGGATGCTAAGCTACTGAGTGC 1381
154627832 AAAGCCTCTTTACTAACTTCCAATAATATCTGAACTCATGTCTAAGGAACATAGAAGCAT 1441
154627892 CTTTCAATTCCTAGTCTCATTTTCTCACCATTCAAATGTTCCTTAGGGAGAATAGAGATC 1501
154627952 CTATGCAAGTCTCAGGAATTGCTCTAGAGAAAACTATCTTTCCCCAAGCTTGAGAAACTT 1561
154628012 ATCCCAGACCAACTTTCTTTCCCCAAATTTTGTTTAAATTAAAAATAGGCAACTAGAGGG 1621
154628072 GGTCAGCTAGTTGGTGTTCCATAGATGCTTTCCTTGAAGCAGGGTAAAGAGCTTTGATTT 1681
154628132 ATTCTTAATTGTTTCTGTCATTTAAGGGTGCATCAGAGACCCTGTTGTAATATTAAATTA 1741
154628192 TATCACTATTGAGGAGCCCATAAGGCTGGGGCCGGTACAAGGATCATCATGAAGTGCAGC 1801
154628252 TATTGAGAGCTACTGACAACTTCTTAGAGGTGTGGCTGAAAGCAGCTAGCAAGAATATAA 1861
154628312 GAGACAAAATCTTGTGTTGAGGGAAGTCACCTGTGAGAGCAGCAGGTAGATGTTGAGGCT 1921
154628372 TGGGAGACCTTATTGTTCTAAGATACCTGGCCTTGTCTTCCAAACAAAAAATCATCCTTG 1981
                                                                SINE (MIR) ->
154628432 ACATGCAGAAAGACTACTGATTTCTCAATAGATTTATAATGGAACAGATATAATGATGTG 2041
154628492 ACCTGGAGCAATTTCTTTAACCTCATACAGAATCAAATTCCCCATCTGCACATAAAAGTA 2101
154628552 TTTACCTCATAGGTTTATTGTGAAGCTTAAGATAAAATATGTAATTTACCTTGCATAATG 2161
             rs28512393A/G
154628612 CTTGGCATATAGTAAGCACCGAATATACATTAGTTATTAATATTGTTATTTTCATGAGCA 2221
154628672 AACCTAGTAGTGATAAAAGCTCAGAAACACTGGGTAGGAGCTTGAGCTAGCTACAGAGAA 2281
154628732 CTTGAATTGTAACCTGGCCTTCAGATTCCTCATCAGAAGGGTGGGACCAGGAATAAGAGC 2341
154628792 GAGCGTTACACTTGGCTCAAGTCCAAACTGCATGACTGCACAAACACTGTAACTGCTCTG 2401
154628852 TATTCAGACACATTTGAAAGGATTTTTCCAGCATAAATTTCTAAATGTGCTGCTTGTTTT 2461
154628912 GTAATTCAATCTAGGGCAGCAGCAGCAGTGCTATCATTATTTACATTTCTAGCTTTTGGT 2521
                                                   rs35301922-/TTC
154628972 CTTGGATCCAAAATATTTGTAGCAGTGTGAATTGGCTCATAAGAAAATTCTTCTCTTTTC 2581
                                             rs28638224A/T
```

```
154629032 TTCAAAAAAGTTTCATGTTACTTTCTTTTATACCTGTAAGCAAGTAAATATTAATAATCT 2641
154629092 TTTACATTGATTTGTTTCTGAGACATTTGGTCAGAAATAATTTGCAGAAATAAATTGATT 2701
154629152 CAGACACTGTAGCTAATTTAACATTAATTTTTTCCTATTTCTTTGATTTTTAGAAATTAT 2761
                                              FAS8165 ->
                                            rs67398165A/G
154629212 TCCAGTAGAGGAGAACCTCATCAGCAACATATTAACTTTCTGATATTGTTAATACAATTT 2821
154629272 TTTCATTTTAATGAACAAAAGGCAATACAAAGACCTTGTAAACAATTAGAAAATCATCTT 2881
154629332 GCCCTCAATTAAATTTTCTATTCTCCAAGATTTAAATTTTCAACAGTAACCTAAATAGCT 2941
          29.3F ->
154629392 GTTAACCCTCTATTCATGAGTGTAAAGGGGAGAAAACATTTTTGTTTATGTATATGTCAA 3001
                              29.4F ->
154629452 GTCTTAAGATTGTTTCAATAGTGACCAGTTCCTTTTCAGTCTTTCCTTTACCTTGGTTGT 3061
154629512 GTTTGAGTCTTAAGTTCTCGGACATTAATAACTGCACATTGGGTACTCAAATTCTGGAAG 3121
154629572 GACTAACCTGGCTGATCTATAATGGACAGCTGTGCATGACCCCAATAAATGAACTCCTGC 3181
154629632 ATAATCTGCAACTAGTCTAGACTACAAGCAGTAAATAGTTGTTGAAGGATGAATGATGGA 3241
154629692 ATGGTCAGGACTTTTCTATGTAAGAAGGCTACTGAGAGTGTTCACTCCACACATAAGGTT 3301
154629752 TCCAGGTACTATATGTTTCTATTTCTGTCTGTATTTCCATTTACTCTTTCTTTAAATCTT 3361
154629812 TTATTGTCTCCTCTTGGCTCCAAACTGTGGATCTGGGACCACATGCGACTCCAAGTAAAC 3421
154629872 TTTCAGCCCTCAAATGACCTCAGTATTACATCCTTCCTGAAGTATCGCAGAATTTTTCTC 3481
154629932 TAAATTAAAGGAGAATCACAGGACGTCACACTTAATGACTGCTGATTCAAAATTAATTTT 3541
                                            LINE (L3) ->
154629992 CCTTCCTTTGATGCCTAAATCTACTTCAGCACATAATTATAGAATTCTGCTTTGGAAGGG 3601
                            rs6642234A/C
154630052 GCCCTAGAGCTTAATCTAATTTTCAACTAATTGCAGGATTCATTATTAGAGAACCAATGA 3661
                                    rs306899A/G
154630112 TAGTTTGTAAATCAGTCTTTGAGCATTTACCTGATAAAACTATTCCATCAGTAGTTAGAA 3721
154630172 ACTATTCCATTAGTACTTAGCTCTTAAAATGAGCTAAGACCCACCTCCTTAAAACTTCCT 3781
154630232 TTGAGCAAAGCTTCATGGCCCCACAAAAGAAAAGATTTTCCTATTTTCATTTAAAATATC 3841
154630292 ATATTATTAAATTATCTCTAAATGGCAAATATAGCATATTTTAACAGGAAGTATTCCTAG 3901
154630352 TGGTTCTATCTCCATTAATCCATACTCCAAAATATTGCCCACTCTTTTAAGCTACCAGCC 3961
                                                 rs306898A/G
154630412 CTATGCAGTATCCCAAATATCAGTGCCAAAACCAGAAGTGCATTCTCCTATTACATTAAT 4021
154630472 AATAATTTTACAGTATGGTGGCAGGAAATGCACTTTCAGTTCACTGGCTATAGCTATTAA 4081
                           <- 30.5R              <- 30.58R
154630532 ACTATGAGGAAATGACTGCCTGGAAGGCAGCCAGGTCCTTGGCTTAAAGGCCAACTCTGA 4141
                                                     rs306897A/G
154630592 GGAAGACAAAGGTAATACAAGAAGGGAAACTATGCCACCTGACCAAAAACGTTATTGAGA 4201
                    DNA (MER5C) ->
154630652 CACTCTAGAGCAGTGCTTCTCAACTCTTTGTGGTGAAGGAACTTTTTTTTTTTAATTTCC 4261
154630712 AATCGGCCATGGACTTTTGTAAAATACGAAATGATGCACTTGGATAAAGCAGCAATGTGA 4321
154630772 AATGCTCTAAAAATGTCTAAATACTTACCATTTGTATACTTGATCAGTATAATTATATAC 4381
154630832 TTATACGATTTATATACTTATTATACTTATCTCATAAGGGACTGGTAACAAACAGTTTTG 4441
154630892 GAACTGGTGTGTGGAACACATGCTGAATAGCACTGTTCCAGAGAGGGAGAAAGGATGAGA 4501
154630952 ATTACATTTGCCATCCTATTCTTTTAGTATGGTTCAAGTGACATGCTAGAAAGGAACTAT 4561
154631012 TCCTGCTGCTGGTAAGTACCAGAGATGTGTGTATGTTAAACATAAGATTGGTGGCCCTGC 4621
            LINE (L3) ->
154631072 TTAACAAGATAGTAAACACTTGGCTATAAAGATAGTATCGATGAAAACAATTGCCATTTA 4681
```

```
154631132 TAAACCATGTTTAAAGTAGTTTAATGATGAAATCTTGGCTAATCATATGAATATGGGGAA 4741
             31.1F ->
154631192 AAGTATGTTTGCCCAGAAGTAGGTGGTATCAAAAAGCATTTTCTCACAGTTATGTAATTA 4801
                     31.2F ->
154631252 CCAAGAATAGAGTTTAGGGAGTGACTCTACCAAGCAGGCAAAAGAGAGGACTTCCAAGAG 4861
                                                            rs3916103G/A
154631312 AGTAATTTTGTCAGCAGAAACCTGGAATAACCCCAAGGTCAGAGGTAGTAAGGAGGGAAA 4921
154631372 GGGACTTGACAATAGGATGATAGATAGTATATTTCATTAAAGGAATGAGGAATAGTTACA 4981
154631432 TGAATTCTTCCTCTATAAATGATGTTTGCTCCAAGACGAAGACTTCTAAAGGAAGAATTC 5041
154631492 TGACACTGAAGGCTATCATATTGGGAATCAGGGCCAAAGAACAGGGCAATACCTCATGGA 5101
154631552 ATGTCTGACTGCTACATGGAATCAGAAGACAAGCTAACCATCCACCATGAAGGGGAAATA 5161
154631612 CTGGCTTTTCTCTCTTTCCACGTACAATAAACAAGAGATAGGGCAATCAGATTCTGCATT 5221
154631672 AATTGCCAAGGGAACAAGGATTTTTCACATGTACAGAAAATATGAATTCCAGCTAAAAAC 5281
154631732 ATTTTCTAATCTAGGTCTTCATATGTAAATGTAAACAGGCAACCGAGAATCATCAAATTG 5341
154631792 AGTGAAAACCAAAAACATTAGATCCACCAGACTCAACCAACACAATAACTAATGGTTAAA 5401
154631852 GAAATAGTATGTGGAGAAAACAGAAGTGAACTATATGGATAAATAGGATTTAGAGAAAGA 5461
154631912 TTTTTCACACTCAATCACACAATGCCCAATACATAATTAAATACCTTGGATCCATAAAAT 5521
                                   rs3909635C/T               <- 31.9R
154631972 AAAAATAAGCTTTTACAAAAAAGGAAGTAATTCAGAGTTCTTAGCTATAAGCAATATAATC 5581
                 <- 32.0R       LINE (L2) ->
154632032 CACTTGGGCTTTTCTAAAAGTAAACAAATACCTAATGAGTGGTAACTATGTGTTGGTTAT 5641
                                              rs35177475-/C
154632092 TCTTTTAGGCACTGGGGATACAGCAGTGAGTAAAATAGACATTCCCTGTCCTTAAAGAGC 5701
154632152 TTATATTCCAGTCAGAGGCTAGAGATTATAAGGAAATAAACATATGATTGAGGTAGTGAT 5761
154632212 GGCAGAGTAAGGGAGACAGCAAGTGACAGGGGTGTGTGTAAGTGTGTCTGTGTATGCATG 5821
154632272 TACGTATTATATAGATAGAGTGGGAAGAGAAGGCCTCTCTGATAAAGTAACATTTGAGCT 5881
                                               rs28725321A/C
154632332 AGGAATCAAAGGAAGTGATGAGGTGAATCAAGAGCAAGAGAAAAATATTCTGAGCAAAAG 5941
154632392 GAACACAAAATAAAAAAAATTTGAGGTGGAAGTATCTGGGGGTTTTGACTAATAGCAAGG 6001
154632452 ACACCAGTTTGGCTTCAGTGGGCTGTGTGAGAAAGAGTGATAGAAATTGAGGTGAGGGCA 6061
                          rs306896C/T
154632512 ATAAGAATAGGCCAAATTATATAGGACCTTGTAGACTATTGTATTTTGGCTTTCACTATG 6121
154632572 ATTTTTGCCTGGGGAACTAGAATAACTAAGGTACTGTTTATTATAATGGAGAAAATTATA 6181
154632632 GGAGAATCAGTTAATTTTTGTTGGAGAATGAAGAGCTGAAGTTTTGGACATGTTAACTAG 6241
154632692 GAGATGTGTTTTAAATATTCAATTGGAGATTTTGAATACCCAGTGGGATACATTATTCTG 6301
154632752 GAGTTCAGAGGAGGGATCGAGTCTATAGATATTAATTTGAGAGTTATTTAAAGTCATGAG 6361
154632812 ACTGCATGAGATCAGCTAAGGAATGAATGTAGATAAACAATAGAAGAGGACCAAGGATTG 6421
                                                              LINE (L1) ->
154632872 AGTCTTACATTACTTCAGTATTTTGCAGTTGGGAATTAAAATATCACCAGAATACATATA 6481
154632932 CACCATAGAATACTACACAGTCATAAAAAAGAGTGAAATCATGTCCTTTGCAGCAACATG 6541
154632992 GATGCAGCTAGAAGCCATTATCCCAAGCAAACTAATGCAGGAACAGAAAACCAAATACCA 6601
                  rs9657865C/A
154633052 CATGTTCTCACTTACAAGTGGGAGCTAAACATTGGGTACACATGAACATAAAGATGGCAA 6661
154633112 CAATAGGCACTGGAGACTAGTGCGGGGGGAAGGGAAGAGTGTAAGGGTTGAAAAACTAAC 6721
154633172 TGTTGGGTACTATGCTCACTACCTGGGTGACGGGATCAGTTGTACCCCAAACCTCAGCAT 6781
154633232 CACGCAATATACTCATGTAACAAACCTGCACATATACCCTCTGAACCTAAAAGTTGAAAT 6841
          LINE (L2) ->
```

```
154633292 CATTTAAAAAAATGAATTTAAAAATATCACCAGAGGACATTGAGATGAAATAAATAGTGA 6901
154633352 GATAGGAAAAGTATAAAGAGGCTACCTGAAAGCCAAGTAAAGAAAGTGTTTCAAGAAGGA 6961
154633412 GAGAATAATCAACCATGTAAATGATACCAACAGGTTAAGTAAGATGAACACTGATAATTA 7021
154633472 ACCTTTGAATTTAGCAATATGGAGATCATTGGTAATGTCAGCAAGGGTAGCTTGTAAAAC 7081
154633532 AATGTCATCAGTTAAGAGTGAGAATGGAGTACTAGATTTTGGAGGTTTAAGGAGAGAGCA 7141
154633592 GATGGTATAAAATAGACGGCTGGTGGGTTATAAAAATGTCACAGCATTGCCAAGCAGCTC 7201
154633652 GAAAAGTCCACTTGAGGTTAATGGTCATGAATTTAAAGAGCCTAACATTAAATGAATACC 7261
154633712 AAACCAGTTAACATGGAAACAGAAAAGGTGGAAAGTTGGATTTAGGCATGTTTGGATTTT 7321
154633772 TTCAGAGGTGACTATGATGAATTGAAAGAAGGGAAAAGAAGTTAATGGTATATGCTTGGA 7381
154633832 AGTGTAATTGTAATGATGAACCATGAAATCTAAAGTCAGGTAAGAAGAGAAGTGAGAACA 7441
154633892 TGTGTAGAGTGAATGGCAGTAAAAAGTAATAGGGGTCAATGAATTTGAGGTCCCAGTAGG 7501
154633952 TTCAAGAAAATGTTGGAATTGGGGCATTAGAAGGAGTGAGCAGAAAAGACAGGAAAAGAT 7561
                                                                     LTR ->
154634012 GGCTTGAAAAGGGGATGCTGAAATTGAGATTGTGTAAGTAGTATAGCTATTTATTAAAAG 7621
154634072 AAGGCTGGAAAACCAAGCTTGAGGTTCAGCTTCCAGAAATAATGTCCAGGACCATTGCCA 7681
154634132 CAGCACTGCCCTGATAAGAAAAACTGTTGCTCCTACAATTAGATGCCAAGAAATTGGCTT 7741
154634192 TACTGCAGCAGCTACTGCTGCCAGCACTGATGCCACTACTTTATTAAGAACTTGACCCTG 7801
154634252 CAATCTCTGCCATCCCCCAAAGTCAAGTCCCTTTGCTACTACCTATTCCAACAATGGAAA 7861
154634312 CCATACAGTTCTACCTTCTGCATGGCTTCAATTCTGAATCAGTCTCATGTCAGGGCATCT 7921
154634372 TACTGGTAGAACATAAGTCATATTCCTACATTGTAGTTTCAACACAGACGTTTGAGTTGT 7981
154634432 GACTTCTACATTTAAGAGGTAAATATTATAATGTAAGAATTTTTTCCAAATATAGAAAGG 8041
154634492 CTATACATGATATGTTGGTCAGACACAAATATGACAAATGTTTACTATATAGATTTTTAA 8101
                                      34.5F ->            34.55F ->
                              rs35654209-/T
154634552 AATTAAAAATACAATTGCCAAAATATTTCCCCTTACTAAAGGGGTGACTAGTGGTATGTA 8161
                                   <- 34.6R        <- 34.67R
154634612 CACAACTGAAAACAGAATTACTTTGTTGAAGATTGGACATAGTGATTCTGCCAGGATGCA 8221
                                                                    LINE (L1)
->
154634672 ATATAAAAAAAGGAAATTGTATGAAAAAAGTAAGAGACATATAGATATTTATGAGAATCC 8281
                                                             (TAAAA)n STR ->
154634732 CAATATCTGTTTAATGAGAGTTGGAAAGAAAAGGAAAGCAAGGGAAGGGAAATAAAATAA 8341
                         ASOs not ordered
                        rs55931901+/-, rs56237601+/-, rs58677056+/-
                                                      rs57809158-/ATAAA
                                                             LINE (L1) ->
154634792 AATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAGGAAAGGAA 8401
                                         rs9645287C/G
154634852 AAGGAAAGGAGGGGATGAAGAAATCAAATAATGGACTAAAATTTCCTTAAATTGAAAAAG 8461
154634912 GACTTAAGTGTTCAGATTGAAAGAGCCCACCAAGTCTCAAGTAGCAACAATGGGAAGAAT 8521
154634972 GACACATGTCTAGACAGATTCTGGTAAAAATTTTAAACTATAAAGTCAAAGTGAATATCC 8581
                                                                  LINE (LI) ->
154635032 TCTTCCAGGCACCTTGGTGTAGCACATAAGCTACATATCCCTATGGGGTAGCTCTATTGG 8641
154635092 ATCGCTAACTCACCCCTTCACCAAAATTAACTCAAATGGATCAAAGAGCTATCAAAAAAT 8701
                                         rs61461091-/A
154635152 ATGAAATCATCAAAAGATAAGAAATACTAAAAAAAAAAGCACAAGGAGATAAGGCACAAA 8761
154635212 ACCTAGATGCTATAATTCCATAATGGAAAACATGATCGATGAGTTTGACTGCATAACAAT 8821
```

```
154635272 TTAAAACATTTGCAAAGCATGGAGTATTATAAGTTAATTAAAAAGACCAATGGCAAAGTG 8881
                                                  ASOs not ordered
                                            rs73249626A/T and rs2742288A/T
                                                  rs71278020AACTA/TACTT
                                            rs73249627A/T and rs2742289A/T
                                                               T (R37 and R72)
154635332 ATAAAAAAGTAATGTTTATGCTATAGGTTTAATGTCTTGATATATATATATATAACTATT 8941
                  <- RAS6487
LINE (L1) ->     rs58176487C/T
154635392 GATATTTAAATGTATCTATTAAGTTCTTACAGATCCATAATAAGAAAAAATAGAAAAAAT 9001
154635452 GGACAAAGAACAAGAAAATACAGTTCTCAAAATGAGACATATAAATAGCCAATATTTATA 9061
154635512 TTTAAAAGTATTGAATCTCACTCAAATTTTTTAAAACTCAAGATATCATTTTTTACTTCT 9121
154635572 CAGATAGAAAAAGGTTGAGAAATTTGATAATATACTTGGTTGGCAAGAGTGTGAGAAAAT 9181
154635632 GAGCATTCTCCTAAACTCCTTACCAGTAAGCTTTTTGGAGAGCAATTTGGCAATATTAAT 9241
154635692 CAACATTTCAAAACATGTATTTGTCCCAGCATTTTCTCTTTAGAAATTTGTGCTATAGAT 9301
154635752 ATATTAGCACAAATATCCAAAATACATGTAAAAGATATTTATTGCAGCATTTTTGTAATA 9361
                                                            35.81F ->
154635812 CCAAAAGTCTGAAAACAATCTAAATATTTATTAGTTAGAGACAGGTACTTCTGGCAAAGA 9421
                                     <- RAS6895
                                     rs306895A/T      <- 35.87R
154635872 GAAGAGTCAAATATAACAGGAGTCTGTTTGCATGAGGACTAAACAGTCTCCCAAATGTAT 9481
            LINE (L1) ->
       <-RAS0570
       rs5940570A/C
154635932 TCTTTAAAAAACAGCTTTATTGAAACATAATTCACATGCAATAAGTCACCCATGTAATGT 9541
154635992 GCACAATTTAATAGTTTTTAGTATATTCTCAACGTTGTACAACTATCACCACTATCTTAT 9601
                                                      rs55996406A/C
154636052 ACCATAATGTTTTCATCAACCCAAAAAGAAACCCTGTACCCATTAGAAGTCACTCACAAT 9661
154636112 TTCTCCTTTCCCTCAGTCCCTGAAACCACAAATTGACTTTTCTTTGTGGATTTGCATTTT 9721
            rs55764735-/T  rs55720186-/T
154636172 CTGGACATTCACATAAATGAATATTATAATATATGACCTTTTGTGTCTGGCTTCTTTCAT 9781
154636232 TTATTGTGATGTTTTCAAGGTTCATTCATATTGTAGGATGTATCAGTGCTTCGTCCTTTT 9841
                                         rs306894C/T
154636292 GTGTCTAGAAAATATTCCATTGTATGGATAGGTCACGTTTTGTTTAGCCTGTCCTACTGA 9901
154636352 ACAATCTTTCAACAATTATTCTGAATAATCCTATTGAACATTCATGTACATGTTTTTGTA 9961
154636412 TAGACACATGTTTTCAACTTTTTTGGTTATATATCTAGGAATGGAATTACTAGATCATAT 10021
                                                  36.4F ->
154636472 GATAACTTCATGTTTAACATTTCAAGAAAATTCCAAACTGTCTTCCAAAGTGGCTGCACC 10081
154636532 ACTCACAAAGTGCTGTGAGATTTACAATCTCACCAGCAGTGTATGAGGCTTCCAATTTCT 10141
154636592 TCACATCCTCACTACCACGTGTGATTGTCTGTTGTTTTGATTTTGGCCATTCTAGTGGGT 10201
154636652 ATTAAGTTGTATCTAATTTTGTTTCTGCTTTGCATTCCCCTAATGACTGATGTTATTGAG 10261
154636712 CATGTTTTCATGCACATATTGGCCATTTGTGTGTCTTTGAAGAGATGTCTATTCAAATCC 10321
          rs306893C/G
154636772 TTTGCGCATTTTTCCATTTGTTTCTTTGTTTTAGTTGTTGAGTTGAAGATCTTTATCTAT 10381
154636832 TGTGGATACAAGTTCCTTATCAGATATATGACTGGTAAACATTTTCTCCTATTCTGTGGG 10441
154636892 TTGCATTTTCACATTATTGATTATATTTTTGGCAGGAAAAAAAGGTTTGATTTTGAAGA 10501
154636952 AACGTAATTTATCTATTTCTTCTTTTGTCACTTGTGCTTTTCTTGTTGCGTCTAGGAATA 10561
```

```
154637012 CTATGCCTAATCCAAGGTCACAAATATTTACTCCTATGTTTTATTCTAAGAGTTTTGTAG 10621
154637072 TTTTGGCTCATACATTTATGTCTATGATTTTGAGTTAATTTTTTGTACAGCATATGATTT 10681
154637132 GAAGTAGGAGTCCAACTTCATTCTTTTGCATGTGGCTATTCAGTTGTTGCAGCACCATTT 10741
154637192 GTTGAAGAGATTATTCTTTCTCCATTAAATTGTGCTGACTCCCTTGTAGAAAATCAATGG 10801
154637252 ATCATAAACGTAAGGTTTTATTTCTGGACTCTCAATTATATTACATTTGACTATATGTAT 10861
154637312 ATCCTTATTCCAGTACTATACCATCTTGATTACTGTAGTATTGTAGTAAGCTTTGAAGTC 10921
                                                                      rs35418317-/A
154637372 AGAAAGATTTAATACTTTTATTTTGTTCTTCCTTTTCAAGATTGTTTTGACTATTCTGTA 10981
                                                 rs35911738-/T  rs5940408C/T
                                                                      rs28411049A/G
                                                                      (not ordered)
154637432 CCCCTTGCATTTCCATATGAATTTTAGGGTTCATTCATGAATTTTTGCAAAAAAATCTGG 11041
                               rs306892A/T
154637492 CTGGGACTTTGATAGGGATTATGTTTTCTGTAGATCACTTTGGAGAGTATTTCTATCTTA 11101
                                                             rs5940409C/T
154637552 ACAATATTAAATCTTTCAATCTACGAATATGGAGTGTCTTTCCATTTATATAGGTCTGCT 11161
154637612 TTAACTTCTTTCAACAATGTTCTGTGGTTTTCAGTGTATGTCTTGTATTTTTTTGTTAAA 11221
            rs184257T/C
154637672 TGTATTCTTATTTTATTCTTTTAGATGCTATTGTAAATGGAGTTGTTTTCTTAATTGCAT 11281
154637732 TTTCAGGTTGTTCATTGCTAGTGTATAGATATACAACTTACTTTTGTATATTGATTAATC 11341
                               (not ordered)
                              rs34243947-/TG               rs306891A/G
154637792 TTGTATATTGCCATGTAGCTGAACTTGTGTGTGTGTGTGTGTATTTTAAATAGGGAATTT 11401
                                                                      <- 37.85R
154637852 GCTATGGAAGTTTGTTACAAATAGATATATATAATTGCAAAAATAATTTTGTGTGCATAC 11461
154637912 TCACACGTGCGTGCACACACACACACACACCCTGTAGGCTGTTGGGTAACAGCAGTGGAA 11521
154637972 AAAATAAGTAAAAAGTCTAACTGGAGAGTAATGAGATTACTAGTGCCTAAATCTCAGAAT 11581
154638032 TCTTAGTCAAAACTTGCAGTAAAAAGTGAAAGTGAACACATCCCAGTTTAATCTTTTACC 11641
154638092 CCAATGTCTTTAAATTGCTTCGCATTTTACTTTATTAAAAGCAAAAGTTAATCTTCTTTA 11701
                                                                <- 38.1R
154638152 TCAGCTCAGAGCTATTTTAGATGCTTCCCAATCTTCAGTCTTCCTTCCCCTGTAACTACT 11761
                        <- 38.2R                                        SINE (MIR) ->
154638212 CCCTCTTTATTTAGAAGGCATACAATAGACATTAACCCTAATGTAGAGAAGTAGTGTGGC 11821
154638272 ATGGGAGTTGAGGGTATGGGCTCTGGAGTTAAACTGCTTGGGTTCAAATCTCATTTATGC 11881
154638332 TATATATTTGCTGTTTTGCCTAAAACATACCTTTACCTCTCTTGGCCTCAGTTGCCCTAT 11941
154638392 TCTAAAATGTGGGTTATTTTGGCACCTTCCTTAGAGGATTTTTGTGAGATTATGTATATA 12001
154638452 AAATACCTGAAACAATACACTGTAAATGCTCAATAAATGCAAGGATGTCATTAATATTAC 12061
154638512 ATGTCAACCAGTTGGCATATAAGGAGATCTGTCTCCAGCCTGGGAAGATTTTAAAATAAG 12121
154638572 CAAACAAAGCATCTTATTTTTAGCCTTTGGGCATCTGCTTTTCCTTGGCTTGTTGGAGAG 12181
154638632 AGTGTTCCTAAGGGATTCCCTTTTTAAAAGTTGGGGGCATGAAGAAGAAAGTGCCAGAGG 12241
                                                                   rs4893817C/A
154638692 GAGACAGTGAAAGACTCCAAGAAAGCAGAGAAGACAAATAGAAAGGAGAAGAAAGTTAGA 12301
154638752 AGGAGGAGAAACCAAAGAGGCCAGGGAGAGGAAAGTAGGTGGCTGGGGAGAGAAAGTATA 12361
154638812 ACAAGAATAAAGTTCATGAAGGTAGGATGGTTGGGGGAGACCTTGTAATACAGAGCTAAG 12421
154638872 AACTTTATTGAAGTTTAACAGAAAAGAAACCCTTGTTTGAAAGGAAGAGATTGGCTATAT 12481
154638932 TTAGTACTGGGTAGGATGTCAGTTCTGAAAATAAACTGCTCCCAGAACTGAAAGCAATCC 12541
                                          rs67918012A/C and rs802637A/C
```

```
154638992 TCTGCACTGTTGTCACCTGCAGCATGCTGGCCTCCCTTATAGAGGAGAGTGGAAAATGGT 12601
154639052 CATATCAGCTACCTCAGAGTTTCTGACATTCACAAGGACTTGCCCCACTTTGTCTCCTTG 12661
154639112 TCCCATTCCTAAGTTTTAAAACTTCTGATTAATGAGAAAATAATTTATACAGCTGAGAAA 12721
154639172 AGTGAAGTATATGTTGGTGTGTATGTGGAGCAGGGAAGTATCCTGCTAAGTTAATACAGG 12781
154639232 AATCTAAACACCATGTGGCCAGGTGCTTAATGCAATATTCACCACATCCATGGTACACAT 12841
154639292 CTGGCATACCACTGTAGACACAGGGACCTAGTCAGAAGTATGACAGGCCATATATAAAGA 12901
154639352 TATATTTGAAGCATCCGATAGGTGTCATAGAGGTAAGTTCTCTATGCAGATGTCATGCTT 12961
154639412 GTGTGACTTAGGGAACTCACATAGGGCATGCAGATAGGCTCAGGTTAGGATACATTAGGA 13021
154639472 AAAGACAGAGAATTCATGATTTGAGCCATCAAAAACAGAATGCATGTGGAGGAGGATGGA 13081
154639532 AATGAGGAGAAAGTTTTCCTTTTCCAACAGGAATAAATATAATTGTGTTCATAGGTACAA 13141
154639592 TGCCATCTACTGCTCCTTTACAAGACACAGAGACTTACAGGTACATTTAGATACATACAC 13201
154639652 ACACAAATACAAACAACATATGTTGTGTTGGGAAAATATAGGTGATGCCTTTCACTGAAT 13261
                                               LINE (L2) ->
154639712 GAGAAAAAAAAACTAGAAAATGGGTATTTTGATGGAACAAAATATAACTAGAATTCTAAC 13321
154639772 TTCTAACTTTAATTCATCACTTCCACTATTACCACTCTCTCATCTAGACAGTTGCGATAG 13381
154639832 CTCCCTAATTGTTTTCCTGCTTCCAACATAAACCCCTCTTCTGAATATATACTCTACAGG 13441
154639892 G                                                            13501
```

Appendix 8.2

1. Primer sequences and PCR conditions used for generating genotyping amplicons

Two rounds of nested PCRs were used to generate all amplicons. Forward primers are listed first, followed by reverse primers. All primers were first optimised using annealing temperature titrations.

Amplicon 1:

Primers	PCR conditions
Primary PCRs- 26.8F - 5'TCATAAAAGGCAAGCTGCCC 3′ 30.58R - 5'TCAGAGTTGGCCTTTAAGCC 3′	96°C for 1min 96°C for 20 sec 62°C for 30 sec } x 26 cycles 66°C for 5:30 min 66°C for 6min
Secondary PCRs- 26.87F - 5' GGAATGCAATTTAGTGTGGC 3′ 30.58R - 5' TCAGAGTTGGCCTTTAAGCC 3′	96°C for 1min 96°C for 20 sec 62°C for 30 sec } x 32 cycles 66°C for 5:30 min 65°C for 6min

Amplicon 2:

Primers	PCR conditions
Primary PCRs- 29.3F - 5' GTTAACCCTCTATTCATGAG 3′ 31.9R - 5' GAAAAGCCCAAGTGGATTAT 3′	96°C for 1min 96°C for 20 sec 52°C for 30 sec } x 32 cycles 62°C for 4 min 62°C for 6min
Secondary PCRs- 29.4F - 5' GTGACCAGTTCCTTTTCAGT 3′ 31.9R - 5' GAAAAGCCCAAGTGGATTAT 3′	96°C for 1min 96°C for 20 sec 57°C for 30 sec } x 32 cycles 62°C for 4 min 62°C for 6min

Amplicon 3:

Primers	PCR conditions
Primary PCRs- 31.1F - 5' GTATGTTTGCCCAGAAGTAG 3' 34.67R - 5' TGCATCCTGGCAGAATCACT 3'	96°C for 1min 96°C for 20 sec 56°C for 30 sec 65°C for 5 min x 32 cycles 65°C for 6min
Secondary PCRs- 31.2F - 5' GTTTAGGGAGTGACTCTACC 3' 34.6R - 5' CACTATGTCCAATCTTCAAC 3'	96°C for 1min 96°C for 20 sec 57°C for 30 sec 65°C for 5 min x 32 cycles 65°C for 6min

Amplicon 4:

Primers	PCR conditions
Primary PCRs- 34.5F - 5' CCCCTTACTAAAGGGGTGAC 3' 38.2R - 5' TGCCTTCTAAATAAAGAGGG 3'	96°C for 1min 96°C for 20 sec 56°C for 30 sec 65°C for 5:30 min x 32 cycles 65°C for 6min
Secondary PCRs- 34.55F - 5' GGGGTGACTAGTGGTATGTA 3' 38.1R - 5' CAGGGGAAGGAAGACTGAAG 3'	96°C for 1min 96°C for 20 sec 60°C for 30 sec 65°C for 5:30 min x 34 cycles 65°C for 6min

2. Primer sequences and PCR conditions used in sperm assays

Two rounds of nested PCRs were used to generate all amplicons. Forward primers are listed first, followed by reverse primers. All primers were first optimised using annealing temperature titrations.

Primary PCRs:

Primers	PCR conditions
Forward primer- FAS0568G/T- 5'gggGACACCATGATATCCTTA**G** 3' 5'gggGACACCATGATATCCTTA**T** 3' Reverse primer- 38.2R - 5' TGCCTTCTAAATAAAGAGGG 3'	96°C for 1min 96°C for 20 sec 52°C for 30 sec 56°C for 12:30 min x 26 cycles 56°C for 6min

Secondary PCRs:

Primers	PCR conditions
Forward primer- FAS8165A/G – 5' CAACATATTAACTTTCT**A** 3' 5' CAACATATTAACTTTCT**G** 3' Reverse primer- 38.2R – 5' TGCCTTCTAAATAAAGAGGG 3'	96°C for 1min $\left.\begin{array}{l}\text{96°C for 20 sec}\\ \text{52°C for 30 sec}\\ \text{56°C for 10:30 min}\end{array}\right\}$ x 26 cycles 56°C for 6min

Tertiary PCRs:

Primers	PCR conditions
Forward primer- 29.3F – 5' GTTAACCCTCTATTCATGAG 3' Reverse primer- 38.1R – 5' CAGGGGAAGGAAGACTGAAG 3'	96°C for 1min $\left.\begin{array}{l}\text{96°C for 20 sec}\\ \text{55°C for 30 sec}\\ \text{58°C for 10 min}\end{array}\right\}$ x 28 cycles 58°C for 6min

Appendix 8.3.

Sequences of ASOs used for SNP typing

SNP	Position	Allele	Oligo ordered
rs34496779-/G	154627073	-	GATTCCGAGGTAATCCTA
		G	GATTCCGA**G**GGTAATCCT
rs5940568G/T	154627209	G	ATCCTTA**G**GATCTCCTGA
		T	ATCCTTA**T**GATCTCCTGA
rs28512393A/G	154628627	A	ATAGTAA**A**CACCGAATAT
		G	ATAGTAA**G**CACCGAATAT
rs35301922-/TTC	154629019	+	AAGAAAA**TTC**TTCTCTTT
		-	AAGAAAATTCTCTTTTCT
rs28638224A/T	154629074	A	GTAAGCA**A**GTAAATATTA
		T	GTAAGCA**T**GTAAATATTA
rs67398165A/G	154629253	A	ACTTTCT**A**ATATTGTTAA
		G	ACTTTCT**G**ATATTGTTAA
rs6642234A/C	154630075	A	TAATTTT**A**AACTAATTGC
		C	TAATTTT**C**AACTAATTGC
rs306899A/G	154630144	A	TTTACCT**A**ATAAAACTAT
		G	TTTACCT**G**ATAAAACTAT
rs306898A/G	154630455	A	AGTGCAT**T**CTCCTATTAC
		G	AGTGCAT**C**CTCCTATTAC
rs306897A/G	154630642	A	CAAAAAC**A**TTATTGAGAC
		G	CAAAAAC**G**TTATTGAGAC
rs3916103G/A	154631368	G	AAGGAGG**G**AAAGGGACTT
		A	AAGGAGG**A**AAAGGGACTT
rs3909635C/T	154632002	C	GAAGTAA**C**CAGAGTTCTT
		T	GAAGTAA**T**CAGAGTTCTT
rs35177475-/C	154632134	-	TAGACATTCCCTGTCCTT
		C	TAGACATT**C**CCCTGTCCT
rs28725321A/C	154632376	A	GAGAAAA**A**TATTCTGAGC
		C	GAGAAAA**C**TATTCTGAGC
rs306896C/T	154632533	C	AATTATA**C**AGGACCTTGT
		T	AATTATA**T**AGGACCTTGT
rs9657865C/A	154633066	C	TCACTTA**C**AAGTGGGAGC
		A	TCACTTA**A**AAGTGGGAGC
rs35654209-/T	154634576	T	CAAAATA**T**TTTCCCCTTA
		-	CAAAATATTTCCCCTTAC
rs9645287C/G	154634887	C	TAATGGA**C**TAAAATTTCC
		G	TAATGGA**G**TAAAATTTCC
rs61461091-/A	154635189	-	AAAAAAAGCACAAGGAGA
		A	AAAAAAA**A**GCACAAGGAG
rs58176487C/T	154635404	C	TTAAATG**C**ATCTATTAAG
		T	TTAAATG**T**ATCTATTAAG
rs306895A/T	154635903	A	TGTTTGC**A**TGAGGACTAA
		T	TGTTTGC**T**TGAGGACTAA
rs5940570A/C	154635933	A	ATGTATT**A**TTTAAAAAAC
		C	ATGTATT**C**TTTAAAAAAC

SNP	Position	Allele	Oligo ordered
rs55996406A/C	154636099	A	CATTAGA**A**GTCACTCACA
		C	CATTAGA**C**GTCACTCACA
rs55764735-/T	154636180	-	TCTGGACATTCACATAAA
		T	TCTGGACA**T**TTCACATAA
rs55720186-/T	154636189	-	ACATAAATGAATATTATA
		T	ACATAAA**T**TGAATATTAT
rs306894C/T	154636327	C	TAGGTCA**C**GTTTTGTTTA
		T	TAGGTCA**T**GTTTTGTTTA
rs306893C/G	154636777	G	CCTTTGC**G**CATTTTTCCA
		C	CCTTTGC**C**CATTTTTCCA
rs35418317-/A	154637426	-	GACTATTCTGTACCCCTT
		A	GACTATT**A**CTGTACCCCT
rs306892A/T	154637511	A	TAGGGAT**A**ATGTTTTCTG
		T	TAGGGAT**T**ATGTTTTCTG
rs5940409C/T	154637596	C	CTTTCCA**C**TTATATAGGT
		T	CTTTCCA**T**TTATATAGGT
rs184257T/C	154637679	C	TGTATTC**C**TATTTTATTC
		T	TGTATTC**T**TATTTTATTC
rs306891A/G	154637845	A	TAAATAG**A**GAATTTGCTA
		G	TAAATAG**G**GAATTTGCTA

Appendix 8.4.

Genotypes of North European and African semen donors. Donor names are highlighted in orange, whilst positions of DNA samples on 96-well plates are in yellow. Donors 1-96 and donors 103, 105, 107, 119, 134, 135, 138, 144, 146, 150, 168, 211, 232 are North European. Others are of African origin. H indicates heterozygous. Positions are according to ChrX, NCBI36/hg18.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
SNP	location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
rs5940568G/T	154627209	G	G	G	G	H	G	G	G	G	G	H	G	G	H	G	H	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G
rs35301922-/TTC	154629019	+	+	+	+	+	+	+	H	+	+	+	+	+	+	+	+	+	H	+	+	H	+	+	+	+	+	+	+	+	+	+	+
rs306899A/G	154630144	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	H	G	G	H	G	G	G	G	G	G	G	G	G	G	G
rs306897A/G	154630642	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	H	G	G	H	G	G	G	G	G	G	G	G	G	G	G
rs28725321A/C	154632376	A	A	A	A	H	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs306896C/T	154632533	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	H	T	T	T	T	T
rs9657865C/A	154633066	C	C	C	C	H	C	C	C	C	C	C	C	C	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs58176487C/T	154635404	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs5940570A/C	154635933	C	C	C	C	H	C	C	C	C	C	C	C	C	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs5940409C/T	154637596	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

		34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66
SNP	location	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66
rs5940568G/T	154627209	G	G	G	G	G	G	G	H	G	H	G	G	G	G	G	H	H	G	H	G	H	G	G	H	G	H	G	G	G	H	H	H
rs35301922-/TTC	154629019	H	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	H	+	+	+	+
rs306899A/G	154630144	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G
rs306897A/G	154630642	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G
rs28725321A/C	154632376	A	A	A	A	A	A	A	A	A	H	A	A	A	A	A	H	H	A	H	A	A	A	A	A	A	A	A	A	A	A	H	A
rs306896C/T	154632533	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	H	T	T	H	T	T	H	T	T	T	T	T	T	T	H
rs9657865C/A	154633066	C	C	C	C	C	C	C	C	C	H	C	C	C	C	C	H	H	C	H	C	C	C	C	C	C	C	C	C	C	C	H	C
rs58176487C/T	154635404	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs5940570A/C	154635933	C	C	C	C	C	C	C	C	C	H	C	C	C	C	C	H	H	C	H	C	C	C	C	C	C	C	C	C	C	C	H	C
rs5940409C/T	154637596	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	H	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

		67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	87	88	89	90	91	92	93	94	96
SNP	location	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	87	88	89	90	91	92	93	94	96
rs5940568G/T	154627209	G	G	G	G	G	G	G	G	H	G	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
rs35301922-/TTC	154629019	+	+	+	H	H	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rs306899A/G	154630144	G	G	G	H	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
rs306897A/G	154630642	G	G	G	H	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
rs28725321A/C	154632376	A	A	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs306896C/T	154632533	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs9657865C/A	154633066	C	C	C	C	C	C	C	C	C	C	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs58176487C/T	154635404	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs5940570A/C	154635933	C	C	C	C	C	C	C	C	C	C	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs5940409C/T	154637596	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

		134	146	150	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	211	232	235	236	237	238	239	240
SNP	location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
rs34496779-/G	154627073	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	G	G	G	G
rs35301922-/TTC	154629019	H	+	+	H	+	+	H	+	+	+	H	+	H	+	+	+	H	+	+	H	H	+	+	+	H	+	+	H	+
rs306899A/G	154630144	H	G	G	H	G	G	H	G	G	G	H	G	H	G	G	G	H	G	G	H	H	G	G	G	H	G	G	H	G
rs306897A/G	154630642	H	G	G	H	G	G	H	G	G	G	H	G	H	G	G	G	H	G	G	H	H	G	G	G	H	G	G	H	G
rs3909635C/T	154632002	T	C	H	H	H	T	T	H	T	T	T	H	H	H	H	T	T	T	T	T	T	H	T	T	H	H	T	H	T
rs28725321A/C	154632376	A	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	H	A	A	A	A	A	A	H	H	H	A	H	H
rs306896C/T	154632533	T	T	T	T	T	H	T	T	T	H	H	T	T	T	T	T	H	T	H	T	T	T	T	H	H	C	T	H	C
rs58176487C/T	154635404	T	T	T	T	T	T	T	T	T	H	H	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	H	H
rs306895A/T	154635903	H	H	H	A	A	H	T	A	A	T	T	H	H	A	A	A	H	H	A	A	H	H	T	H	H	H	H	H	H
rs5940570A/C	154635933	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	H	C	C	C	H	C	C	H	H	H	C	C	C
rs306894C/T	154636327	H	H	H	C	C	H	T	C	C	H	H	H	H	C	C	C	C	H	C	C	C	H	T	C	C	C	H	C	C
rs306893C/G	154636777	H	H	H	G	G	H	C	G	G	C	C	H	H	G	G	H	H	H	H	G	H	H	C	H	C	C	G	H	H
rs306892A/T	154637511	H	T	H	T	T	H	A	T	T	H	H	H	H	T	T	H	T	H	H	H	T	H	A	T	H	H	T	T	T
rs184257T/C	154637679	H	T	H	T	T	H	C	T	T	H	H	H	H	T	T	H	H	H	H	H	T	H	C	T	H	H	T	T	T
rs306891A/G	154637845	H	G	H	G	G	H	A	H	G	H	H	A	H	G	G	H	G	H	H	G	G	H	A	G	G	H	G	G	G

		241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269
SNP	location	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58
rs34496779-/G	154627073	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	G	H	G	G
rs35301922-/TTC	154629019	+	+	+	+	H	+	+	+	+	H	+	+	+	H	+	+	+	+	+	H	H	H	+	+	+	+	+	+	+
rs306899A/G	154630144	G	H	G	G	H	G	H	H	H	H	G	G	G	H	G	G	G	G	G	H	H	H	G	G	H	G	G	G	G
rs306897A/G	154630642	G	G	G	G	H	G	G	H	H	H	G	G	G	H	G	G	G	G	G	H	H	H	G	G	H	G	G	G	G
rs3909635C/T	154632002	T	T	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	T	T	H	H	H	T	H	H	T	T	T	T
rs28725321A/C	154632376	A	A	C	H	H	H	H	H	H	H	A	A	A	A	H	A	A	H	A	C	H	H	A	A	A	A	H	A	A
rs306896C/T	154632533	T	T	C	C	H	H	H	C	T	H	T	H	T	T	C	T	H	H	H	C	H	H	H	H	T	H	H	T	C
rs58176487C/T	154635404	T	T	T	H	T	T	T	T	T	C	T	T	T	T	T	T	C	T	C	H	T	H	H	T	T	H	T	T	T
rs306895A/T	154635903	A	H	T	T	H	H	T	A	A	T	A	H	H	A	T	A	T	T	T	T	H	H	H	H	A	H	T	A	H
rs5940570A/C	154635933	C	C	H	H	H	H	H	C	C	C	C	C	C	C	H	C	C	H	C	H	H	C	C	C	C	C	H	C	C
rs306894C/T	154636327	C	H	C	C	C	C	H	C	C	C	C	H	H	C	H	C	C	H	C	C	C	C	C	H	C	C	H	C	H
rs306893C/G	154636777	G	H	C	C	H	H	C	G	G	C	G	H	H	G	C	G	C	C	C	C	C	H	H	H	G	H	C	H	H
rs306892A/T	154637511	T	H	T	T	T	T	H	T	T	T	T	H	H	T	H	T	T	H	T	T	H	T	T	H	T	T	H	H	H
rs184257T/C	154637679	T	H	T	T	T	T	H	H	H	H	T	H	H	H	H	T	T	H	T	T	H	T	T	H	T	T	H	H	H
rs306891A/G	154637845	G	H	G	G	G	G	H	G	G	G	G	H	H	G	H	G	G	H	G	G	H	G	G	H	G	G	H	H	H

		270	271	272	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	103	105	107	119	135	138	144	168
SNP	location	59	60	61	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88
rs34496779-/G	154627073	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	?
rs35301922-/TTC	154629019	+	+	+	H	H	+	+	H	+	+	H	+	+	H	+	H	+	+	+	+	H	+	+	+	+	+	+	+	?
rs306899A/G	154630144	G	G	G	H	H	G	G	H	G	G	H	G	H	H	G	H	G	G	G	G	H	G	G	G	G	G	G	G	H
rs306897A/G	154630642	G	G	G	H	H	G	G	H	G	G	H	G	G	H	G	H	G	G	G	G	H	G	G	G	G	G	G	G	H
rs3909635C/T	154632002	H	T	T	T	H	T	H	T	T	T	T	T	T	T	H	T	T	T	T	T	T	T	H	T	H	C	H	H	T
rs28725321A/C	154632376	A	H	A	A	A	A	H	A	H	A	A	A	H	H	C	H	H	A	A	A	A	A	A	A	A	A	A	A	A
rs306896C/T	154632533	H	H	T	T	T	T	T	H	H	T	H	H	H	H	T	H	C	H	T	T	T	T	T	T	T	H	T	T	T
rs58176487C/T	154635404	T	T	T	T	H	H	H	H	T	H	H	H	H	T	T	H	T	T	T	T	T	T	T	T	T	T	T	T	T
rs306895A/T	154635903	H	H	A	H	H	T	H	H	T	A	H	H	H	H	H	H	T	A	H	T	A	H	H	T	H	H	H	H	H
rs5940570A/C	154635933	C	H	C	C	C	C	C	C	A	C	C	C	C	H	C	C	H	C	C	C	C	C	C	C	C	C	C	C	C
rs306894C/T	154636327	H	C	C	H	C	H	C	C	C	C	C	C	C	C	C	C	H	C	H	T	C	H	H	T	H	H	H	H	H
rs306893C/G	154636777	H	H	G	G	H	C	H	H	C	H	H	H	H	H	C	H	C	G	H	C	G	H	H	C	H	H	H	H	H
rs306892A/T	154637511	H	T	T	T	T	H	H	T	T	H	T	T	T	T	A	T	H	T	H	A	T	H	H	A	H	H	H	H	H
rs184257T/C	154637679	H	T	T	T	T	H	H	T	T	H	H	T	T	T	C	T	H	T	H	C	T	H	H	C	H	H	H	H	H
rs306891A/G	154637845	H	G	G	G	G	H	G	G	G	H	G	G	G	G	A	G	H	G	H	A	G	H	H	A	H	H	H	H	H

References

ALANI, E., PADMORE, R. & KLECKNER, N. 1990. Analysis of wild-type and rad50 mutants of yeast suggests an intimate relationship between meiotic chromosome synapsis and recombination. Cell.

ALBINI, S. 1987. Synaptonemal complex spreading in Allium cepa and A. fistulosum. I: The initiation and sequence of pairing. Chromosoma.

ALLERS, T. & LICHTEN, M. 2001. Differential timing and control of noncrossover and crossover recombination during meiosis. Cell.

ANDERSON, L. K., OFFENBERG, H. H., VERKUIJLEN, W. M. & HEYTING, C. 1997. RecA-like proteins are components of early meiotic nodules in lily. Proc Natl Acad Sci U S A.

ANDERSON, L. K., REEVES, A., WEBB, L. M. & ASHLEY, T. 1999. Distribution of crossing over on mouse synaptonemal complexes using immunofluorescent localization of MLH1 protein. Genetics.

ARGUESO, J. L., WANAT, J., GEMICI, Z. & ALANI, E. 2004. Competing crossover pathways act during meiosis in Saccharomyces cerevisiae. Genetics.

ARMSTRONG, S. J., KIRKHAM, A. J. & HULTÉN, M. A. 1994. XY chromosome behaviour in the germ-line of the human male: a FISH analysis of spatial orientation, chromatin condensation and pairing. Chromosome Res.

ARNDT, K. T., STYLES, C. & FINK, G. R. 1987. Multiple global regulators control HIS4 transcription in yeast. Science.

ARNHEIM, N., CALABRESE, P. & NORDBORG, M. 2003. Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. Am J Hum Genet.

BAGSHAW, A. T. M., PITT, J. P. W. & GEMMELL, N. J. 2006. Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots. BMC Genomics.

BARLOW, A. L. & HULTÉN, M. A. 1998. Crossing over analysis at pachytene in man. Eur J Hum Genet.

BAUDAT, F., BUARD, J., GREY, C., FLEDEL-ALON, A., OBER, C., PRZEWORSKI, M., COOP, G. & DE MASSY, B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science.

BAUDAT, F. & DE MASSY, B. 2007. Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot. PLoS Genetics.

BAUDAT, F. & DE MASSY, B. 2007. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. Chromosome Res.

BAUDAT, F., MANOVA, K., YUEN, J. P., JASIN, M. & KEENEY, S. 2000. Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. Mol Cell.

BAUDAT, F. & NICOLAS, A. 1997. Clustering of meiotic double-strand breaks on yeast chromosome III. Proc Natl Acad Sci U S A.

BELLANI, M. A., BOATENG, K. A., MCLEOD, D. & CAMERINI-OTERO, R. D. 2010. The expression profile of the major mouse SPO11 isoforms indicates that SPO11beta introduces double strand breaks and suggests that SPO11alpha has an additional role in prophase in both spermatocytes and oocytes. Mol Cell Biol.

BERG, I. L., NEUMANN, R., LAM, K.-W. G., SARBAJNA, S., ODENTHAL-HESSE, L., MAY, C. A. & JEFFREYS, A. J. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet.

BERG, I. L., NEUMANN, R., SARBAJNA, S., ODENTHAL-HESSE, L., BUTLER, N. J. & JEFFREYS, A. J. 2011. Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proc Natl Acad Sci U S A.

BERGERAT, A., DE MASSY, B., GADELLE, D., VAROUTAS, P. C., NICOLAS, A. & FORTERRE, P. 1997. An atypical topoisomerase II from Archaea with implications for meiotic recombination. Nature.

BHALLA, N. & DERNBURG, A. F. 2008. Prelude to a Division. Annu. Rev. Cell Dev. Biol.

BIANCO, P. R., TRACY, R. B. & KOWALCZYKOWSKI, S. C. 1998. DNA strand exchange proteins: a biochemical and physical comparison. Front Biosci.

BILLINGS, T., SARGENT, E. E., SZATKIEWICZ, J. P., LEAHY, N., KWAK, I.-Y., BEKTASSOVA, N., WALKER, M., HASSOLD, T., GRABER, J. H., BROMAN, K. W. & PETKOV, P. M. 2010. Patterns of Recombination Activity on Mouse Chromosome 11 Revealed by High Resolution Mapping. PLoS ONE.

BISHOP, D. K. 1994. RecA homologs Dmc1 and Rad51 interact to form multiple nuclear complexes prior to meiotic chromosome synapsis. Cell.

BISHOP, D. K., PARK, D., XU, L. & KLECKNER, N. 1992. DMC1: a meiosis-specific yeast homolog of E. coli recA required for recombination, synaptonemal complex formation, and cell cycle progression. Cell.

BISHOP, D. K. & ZICKLER, D. 2004. Early decision; meiotic crossover interference prior to stable strand exchange and synapsis. Cell.

BLASCHKE, R. J. & RAPPOLD, G. 2006. The pseudoautosomal regions, SHOX and disease. Curr Opin Genet Dev.

BORDE, V. 2007. The multiple roles of the Mre11 complex for meiotic recombination. Chromosome Res.

BORDE, V., WU, T. C. & LICHTEN, M. 1999. Use of a recombination reporter insert to define meiotic recombination domains on chromosome III of Saccharomyces cerevisiae. Mol Cell Biol.

BÖRNER, G. V., KLECKNER, N. & HUNTER, N. 2004. Crossover/noncrossover differentiation, synaptonemal complex formation, and regulatory surveillance at the leptotene/zygotene transition of meiosis. Cell.

BOTSTEIN, D., WHITE, R. L., SKOLNICK, M. & DAVIS, R. W. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet.

BROMAN, K. W., MURRAY, J. C., SHEFFIELD, V. C., WHITE, R. L. & WEBER, J. L. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet.

BROMAN, K. W., ROWE, L. B., CHURCHILL, G. A. & PAIGEN, K. 2002. Crossover interference in the mouse. Genetics.

BROWN, W. R. 1988. A physical map of the human pseudoautosomal region. EMBO J.

BUARD, J. & VERGNAUD, G. 1994. Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). EMBO J.

BUHLER, C., BORDE, V. & LICHTEN, M. 2007. Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in Saccharomyces cerevisiae. PLoS Biol.

BURGOYNE, P. 1982. Genetic homology and crossing over in the X and Y chromosomes of mammals. Hum Genet.

CALLENDER, T. L. & HOLLINGSWORTH, N. M. 2010. Mek1 suppression of meiotic double-strand break repair is specific to sister chromatids, chromosome autonomous and independent of Rec8 cohesin complexes. Genetics.

CAO, L., ALANI, E. & KLECKNER, N. 1990. A pathway for generation and processing of double-strand breaks during meiotic recombination in S. cerevisiae. Cell.

CARBALLO, J. A., JOHNSON, A. L., SEDGWICK, S. G. & CHA, R. S. 2008. Phosphorylation of the axial element protein Hop1 by Mec1/Tel1 ensures meiotic interhomolog recombination. Cell.

CARRINGTON, M. & CULLEN, M. 2004. Justified chauvinism: advances in defining meiotic recombination through sperm typing. Trends Genet.

CHANDLEY, A. C., GOETZ, P., HARGREAVE, T. B., JOSEPH, A. M. & SPEED, R. M. 1984. On the nature and extent of XY pairing at meiotic prophase in man. Cytogenet Cell Genet.

CHANDLEY, A. C., HARGREAVE, T. B., MCBEATH, S., MITCHELL, A. R. & SPEED, R. M. 1987. Ring XY bivalent: a new phenomenon at metaphase I of meiosis in man. J Med Genet.

CHARCHAR, F. J. 2003. Complex Events in the Evolution of the Human Pseudoautosomal Region 2 (PAR2). Genome Research.

CHENG, E. Y., HUNT, P. A., NALUAI-CECCHINI, T. A., FLIGNER, C. L., FUJIMOTO, V. Y., PASTERNACK, T. L., SCHWARTZ, J. M., STEINAUER, J. E., WOODRUFF, T. J., CHERRY, S. M., HANSEN, T. A., VALLENTE, R. U., BROMAN, K. W. & HASSOLD, T. J. 2009. Meiotic Recombination in Human Oocytes. PLoS Genetics.

CHEUNG, V., BURDICK, J., HIRSCHMANN, D. & MORLEY, M. 2007. Polymorphic Variation in Human Meiotic Recombination. The American Journal of Human Genetics.

CHOWDHURY, R., BOIS, P. R. J., FEINGOLD, E., SHERMAN, S. L. & CHEUNG, V. G. 2009. Genetic analysis of variation in human meiotic recombination. PLoS Genetics.

CHU, S. & HERSKOWITZ, I. 1998. Gametogenesis in yeast is regulated by a transcriptional cascade dependent on Ndt80. Mol Cell.

CHUZHANOVA, N., CHEN, J.-M., BACOLLA, A., PATRINOS, G. P., FÃ©REC, C., WELLS, R. D. & COOPER, D. N. 2009. Gene conversion causing human inherited disease: Evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. Hum. Mutat.

CICCODICOLA, A., D'ESPOSITO, M., ESPOSITO, T., GIANFRANCESCO, F., MIGLIACCIO, C., MIANO, M. G., MATARAZZO, M. R., VACCA, M., FRANZÈ, A., CUCCURESE, M., COCCHIA, M., CURCI, A., TERRACCIANO, A., TORINO, A., COCCHIA, S., MERCADANTE, G., PANNONE, E., ARCHIDIACONO, N., ROCCHI, M., SCHLESSINGER, D. & D'URSO, M. 2000. Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. Hum Mol Genet.

COLE, F., KEENEY, S. & JASIN, M. 2010. Evolutionary conservation of meiotic DSB proteins: more than just Spo11. Genes Dev.

COLE, F., KEENEY, S. & JASIN, M. 2010. Comprehensive, Fine-Scale Dissection of Homologous Recombination Outcomes at a Hot Spot in Mouse Meiosis. Mol Cell.

COLLINS, A. & MORTON, N. E. 1998. Mapping a disease locus by allelic association. Proc Natl Acad Sci U S A.

COLLINS, I. & NEWLON, C. S. 1994. Meiosis-specific formation of joint DNA molecules containing sequences from homologous chromosomes. Cell.

CONSORTIUM, I. H. 2005. A haplotype map of the human genome. Nature.

CONSORTIUM, I. H., FRAZER, K. A., BALLINGER, D. G., COX, D. R., HINDS, D. A., STUVE, L. L., GIBBS, R. A., BELMONT, J. W., BOUDREAU, A., HARDENBOL, P., LEAL, S. M., PASTERNAK, S., WHEELER, D. A., WILLIS, T. D., YU, F., YANG, H., ZENG, C., GAO, Y., HU, H., HU, W., LI, C., LIN, W., LIU, S., PAN, H., TANG, X., WANG, J., WANG, W., YU, J., ZHANG, B., ZHANG, Q., ZHAO, H., ZHAO, H., ZHOU, J., GABRIEL, S. B., BARRY, R., BLUMENSTIEL, B., CAMARGO, A., DEFELICE, M., FAGGART, M., GOYETTE, M., GUPTA, S., MOORE, J., NGUYEN, H., ONOFRIO, R. C., PARKIN, M., ROY, J., STAHL, E., WINCHESTER, E., ZIAUGRA, L., ALTSHULER, D., SHEN, Y., YAO, Z., HUANG, W., CHU, X., HE, Y., JIN, L., LIU, Y., SHEN, Y., SUN, W., WANG, H., WANG, Y., WANG, Y., XIONG, X., XU, L., WAYE, M. M. Y., TSUI, S. K. W., XUE, H., WONG, J. T.-F., GALVER, L. M., FAN, J.-B., GUNDERSON, K., MURRAY, S. S., OLIPHANT, A. R., CHEE, M. S., MONTPETIT, A., CHAGNON, F., FERRETTI, V., LEBOEUF, M., OLIVIER, J.-F., PHILLIPS, M. S., ROUMY, S., SALLÉE, C., VERNER, A., HUDSON, T. J., KWOK, P.-Y., CAI, D., KOBOLDT, D. C., MILLER, R. D., PAWLIKOWSKA, L., TAILLON-MILLER, P., XIAO, M., TSUI, L.-C., MAK, W., SONG, Y. Q., TAM, P. K. H., NAKAMURA, Y., KAWAGUCHI, T., KITAMOTO, T., MORIZONO, T., NAGASHIMA, A., et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature.

COOKE, H. J., BROWN, W. R. & RAPPOLD, G. A. 1985. Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. Nature.

COOP, G., WEN, X., OBER, C., PRITCHARD, J. K. & PRZEWORSKI, M. 2008. High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. Science.

COX, M. E., CAMPBELL, J. K. & LANGEFELD, C. D. 2005. An exploration of sex-specific linkage disequilibrium on chromosome X in Caucasians from the COGA study. BMC Genet.

CRAWFORD, D. C., BHANGALE, T., LI, N., HELLENTHAL, G., RIEDER, M. J., NICKERSON, D. A. & STEPHENS, M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet.

CUI, X. F., LI, H. H., GORADIA, T. M., LANGE, K., KAZAZIAN, H. H., GALAS, D. & ARNHEIM, N. 1989. Single-sperm typing: determination of genetic distance between the G gamma-globin and parathyroid hormone loci by using the polymerase chain reaction and allele-specific oligomers. Proc Natl Acad Sci U S A.

D'ESPOSITO, M., CICCODICOLA, A., GIANFRANCESCO, F., ESPOSITO, T., FLAGIELLO, L., MAZZARELLA, R., SCHLESSINGER, D. & D'URSO, M. 1996. A synaptobrevin-like gene in the Xq28 pseudoautosomal region undergoes X inactivation. Nat Genet.

D'ESPOSITO, M., MATARAZZO, M. R., CICCODICOLA, A., STRAZZULLO, M., MAZZARELLA, R., QUADERI, N. A., FUJIWARA, H., KO, M. S., ROWE, L. B., RICCO, A., ARCHIDIACONO, N., ROCCHI, M., SCHLESSINGER, D. & D'URSO, M. 1997. Differential expression pattern of XqPAR-linked genes SYBL1 and IL9R correlates with the structure and evolution of the region. Hum Mol Genet.

DAUSSET, J., CANN, H., COHEN, D., LATHROP, M., LALOUEL, J. M. & WHITE, R. 1990. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics.

DE BONIS, M. L. 2006. Maintenance of X- and Y-inactivation of the pseudoautosomal (PAR2) gene SPRY3 is independent from DNA methylation and associated to multiple layers of epigenetic modifications. Hum Mol Genet.

DE MASSY, B., BAUDAT, F. & NICOLAS, A. 1994. Initiation of recombination in Saccharomyces cerevisiae haploid meiosis. Proc Natl Acad Sci U S A.

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.

DE MASSY, B., ROCCO, V. & NICOLAS, A. 1995. The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in Saccharomyces cerevisiae. EMBO J.

DERNBURG, A. F., MCDONALD, K., MOULDER, G., BARSTEAD, R., DRESSER, M. & VILLENEUVE, A. M. 1998. Meiotic recombination in C. elegans initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis. Cell.

DETLOFF, P., WHITE, M. A. & PETES, T. D. 1992. Analysis of a gene conversion gradient at the HIS4 locus in Saccharomyces cerevisiae. Genetics.

DEVLIN, B. & RISCH, N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics.

DRAY, E., DUNLOP, M. H., KAUPPI, L., SAN FILIPPO, J., WIESE, C., TSAI, M.-S., BEGOVIC, S., SCHILD, D., JASIN, M., KEENEY, S. & SUNG, P. 2011. Molecular basis for enhancement of the meiotic DMC1 recombinase by RAD51 associated protein 1 (RAD51AP1). Proc Natl Acad Sci U S A.

DRESSER, M. E., EWING, D. J., CONRAD, M. N., DOMINGUEZ, A. M., BARSTEAD, R., JIANG, H. & KODADEK, T. 1997. DMC1 functions in a Saccharomyces cerevisiae meiotic pathway that is largely independent of the RAD51 pathway. Genetics.

DUFFY, D. L. 2006. An Integrated Genetic Map for Linkage Analysis. Behav Genet.

EHMSEN, K.T., HEYER, W. 2008. Biochemistry of meiotic recombination: Formation, processing and resolution of recombination intermediates. Genome Dyn Stab (3). R. Egel, D.H. Lankenau: Recombination and meiosis. Springer Heidelberg.

EGEL, R. 1995. The synaptonemal complex and the distribution of meiotic recombination events. Trends Genet.

EISENBARTH, I., VOGEL, G., KRONE, W., VOGEL, W. & ASSUM, G. 2000. An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. Am J Hum Genet.

ELLIS, N. A., GOODFELLOW, P. J., PYM, B., SMITH, M., PALMER, M., FRISCHAUF, A. M. & GOODFELLOW, P. N. 1989. The pseudoautosomal boundary in man is defined by an Alu repeat sequence inserted on the Y chromosome. Nature.

ELLISON, J. W., WARDAK, Z., YOUNG, M. F., GEHRON ROBEY, P., LAIG-WEBSTER, M. & CHIONG, W. 1997. PHOG, a candidate gene for involvement in the short stature of Turner syndrome. Hum Mol Genet.

FLAQUER, A., FISCHER, C. & WIENKER, T. F. 2009. A new sex-specific genetic map of the human pseudoautosomal regions (PAR1 and PAR2). Hum Hered.

FLAQUER, A., RAPPOLD, G. A., WIENKER, T. F. & FISCHER, C. 2008. The human pseudoautosomal regions: a review for genetic epidemiologists. Eur J Hum Genet.

FOGEL, S. & MORTIMER, R. K. 1969. Informational transfer in meiotic gene conversion. Proc Natl Acad Sci U S A.

FOSS, E., LANDE, R., STAHL, F. W. & STEINBERG, C. M. 1993. Chiasma interference as a function of genetic distance. Genetics.

FREIJE, D., HELMS, C., WATSON, M. S. & DONIS-KELLER, H. 1992. Identification of a second pseudoautosomal region near the Xq and Yq telomeres. Science.

FROENICKE, L., ANDERSON, L. K., WIENBERG, J. & ASHLEY, T. 2002. Male mouse recombination maps for each autosome identified by chromosome painting. Am J Hum Genet.

FUNG, J. C., ROCKMILL, B., ODELL, M. & ROEDER, G. S. 2004. Imposition of crossover interference through the nonrandom distribution of synapsis initiation complexes. Cell.

GARDINER, J. M., BULLARD, S. A., CHROME, C. & MALONE, R. E. 1997. Molecular and genetic analysis of REC103, an early meiotic recombination gene in yeast. Genetics.

GERTON, J. L., DERISI, J., SHROFF, R., LICHTEN, M., BROWN, P. O. & PETES, T. D. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proc Natl Acad Sci U S A.

GETUN, I. V., WU, Z. K., KHALIL, A. M. & BOIS, P. R. J. 2010. Nucleosome occupancy landscape and dynamics at mouse recombination hotspots. EMBO reports.

GILBERTSON, L. A. & STAHL, F. W. 1996. A test of the double-strand break repair model for meiotic recombination in Saccharomyces cerevisiae. Genetics.

GOLDFARB, T. & LICHTEN, M. 2010. Frequent and Efficient Use of the Sister Chromatid for DNA Double-Strand Break Repair during Budding Yeast Meiosis. PLoS Biol.

GOLDMAN, A. S. & LICHTEN, M. 2000. Restriction of ectopic recombination by interhomolog interactions during Saccharomyces cerevisiae meiosis. Proc Natl Acad Sci U S A.

GOLDSTEIN, P. 1982. The synaptonemal complexes of Caenorhabditis elegans: pachytene karyotype analysis of male and hermaphrodite wild-type and him mutants. Chromosoma.

GONSALVES, J., SUN, F., SCHLEGEL, P. N., TUREK, P. J., HOPPS, C. V., GREENE, C., MARTIN, R. H. & PERA, R. A. R. 2004. Defective recombination in infertile men. Hum Mol Genet.

GOYON, C. & LICHTEN, M. 1993. Timing of molecular events in meiosis in Saccharomyces cerevisiae: stable heteroduplex DNA is formed late in meiotic prophase. Mol Cell Biol.

GRAVES, J. A., WAKEFIELD, M. J. & TODER, R. 1998. The origin and evolution of the pseudoautosomal regions of human sex chromosomes. Hum Mol Genet.

GREY, C., BAUDAT, F. & DE MASSY, B. 2009. Genome-wide control of the distribution of meiotic recombination. PLoS Biol.

GUILLON, H., BAUDAT, F., GREY, C., LISKAY, R. M. & DE MASSY, B. 2005. Crossover and noncrossover pathways in mouse meiosis. Mol Cell.

GUILLON, H. & DE MASSY, B. 2002. An initiation site for meiotic crossing-over and gene conversion in the mouse. Nat Genet.

GUILLON, H. & DE MASSY, B. 2002. An initiation site for meiotic crossing-over and gene conversion in the mouse. Nat Genet.

GYAPAY, G., MORISSETTE, J., VIGNAL, A., DIB, C., FIZAMES, C., MILLASSEAU, P., MARC, S., BERNARDI, G., LATHROP, M. & WEISSENBACH, J. 1994. The 1993-94 Généthon human genetic linkage map. Nat Genet.

HASSOLD, T. & SHERMAN, S. 2000. Down syndrome: genetic recombination and the origin of the extra chromosome 21. Clin Genet.

HASSOLD, T. J., SHERMAN, S. L., PETTAY, D., PAGE, D. C. & JACOBS, P. A. 1991. XY chromosome nondisjunction in man is associated with diminished recombination in the pseudoautosomal region. Am J Hum Genet.

HAYASHI, K., YOSHIDA, K. & MATSUI, Y. 2005. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature.

HENDERSON, K. A., KEE, K., MALEKI, S., SANTINI, P. A. & KEENEY, S. 2006. Cyclin-dependent kinase directly regulates initiation of meiotic recombination. Cell.

HENKE, A., FISCHER, C. & RAPPOLD, G. A. 1993. Genetic map of the human pseudoautosomal region reveals a high rate of recombination in female meiosis at the Xp telomere. Genomics.

HER, C., WU, X., WAN, W. & DOGGETT, N. A. 1999. Identification and characterization of the mouse MutS homolog 5: Msh5. Mamm Genome.

HEYTING, C. & DIETRICH, A. J. 1992. Synaptosomal complexes and the organization of chromatin during meiotic prophase. Cell Biol Int Rep.

HILL, W. G. & ROBERTSON, A. 1968. The effects of inbreeding at loci with heterozygote advantage. Genetics.

HINCH, A. G., TANDON, A., PATTERSON, N., SONG, Y., ROHLAND, N., PALMER, C. D., CHEN, G. K., WANG, K., BUXBAUM, S. G., AKYLBEKOVA, E. L., ALDRICH, M. C., AMBROSONE, C. B., AMOS, C., BANDERA, E. V., BERNDT, S. I., BERNSTEIN, L., BLOT, W. J., BOCK, C. H., BOERWINKLE, E., CAI, Q., CAPORASO, N., CASEY, G., CUPPLES, L. A., DEMING, S. L., DIVER, W. R., DIVERS, J., FORNAGE, M., GILLANDERS, E. M., GLESSNER, J., HARRIS, C. C., HU, J. J., INGLES, S. A., ISAACS, W., JOHN, E. M., KAO, W. H. L., KEATING, B., KITTLES, R. A., KOLONEL, L. N., LARKIN, E., LE MARCHAND, L., MCNEILL, L. H., MILLIKAN, R. C., MURPHY, A., MUSANI, S., NESLUND-DUDAS, C., NYANTE, S., PAPANICOLAOU, G. J., PRESS, M. F., PSATY, B. M., REINER, A. P., RICH, S. S., RODRIGUEZ-GIL, J. L., ROTTER, J. I., RYBICKI, B. A., SCHWARTZ, A. G., SIGNORELLO, L. B., SPITZ, M., STROM, S. S., THUN, M. J., TUCKER, M. A., WANG, Z., WIENCKE, J. K., WITTE, J. S., WRENSCH, M., WU, X., YAMAMURA, Y., ZANETTI, K. A., ZHENG, W., ZIEGLER, R. G., ZHU, X., REDLINE, S., HIRSCHHORN, J. N., HENDERSON, B. E., TAYLOR, H. A., PRICE, A. L., HAKONARSON, H., CHANOCK, S. J., HAIMAN, C. A., WILSON, J. G., REICH, D. & MYERS, S. R. 2011. The landscape of recombination in African Americans. Nature.

HINDS, D. A. 2005. Whole-Genome Patterns of Common DNA Variation in Three Human Populations. Science.

HOLLINGSWORTH, N. M. & BRILL, S. J. 2004. The Mus81 solution to resolution: generating meiotic crossovers without Holliday junctions. Genes Dev.

HOLLOWAY, J. K., BOOTH, J., EDELMANN, W., MCGOWAN, C. H. & COHEN, P. E. 2008. MUS81 generates a subset of MLH1-MLH3-independent crossovers in mammalian meiosis. PLoS Genetics.

HOLLOWAY, K., LAWSON, V.E., JEFFREYS, A. J., 2006. Allelic recombination and de novo deletions in sperm in the human -globin gene region. Hum Mol Genet.

HULTÉN, M. 1974. Chiasma distribution at diakinesis in the normal human male. Hereditas.

HUNTER, N. & BORTS, R. H. 1997. Mlh1 is unique among mismatch repair proteins in its ability to promote crossing-over during meiosis. Genes Dev.

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.

JAGIELLO, G. & FANG, J. S. 1979. Analyses of diplotene chiasma frequencies in mouse oocytes and spermatocytes in relation to ageing and sexual dimorphism. Cytogenet Cell Genet.

JEFFREYS, A. J., NEUMANN, R. 2005. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. Hum Mol Genet.

JEFFREYS, A. J., HOLLOWAY, J. K., KAUPPI, L., MAY, C. A., NEUMANN, R., SLINGSBY, M. T. & WEBB, A. J. 2004. Meiotic recombination hot spots and human DNA diversity. Philos Trans R Soc Lond, B, Biol Sci.

JEFFREYS, A. J., KAUPPI, L. & NEUMANN, R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet.

JEFFREYS, A. J. & MAY, C. A. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat Genet.

JEFFREYS, A. J., MURRAY, J. & NEUMANN, R. 1998. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Mol Cell.

JEFFREYS, A. J. & NEUMANN, R. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat. Genet.

JEFFREYS, A. J. & NEUMANN, R. 2009. The rise and fall of a human recombination hot spot. Nat Genet.

JEFFREYS, A. J., NEUMANN, R., PANAYI, M., MYERS, S. & DONNELLY, P. 2005. Human recombination hot spots hidden in regions of strong marker association. Nat Genet.

JEFFREYS, A. J., RITCHIE, A. & NEUMANN, R. 2000. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum Mol Genet.

JONES, G. 1984. The control of chiasma distribution. Symposia of the Society for Experimental Biology.

JORGENSON, E., TANG, H., GADDE, M., PROVINCE, M., LEPPERT, M., KARDIA, S., SCHORK, N., COOPER, R., RAO, D. C., BOERWINKLE, E. & RISCH, N. 2005. Ethnicity and human genetic linkage maps. Am J Hum Genet.

KADYK, L. C. & HARTWELL, L. H. 1992. Sister chromatids are preferred over homologs as substrates for recombinational repair in Saccharomyces cerevisiae. Genetics.

KAUPPI, L., BARCHI, M., BAUDAT, F., ROMANIENKO, P. J., KEENEY, S. & JASIN, M. 2011. Distinct Properties of the XY Pseudoautosomal Region Crucial for Male Meiosis. Science.

KAUPPI, L., JASIN, M. & KEENEY, S. 2007. Meiotic crossover hotspots contained in haplotype block boundaries of the mouse genome. Proc Natl Acad Sci U S A.

KAUPPI, L., STUMPF, M. P. H. & JEFFREYS, A. J. 2005. Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region. Genomics.

KEELAGHER, R. E., COTTON, V. E., GOLDMAN, A. S. H. & BORTS, R. H. 2011. Separable roles for Exonuclease I in meiotic DNA double-strand break repair. DNA Repair.

KEENEY, S. 2007. Spo11 and the formation of DNA double-strand breaks in meiosis. Genome Dyn Stab (2). R. Egel, D.H. Lankenau: Recombination and Meiosis. Springer Heidelberg.

KEENEY, S., BAUDAT, F., ANGELES, M., ZHOU, Z. H., COPELAND, N. G., JENKINS, N. A., MANOVA, K. & JASIN, M. 1999. A mouse homolog of the Saccharomyces cerevisiae meiotic recombination DNA transesterase Spo11p. Genomics.

KEENEY, S., GIROUX, C. N. & KLECKNER, N. 1997. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. Cell.

KEENEY, S. & KLECKNER, N. 1995. Covalent protein-DNA complexes at the 5' strand termini of meiosis-specific double-strand breaks in yeast. Proc Natl Acad Sci U S A.

KIRKPATRICK, D. T., FAN, Q. & PETES, T. D. 1999. Maximal stimulation of meiotic recombination by a yeast transcription factor requires the transcription activation domain and a DNA-binding domain. Genetics.

KIRKPATRICK, D. T., WANG, Y. H., DOMINSKA, M., GRIFFITH, J. D. & PETES, T. D. 1999. Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes. Mol Cell Biol.

KLECKNER, N. 1996. Meiosis: how could it work? Proc Natl Acad Sci U S A.

KLECKNER, N. 2006. Chiasma formation: chromatin/axis interplay and the role(s) of the synaptonemal complex. Chromosoma.

KLECKNER, N. 2006. Chiasma formation: chromatin/axis interplay and the role(s) of the synaptonemal complex. Chromosoma.

KLECKNER, N., STORLAZZI, A. & ZICKLER, D. 2003. Coordinate variation in meiotic pachytene SC length and total crossover/chiasma frequency under conditions of constant DNA length. Trends Genet.

KLECKNER, N., ZICKLER, D., JONES, G. H., DEKKER, J., PADMORE, R., HENLE, J. & HUTCHINSON, J. 2004. A mechanical basis for chromosome function. Proc Natl Acad Sci U S A.

KNEITZ, B., COHEN, P. E., AVDIEVICH, E., ZHU, L., KANE, M. F., HOU, H., KOLODNER, R. D., KUCHERLAPATI, R., POLLARD, J. W. & EDELMANN, W. 2000. MutS homolog 4 localization to meiotic chromosomes is required for chromosome pairing during meiosis in male and female mice. Genes Dev.

KOEHLER, K. E., CHERRY, J. P., LYNN, A., HUNT, P. A. & HASSOLD, T. J. 2002. Genetic control of mammalian meiotic recombination. I. Variation in exchange frequencies among males from inbred mouse strains. Genetics.

KOEHLER, K. E., SCHRUMP, S. E., CHERRY, J. P., HASSOLD, T. J. & HUNT, P. A. 2006. Near-human aneuploidy levels in female mice with homeologous chromosomes. Curr Biol.

KOLAS, N. K., SVETLANOV, A., LENZI, M. L., MACALUSO, F. P., LIPKIN, S. M., LISKAY, R. M., GREALLY, J., EDELMANN, W. & COHEN, P. E. 2005. Localization of MMR proteins on meiotic chromosomes in mice indicates distinct functions during prophase I. J Cell Biol.

KONG, A., BARNARD, J., GUDBJARTSSON, D. F., THORLEIFSSON, G., JONSDOTTIR, G., SIGURDARDOTTIR, S., RICHARDSSON, B., JONSDOTTIR, J., THORGEIRSSON, T., FRIGGE, M. L., LAMB, N. E., SHERMAN, S., GULCHER, J. R. & STEFANSSON, K. 2004. Recombination rate and reproductive success in humans. Nat Genet.

KONG, A., GUDBJARTSSON, D. F., SAINZ, J., JONSDOTTIR, G. M., GUDJONSSON, S. A., RICHARDSSON, B., SIGURDARDOTTIR, S., BARNARD, J., HALLBECK, B., MASSON, G., SHLIEN, A., PALSSON, S. T., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R. & STEFANSSON, K. 2002. A high-resolution recombination map of the human genome. Nat Genet.

KONG, A., THORLEIFSSON, G., GUDBJARTSSON, D. F., MASSON, G., SIGURDSSON, A., JONASDOTTIR, A., WALTERS, G. B., JONASDOTTIR, A., GYLFASON, A., KRISTINSSON, K. T., GUDJONSSON, S. A., FRIGGE, M. L., HELGASON, A., THORSTEINSDOTTIR, U. & STEFANSSON, K. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature.

KONG, A., THORLEIFSSON, G., STEFANSSON, H., MASSON, G., HELGASON, A., GUDBJARTSSON, D. F., JONSDOTTIR, G. M., GUDJONSSON, S. A., SVERRISSON, S., THORLACIUS, T., JONASDOTTIR, A., HARDARSON, G. A., PALSSON, S. T., FRIGGE, M. L., GULCHER, J. R., THORSTEINSDOTTIR, U. & STEFANSSON, K. 2008. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. Science.

KVALØY, K., GALVAGNI, F. & BROWN, W. R. 1994. The sequence organization of the long arm pseudoautosomal region of the human sex chromosomes. Hum Mol Genet.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, N., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. 2001. Initial sequencing and analysis of the human genome. Nature.

LAO, J. P. & HUNTER, N. 2010. Trying to Avoid Your Sister. PLoS Biol.

LAURIE, D. A., HULTÉN, M. & JONES, G. H. 1981. Chiasma frequency and distribution in a sample of human males: chromosomes 1, 2, and 9. Cytogenet Cell Genet.

LAURIE, D. A. & HULTÉN, M. A. 1985. Further studies on chiasma distribution and interference in the human male. Ann Hum Genet.

LAWRIE, N. M., TEASE, C. & HULTÉN, M. A. 1995. Chiasma frequency, distribution and interference maps of mouse autosomes. Chromosoma.

LAZZERONI, L. C., ARNHEIM, N., SCHMITT, K. & LANGE, K. 1994. Multipoint mapping calculations for sperm-typing data. Am J Hum Genet.

LENZI, M. L., SMITH, J., SNOWDEN, T., KIM, M., FISHEL, R., POULOS, B. K. & COHEN, P. E. 2005. Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis i in human oocytes. Am J Hum Genet.

LEWONTIN, R. C. 1988. On measures of gametic disequilibrium. Genetics.

LI, H. H., GYLLENSTEN, U. B., CUI, X. F., SAIKI, R. K., ERLICH, H. A. & ARNHEIM, N. 1988. Amplification and analysis of DNA sequences in single human sperm and diploid cells. Nature.

LI, J., AGARWAL, S. & ROEDER, G. S. 2007. SSP2 and OSW1, two sporulation-specific genes involved in spore morphogenesis in Saccharomyces cerevisiae. Genetics.

LI, L. & HAMER, D. H. 1995. Recombination and allelic association in the Xq/Yq homology region. Hum Mol Genet.

LI, Y. F., NUMATA, M., WAHLS, W. P. & SMITH, G. R. 1997. Region-specific meiotic recombination in Schizosaccharomyces pombe: the rec11 gene. Mol Microbiol.

LICHTEN, M. & GOLDMAN, A. S. 1995. Meiotic recombination hotspots. Annu Rev Genet.

LIEN, S., KAMIŃSKI, S., ALESTRÖM, P. & ROGNE, S. 1993. A simple and powerful method for linkage analysis by amplification of DNA from single sperm cells. Genomics.

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.

LIM, D. S. & HASTY, P. 1996. A mutation in mouse rad51 results in an early embryonic lethal that is suppressed by a mutation in p53. Mol Cell Biol.

LIU, J., WU, T. C. & LICHTEN, M. 1995. The location and structure of double-strand DNA breaks induced during yeast meiosis: evidence for a covalently linked DNA-protein intermediate. EMBO J.

LONGHESE, M. P., BONETTI, D., GUERINI, I., MANFRINI, N. & CLERICI, M. 2009. DNA double-strand breaks in meiosis: checking their formation, processing and repair. DNA Repair.

LYNN, A. 2002. Covariation of Synaptonemal Complex Length and Mammalian Meiotic Exchange Rates. Science.

LYNN, A., SCHRUMP, S., CHERRY, J., HASSOLD, T. & HUNT, P. 2005. Sex, not genotype, determines recombination levels in mice. Am J Hum Genet.

LYNN, A., SOUCEK, R. & BÖRNER, G. V. 2007. ZMM proteins during meiosis: Crossover artists at work. Chromosome Res.

MAGUIRE, M. P., RIESS, R. W. & PAREDES, A. M. 1993. Evidence from a maize desynaptic mutant points to a probable role of synaptonemal complex central region components in provision for subsequent chiasma maintenance. Genome.

MAHADEVAIAH, S. K., TURNER, J. M., BAUDAT, F., ROGAKOU, E. P., DE BOER, P., BLANCO-RODRÍGUEZ, J., JASIN, M., KEENEY, S., BONNER, W. M. & BURGOYNE, P. S. 2001. Recombinational DNA double-strand breaks in mice precede synapsis. Nat Genet.

MANCERA, E., BOURGON, R., BROZZI, A., HUBER, W. & STEINMETZ, L. M. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature.

MANGS, A. H. & MORRIS, B. J. 2008. ZRANB2: structural and functional insights into a novel splicing protein. Int J Biochem Cell Biol.

MAO-DRAAYER, Y., GALBRAITH, A. M., PITTMAN, D. L., COOL, M. & MALONE, R. E. 1996. Analysis of meiotic recombination pathways in the yeast Saccharomyces cerevisiae. Genetics.

MARCON, E. & MOENS, P. 2003. MLH1p and MLH3p localize to precociously induced chiasmata of okadaic-acid-treated mouse spermatocytes. Genetics.

MARTINI, E., DIAZ, R. L., HUNTER, N. & KEENEY, S. 2006. Crossover homeostasis in yeast meiosis. Cell.

MATISE, T. C., PERLIN, M. & CHAKRAVARTI, A. 1994. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. Nat Genet.

MATISE, T. C., SACHIDANANDAM, R., CLARK, A. G., KRUGLYAK, L., WIJSMAN, E., KAKOL, J., BUYSKE, S., CHUI, B., COHEN, P., DE TOMA, C., EHM, M., GLANOWSKI, S., HE, C., HEIL, J., MARKIANOS, K., MCMULLEN, I., PERICAK-VANCE, M. A., SILBERGLEIT, A., STEIN, L., WAGNER, M., WILSON, A. F., WINICK, J. D., WINN-DEEN, E. S., YAMASHIRO, C. T., CANN, H. M., LAI, E. & HOLDEN, A. L. 2003. A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am J Hum Genet.

MAY, C. A., SHONE, A. C., KALAYDJIEVA, L., SAJANTILA, A. & JEFFREYS, A. J. 2002. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. Nat Genet.

MCDERMOTT, A. 1973. The frequency and distribution of chiasmata in man. Ann Hum Genet.

MCKIM, K. S., GREEN-MARROQUIN, B. L., SEKELSKY, J. J., CHIN, G., STEINBERG, C., KHODOSH, R. & HAWLEY, R. S. 1998. Meiotic synapsis in the absence of recombination. Science.

MCVEAN, G., AWADALLA, P. & FEARNHEAD, P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics.

MCVEAN, G. A. T. 2004. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Science.

MOENS, P. B., CHEN, D. J., SHEN, Z., KOLAS, N., TARSOUNAS, M., HENG, H. H. & SPYROPOULOS, B. 1997. Rad51 immunocytology in rat and mouse spermatocytes and oocytes. Chromosoma.

MOENS, P. B., KOLAS, N. K., TARSOUNAS, M., MARCON, E., COHEN, P. E. & SPYROPOULOS, B. 2002. The time course and chromosomal localization of recombination-related proteins at meiosis in the mouse are compatible with models that can resolve the early DNA-DNA interactions without reciprocal recombination. J Cell Sci.

MOHANDAS, T. K., SPEED, R. M., PASSAGE, M. B., YEN, P. H., CHANDLEY, A. C. & SHAPIRO, L. J. 1992. Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp. Am J Hum Genet.

MURDOCH, B., OWEN, N., SHIRLEY, S., CRUMB, S., BROMAN, K. W. & HASSOLD, T. 2010. Multiple loci contribute to genome-wide recombination levels in male mice. Mamm Genome.

MYERS, S. 2005. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science.

MYERS, S., BOWDEN, R., TUMIAN, A., BONTROP, R. E., FREEMAN, C., MACFIE, T. S., MCVEAN, G. & DONNELLY, P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science.

MYERS, S., FREEMAN, C., AUTON, A., DONNELLY, P. & MCVEAN, G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet.

NAG, D. K. & PETES, T. D. 1993. Physical detection of heteroduplexes during meiotic recombination in the yeast Saccharomyces cerevisiae. Mol Cell Biol.

NEALE, M. J. & KEENEY, S. 2006. Clarifying the mechanics of DNA strand exchange in meiotic recombination. Nature.

NEALE, M. J., PAN, J. & KEENEY, S. 2005. Endonucleolytic processing of covalent protein-linked DNA double-strand breaks. Nature.

NEUMANN, R. & JEFFREYS, A. J. 2006. Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. Hum Mol Genet.

NICKLAS, R. B. 1977. Chromosome Distribution: Experiments on Cell Hybrids and In vitro [and Discussion]. Phil Trans R Soc Lond. B.

NICOLAS, A. & PETES, T. D. 1994. Polarity of meiotic gene conversion in fungi: contrasting views. Experientia.

NICOLAS, A., TRECO, D., SCHULTES, N. P. & SZOSTAK, J. W. 1989. An initiation site for meiotic gene conversion in the yeast Saccharomyces cerevisiae. Nature.

NIU, H., LI, X., JOB, E., PARK, C., MOAZED, D., GYGI, S. P. & HOLLINGSWORTH, N. M. 2007. Mek1 kinase is regulated to suppress double-strand break repair between sister chromatids during budding yeast meiosis. Mol Cell Biol.

NIU, H., WAN, L., BAUMGARTNER, B., SCHAEFER, D., LOIDL, J. & HOLLINGSWORTH, N. M. 2005. Partner choice during meiosis is regulated by Hop1-promoted dimerization of Mek1. Mol Biol Cell.

OHTA, K., SHIBATA, T. & NICOLAS, A. 1994. Changes in chromatin structure at recombination initiation sites during yeast meiosis. EMBO J.

OLIVER-BONET, M., TUREK, P. J., SUN, F., KO, E. & MARTIN, R. H. 2005. Temporal progression of recombination in human males. Mol Hum Reprod.

OSMAN, F., DIXON, J., DOE, C. L. & WHITBY, M. C. 2003. Generating crossovers by resolution of nicked Holliday junctions: a role for Mus81-Eme1 in meiosis. Mol Cell.

PAGE, D. C., BIEKER, K., BROWN, L. G., HINTON, S., LEPPERT, M., LALOUEL, J. M., LATHROP, M., NYSTROM-LAHTI, M., DE LA CHAPELLE, A. & WHITE, R. 1987. Linkage, physical mapping, and DNA sequence analysis of pseudoautosomal loci on the human X and Y chromosomes. Genomics.

PAGE, S. L. & HAWLEY, R. S. 2003. Chromosome choreography: the meiotic ballet. Science.

PAIGEN, K., SZATKIEWICZ, J. P., SAWYER, K., LEAHY, N., PARVANOV, E. D., NG, S. H. S., GRABER, J. H., BROMAN, K. W. & PETKOV, P. M. 2008. The Recombinational Anatomy of a Mouse Chromosome. PLoS Genetics.

PAN, J., SASAKI, M., KNIEWEL, R., MURAKAMI, H., BLITZBLAU, HANNAH G., TISCHFIELD, SAM E., ZHU, X., NEALE, MATTHEW J., JASIN, M., SOCCI, NICHOLAS D., HOCHWAGEN, A. & KEENEY, S. 2011. A Hierarchical Combination of Factors Shapes the Genome-wide Topography of Yeast Meiotic Recombination Initiation. Cell.

PARVANOV, E. D., NG, S. H. S., PETKOV, P. M. & PAIGEN, K. 2009. Trans-regulation of mouse meiotic recombination hotspots by Rcr1. PLoS Biol.

PARVANOV, E. D., PETKOV, P. M. & PAIGEN, K. 2010. Prdm9 controls activation of mammalian recombination hotspots. Science.

PEOPLES, T. L., DEAN, E., GONZALEZ, O., LAMBOURNE, L. & BURGESS, S. M. 2002. Close, stable homolog juxtaposition during meiosis in budding yeast is dependent on meiotic recombination, occurs independently of synapsis, and is distinct from DSB-independent pairing contacts. Genes Dev.

PEOPLES-HOLST, T. L. & BURGESS, S. M. 2005. Multiple branches of the meiotic recombination pathway contribute independently to homolog pairing and stable juxtaposition during meiosis in budding yeast. Genes Dev.

PETES, T. D. 2001. Meiotic recombination hot spots and cold spots. Nat Rev Genet.

PETIT, C., LEVILLIERS, J. & WEISSENBACH, J. 1988. Physical mapping of the human pseudo-autosomal region; comparison with genetic linkage map. EMBO J.

PETKOV, P. M., BROMAN, K. W., SZATKIEWICZ, J. P. & PAIGEN, K. 2007. Crossover interference underlies sex differences in recombination rates. Trends Genet.

PHILLIPS, C. M., MENG, X., ZHANG, L., CHRETIEN, J. H., URNOV, F. D. & DERNBURG, A. F. 2009. Identification of chromosome sequence motifs that mediate meiotic pairing and synapsis in C. elegans. Nat Cell Biol.

PITTMAN, D. L., COBB, J., SCHIMENTI, K. J., WILSON, L. A., COOPER, D. M., BRIGNULL, E., HANDEL, M. A. & SCHIMENTI, J. C. 1998. Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for Dmc1, a germline-specific RecA homolog. Mol Cell.

PLUG, A. W., XU, J., REDDY, G., GOLUB, E. I. & ASHLEY, T. 1996. Presynaptic association of Rad51 protein with selected sites in meiotic chromatin. Proc Natl Acad Sci U S A.

POLANI, P. E. 1972. Centromere localization at meiosis and the position of chiasmata in the male and female mouse. Chromosoma.

PTAK, S. E., HINDS, D. A., KOEHLER, K., NICKEL, B., PATIL, N., BALLINGER, D. G., PRZEWORSKI, M., FRAZER, K. A. & PÄÄBO, S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet.

PTAK, S. E., ROEDER, A. D., STEPHENS, M., GILAD, Y., PÄÄBO, S. & PRZEWORSKI, M. 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. PLoS Biol.

PTAK, S. E., HINDS, D. A., KOEHLER, K., NICKEL, B., PATIL, N., BALLINGER, D. G., PRZEWORSKI, M., FRAZER, K. A. & PÄÄBO, S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet.

RAO, E., WEISS, B., FUKAMI, M., RUMP, A., NIESLER, B., MERTZ, A., MUROYA, K., BINDER, G., KIRSCH, S., WINKELMANN, M., NORDSIEK, G., HEINRICH, U., BREUNING, M. H., RANKE, M. B., ROSENTHAL, A., OGATA, T. & RAPPOLD, G. A. 1997. Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. Nat Genet.

RAPPOLD, G. A. 1993. The pseudoautosomal regions of the human sex chromosomes. Hum Genet.

RAPPOLD, G. A., KLINK, A., WEISS, B. & FISCHER, C. 1994. Double crossover in the human Xp/Yp pseudoautosomal region and its bearing on interference. Hum Mol Genet.

RASMUSSEN, S. 1977. Chromosome pairing in triploid females of Bombyx mori analyzed by three dimensional reconstructions of synaptonemal complexes [silkworm, pachytene, bivalents, …. Carlsberg Research Communications (Denmark).

RATTNER, J.B., GOLDSMITH, M.R.., & HAMKALO, B.A., 1981. Chromosome organization during male meiosis in Bombyx mori. Chromosoma.

RENWICK, J. H. 1969. Progress in mapping human autosomes. Br Med Bull.

RESNICK, M.A. 1975. The repair of double-strand breaks in chromosomal DNA of yeast. Basic Life Sci.

REVENKOVA, E., EIJPE, M., HEYTING, C., HODGES, C. A., HUNT, P. A., LIEBE, B., SCHERTHAN, H. & JESSBERGER, R. 2004. Cohesin SMC1 beta is required for meiotic chromosome dynamics, sister chromatid cohesion and DNA recombination. Nat Cell Biol.

ROEDER, G. S. 1997. Meiotic chromosomes: it takes two to tango. Genes Dev.

ROMANIENKO, P. J. & CAMERINI-OTERO, R. D. 1999. Cloning, characterization, and localization of mouse and human SPO11. Genomics.

ROMANIENKO, P. J. & CAMERINI-OTERO, R. D. 2000. The mouse Spo11 gene is required for meiotic chromosome synapsis. Mol Cell.

ROSS, M. T., GRAFHAM, D. V., COFFEY, A. J., SCHERER, S., MCLAY, K., MUZNY, D., PLATZER, M., HOWELL, G. R., BURROWS, C., BIRD, C. P., FRANKISH, A., LOVELL, F. L., HOWE, K. L., ASHURST, J. L., FULTON, R. S., SUDBRAK, R., WEN, G., JONES, M. C., HURLES, M. E., ANDREWS, T. D., SCOTT, C. E., SEARLE, S., RAMSER, J., WHITTAKER, A., DEADMAN, R., CARTER, N. P., HUNT, S. E., CHEN, R., CREE, A., GUNARATNE, P., HAVLAK, P., HODGSON, A., METZKER, M. L., RICHARDS, S., SCOTT, G., STEFFEN, D., SODERGREN, E., WHEELER, D. A., WORLEY, K. C., AINSCOUGH, R., AMBROSE, K. D., ANSARI-LARI, M. A., ARADHYA, S., ASHWELL, R. I. S., BABBAGE, A. K., BAGGULEY, C. L., BALLABIO, A., BANERJEE, R., BARKER, G. E., BARLOW, K. F., BARRETT, I. P., BATES, K. N., BEARE, D. M., BEASLEY, H., BEASLEY, O., BECK, A., BETHEL, G., BLECHSCHMIDT, K., BRADY, N., BRAY-ALLEN, S., BRIDGEMAN, A. M., BROWN, A. J., BROWN, M. J., BONNIN, D., BRUFORD, E. A., BUHAY, C., BURCH, P., BURFORD, D., BURGESS, J., BURRILL, W., BURTON, J., BYE, J. M., CARDER, C., CARREL, L., CHAKO, J., CHAPMAN, J. C., CHAVEZ, D., CHEN, E., CHEN, G., CHEN, Y., CHEN, Z., CHINAULT, C., CICCODICOLA, A., CLARK, S. Y., CLARKE, G., CLEE, C. M., CLEGG, S., CLERC-BLANKENBURG, K., CLIFFORD, K., COBLEY, V., COLE, C. G., CONQUER, J. S., CORBY, N., CONNOR, R. E., DAVID, R., DAVIES, J., DAVIS, C., DAVIS, J., DELGADO, O., DESHAZO, D., et al. 2005. The DNA sequence of the human X chromosome. Nature.

ROSS-MACDONALD, P. & ROEDER, G. S. 1994. Mutation of a meiosis-specific MutS homolog decreases crossing over but not mismatch correction. Cell.

ROUYER, F., DE LA CHAPELLE, A., ANDERSSON, M. & WEISSENBACH, J. 1990. An interspersed repeated sequence specific for human subtelomeric regions. EMBO J.

ROUYER, F., SIMMLER, M. C., JOHNSSON, C., VERGNAUD, G., COOKE, H. J. & WEISSENBACH, J. 1986. A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. Nature.

ROUYER, F., SIMMLER, M. C., VERGNAUD, G., JOHNSSON, C., LEVILLIERS, J., PETIT, C. & WEISSENBACH, J. 1986. The pseudoautosomal region of the human sex chromosomes. Cold Spring Harb Symp Quant Biol.

SAGE, J., MARTIN, L., CUZIN, F. & RASSOULZADEGAN, M. 1995. cDNA sequence of the murine synaptonemal complex protein 1 (SCP1). Biochim Biophys Acta.

SAJANTILA, A., LAHERMO, P., ANTTINEN, T., LUKKA, M., SISTONEN, P., SAVONTAUS, M. L., 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 Research.

SASANUMA, H., MURAKAMI, H., FUKUDA, T., SHIBATA, T., NICOLAS, A. & OHTA, K. 2007. Meiotic association between Spo11 regulated by Rec102, Rec104 and Rec114. Nucleic Acids Res.

SCHMITT, K., LAZZERONI, L. C., FOOTE, S., VOLLRATH, D., FISHER, E. M., GORADIA, T. M., LANGE, K., PAGE, D. C. & ARNHEIM, N. 1994. Multipoint linkage map of the human pseudoautosomal region, based on single-sperm typing: do double crossovers occur during male meiosis? Am J Hum Genet.

SCHULTES, N. P. & SZOSTAK, J. W. 1990. Decreasing gradients of gene conversion on both sides of the initiation site for meiotic recombination at the ARG4 locus in yeast. Genetics.

SCHWACHA, A. & KLECKNER, N. 1994. Identification of joint molecules that form frequently between homologs but rarely between sister chromatids during yeast meiosis. Cell.

SCHWACHA, A. & KLECKNER, N. 1997. Interhomolog bias during meiotic recombination: meiotic functions promote a highly differentiated interhomolog-only pathway. Cell.

SEKELSKY, J. J., MCKIM, K. S., CHIN, G. M. & HAWLEY, R. S. 1995. The Drosophila meiotic recombination gene mei-9 encodes a homologue of the yeast excision repair protein Rad1. Genetics.

SHERIDAN, S. & BISHOP, D. K. 2006. Red-Hed regulation: recombinase Rad51, though capable of playing the leading role, may be relegated to supporting Dmc1 in budding yeast meiosis. Genes Dev.

SHERMAN, J. D., HERICKHOFF, L. A. & STACK, S. M. 1992. Silver staining two types of meiotic nodules. Genome.

SHERMAN, J. D. & STACK, S. M. 1995. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (Lycopersicon esculentum). Genetics.

SHI, Q., SPRIGGS, E., FIELD, L. L., KO, E., BARCLAY, L. & MARTIN, R. H. 2001. Single sperm typing demonstrates that reduced recombination is associated with the production of aneuploid 24,XY human sperm. Am J Med Genet.

SHIFMAN, S., BELL, J. T., COPLEY, R. R., TAYLOR, M. S., WILLIAMS, R. W., MOTT, R. & FLINT, J. 2006. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. PLoS Biol.

SHINOHARA, A., GASIOR, S., OGAWA, T., KLECKNER, N. & BISHOP, D. K. 1997. Saccharomyces cerevisiae recA homologues RAD51 and DMC1 have both distinct and overlapping roles in meiotic recombination. Genes Cells.

SHINOHARA, M., OH, S. D., HUNTER, N. & SHINOHARA, A. 2008. Crossover assurance and crossover interference are distinctly regulated by the ZMM proteins during yeast meiosis. Nat Genet.

SHINOHARA, M., SAKAI, K., SHINOHARA, A. & BISHOP, D. K. 2003. Crossover interference in Saccharomyces cerevisiae requires a TID1/RDH54- and DMC1-dependent pathway. Genetics.

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.

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.

SIMMLER, M. C., ROUYER, F., VERGNAUD, G., NYSTRÖM-LAHTI, M., NGO, K. Y., DE LA CHAPELLE, A. & WEISSENBACH, J. 1985. Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes. Nature.

SLATKIN, M. 1994. Linkage disequilibrium in growing and stable populations. Genetics.

SMAGULOVA, F., GREGORETTI, I. V., BRICK, K., KHIL, P., CAMERINI-OTERO, R. D. & PETUKHOVA, G. V. 2011. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature.

SPEED, R. M. 1977. The effects of ageing on the meiotic chromosomes of male and female mice. Chromosoma.

SPEED, R. M. & CHANDLEY, A. C. 1990. Prophase of meiosis in human spermatocytes analysed by EM microspreading in infertile men and their controls and comparisons with human oocytes. Hum Genet.

STADLER, D. R. 1959. THE RELATIONSHIP OF GENE CONVERSION TO CROSSING OVER IN NEUROSPORA. Proc Natl Acad Sci U S A.

STEFANSSON, H., HELGASON, A., THORLEIFSSON, G., STEINTHORSDOTTIR, V., MASSON, G., BARNARD, J., BAKER, A., JONASDOTTIR, A., INGASON, A., GUDNADOTTIR, V. G., DESNICA, N., HICKS, A., GYLFASON, A., GUDBJARTSSON, D. F., JONSDOTTIR, G. M., SAINZ, J., AGNARSSON, K., BIRGISDOTTIR, B., GHOSH, S., OLAFSDOTTIR, A., CAZIER, J.-B., KRISTJANSSON, K., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R., KONG, A. & STEFANSSON, K. 2005. A common inversion under selection in Europeans. Nat Genet.

STEINER, W. W., SCHRECKHISE, R. W. & SMITH, G. R. 2002. Meiotic DNA breaks at the S. pombe recombination hot spot M26. Mol Cell.

STEINMETZ, M., MINARD, K., HORVATH, S., MCNICHOLAS, J., SRELINGER, J., WAKE, C., LONG, E., MACH, B. & HOOD, L. 1982. A molecular map of the immune response region from the major histocompatibility complex of the mouse. Nature.

STUMPF, M. P. H. & MCVEAN, G. A. T. 2003. Estimating recombination rates from population-genetic data. Nat Rev Genet.

SUN, F., OLIVER-BONET, M., LIEHR, T., STARKE, H., TUREK, P., KO, E., RADEMAKER, A. & MARTIN, R. H. 2006. Variation in MLH1 distribution in recombination maps for individual chromosomes from human males. Hum Mol Genet.

SUN, H., TRECO, D. & SZOSTAK, J. W. 1991. Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the ARG4 recombination initiation site. Cell.

SYM, M., ENGEBRECHT, J. A. & ROEDER, G. S. 1993. ZIP1 is a synaptonemal complex protein required for meiotic chromosome synapsis. Cell.

SYM, M. & ROEDER, G. S. 1994. Crossover interference is abolished in the absence of a synaptonemal complex protein. Cell.

SZOSTAK, J. W., ORR-WEAVER, T. L., ROTHSTEIN, R. J. & STAHL, F. W. 1983. The double-strand-break repair model for recombination. Cell.

TAPPER, W., COLLINS, A., GIBSON, J., MANIATIS, N., ENNIS, S. & MORTON, N. E. 2005. A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci U S A.

TARSOUNAS, M., MORITA, T., PEARLMAN, R. E. & MOENS, P. B. 1999. RAD51 and DMC1 form mixed complexes associated with mouse meiotic chromosome cores and synaptonemal complexes. J Cell Biol.

TEASE, C., HARTSHORNE, G. & HULTÉN, M. 2006. Altered patterns of meiotic recombination in human fetal oocytes with asynapsis and/or synaptonemal complex fragmentation at pachytene. Reprod Biomed Online.

TEASE, C. & HULTÉN, M. A. 2004. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. Cytogenet Genome Res.

TESSÉ, S., STORLAZZI, A., KLECKNER, N., GARGANO, S. & ZICKLER, D. 2003. Localization and roles of Ski8p protein in Sordaria meiosis and delineation of three mechanistically distinct steps of meiotic homolog juxtaposition. Proc Natl Acad Sci U S A.

THOMPSON, D. A. & STAHL, F. W. 1999. Genetic control of recombination partner preference in yeast meiosis. Isolation and characterization of mutants elevated for meiotic unequal sister-chromatid recombination. Genetics.

TSUKAMOTO, Y., MITSUOKA, C., TERASAWA, M., OGAWA, H. & OGAWA, T. 2005. Xrs2p regulates Mre11p translocation to the nucleus and plays a role in telomere elongation and meiotic recombination. Mol Biol Cell.

USUI, T., OHTA, T., OSHIUMI, H., TOMIZAWA, J., OGAWA, H. & OGAWA, T. 1998. Complex formation and functional versatility of Mre11 of budding yeast in recombination. Cell.

VALLENTE, R. U., CHENG, E. Y. & HASSOLD, T. J. 2006. The synaptonemal complex and meiotic recombination in humans: new approaches to old questions. Chromosoma.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW,

R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z., WANG, A., WANG, X., WANG, J., WEI, M., WIDES, R., XIAO, C., YAN, C., et al. 2001. The sequence of the human genome. Science.

WANG, T. F., KLECKNER, N. & HUNTER, N. 1999. Functional specificity of MutL homologs in yeast: evidence for three Mlh1-based heterocomplexes with distinct roles during meiosis in recombination and mismatch correction. Proc Natl Acad Sci U S A.

WEBB, A. J., BERG, I. L. & JEFFREYS, A. 2008. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proc Natl Acad Sci U S A.

WEGMANN, D., KESSNER, D. E., VEERAMAH, K. R., MATHIAS, R. A., NICOLAE, D. L., YANEK, L. R., SUN, Y. V., TORGERSON, D. G., RAFAELS, N., MOSLEY, T., BECKER, L. C., RUCZINSKI, I., BEATY, T. H., KARDIA, S. L. R., MEYERS, D. A., BARNES, K. C., BECKER, D. M., FREIMER, N. B. & NOVEMBRE, J. 2011. Recombination rates in admixed individuals identified by ancestry-based inference. Nat Genet.

WEINER, B. & KLECKNER, N. 1994. Chromosome pairing via multiple interstitial interactions before and during meiosis in yeast. Cell.

WEISSENBACH, J., LEVILLIERS, J., PETIT, C., ROUYER, F. & SIMMLER, M. C. 1987. Normal and abnormal interchanges between the human X and Y chromosomes. Development.

WETTSTEIN, D. , RASMUSSEN, S. W. & HOLM, P. B. 1984. The synaptonemal complex in genetic segregation. Annu Rev Genet.

WHITE, M. A., DETLOFF, P., STRAND, M. & PETES, T. D. 1992. A promoter deletion reduces the rate of mitotic, but not meiotic, recombination at the HIS4 locus in yeast. Curr Genet.

WILTZIUS, J. J. W., HOHL, M., FLEMING, J. C. & PETRINI, J. H. J. 2005. The Rad50 hook domain is a critical determinant of Mre11 complex functions. Nat Struct Mol Biol.

WINCKLER, W. 2005. Comparison of Fine-Scale Recombination Rates in Humans and Chimpanzees. Science.

WOODS, L. M., HODGES, C. A., BAART, E., BAKER, S. M., LISKAY, M. & HUNT, P. A. 1999. Chromosomal influence on meiotic spindle assembly: abnormal meiosis I in female Mlh1 mutant mice. J Cell Biol.

WU, T. C. & LICHTEN, M. 1994. Meiosis-induced double-strand break sites determined by yeast chromatin structure. Science.

XU, L., WEINER, B. M. & KLECKNER, N. 1997. Meiotic cells monitor the status of the interhomolog recombination complex. Genes Dev.

YAUK, C. L., BOIS, P. R. J. & JEFFREYS, A. J. 2003. High-resolution sperm typing of meiotic recombination in the mouse MHC Ebeta gene. EMBO J.

YOSHIDA, K., KONDOH, G., MATSUDA, Y., HABU, T., NISHIMUNE, Y. & MORITA, T. 1998. The mouse RecA-like gene Dmc1 is required for homologous chromosome synapsis during meiosis. Mol Cell.

YU, A., ZHAO, C., FAN, Y., JANG, W., MUNGALL, A. J., DELOUKAS, P., OLSEN, A., DOGGETT, N. A., GHEBRANIOUS, N., BROMAN, K. W. & WEBER, J. L. 2001. Comparison of human genetic and sequence-based physical maps. Nature.

ZAKHARYEVICH, K., MA, Y., TANG, S., HWANG, P. Y.-H., BOITEUX, S. & HUNTER, N. 2010. Temporally and Biochemically Distinct Activities of Exo1 during Meiosis: Double-Strand Break Resection and Resolution of Double Holliday Junctions. Mol Cell.

ZENVIRTH, D., RICHLER, C., BARDHAN, A., BAUDAT, F., BARZILAI, A., WAHRMAN, J. & SIMCHEN, G. 2003. Mammalian meiosis involves DNA double-strand breaks with 3′ overhangs. Chromosoma.

ZHANG, L., CUI, X., SCHMITT, K., HUBERT, R., NAVIDI, W. & ARNHEIM, N. 1992. Whole genome amplification from a single cell: implications for genetic analysis. Proc Natl Acad Sci U S A.

ZICKLER, D. & KLECKNER, N. 1998. The leptotene-zygotene transition of meiosis. Annu Rev Genet.

ZICKLER, D. & KLECKNER, N. 1999. Meiotic chromosomes: integrating structure and function. Annu Rev Genet.

ZIERHUT, C., BERLINGER, M., RUPP, C., SHINOHARA, A. & KLEIN, F. 2004. Mnd1 is required for meiotic interhomolog repair. Curr Biol.