The Landscape of Recombination in African Americans

Leveraging human population variation to investigate homologous recombination

Anjali Gupta Hinch
Merton College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas 2012

Acknowledgements

First of all, I'd like to thank you, Simon, for being a brilliant and inspiring supervisor, and for your advice, guidance, and support over the last four years.

Thank you, David, for your wisdom and your mentorship. It's also always incredible fun visiting your Harvard lab.

Thanks especially to Arti Tandon, Nick Patterson, Nadin Rohland, Yunli Song, Jim Wilson, and Rosie Butler, for your contributions towards making this work happen, and for help and fun conversations along the way.

I would also like to thank Peter Donnelly, Gil McVean, Adam Auton, Chris Holmes, Nudrat Noor, Iain Mathieson, Jonathan Flint, Nicolas Altemose, Colin Freeman, Liz Batty, and Graham Coop. I have always enjoyed our conversations, and I have learnt a lot from you. Thanks to Mikkel Schierup for generously letting me use materials from his work.

Thank you to the Wellcome Trust and the Clarendon fund for funding and supporting my DPhil education, my scientific learning, and this work.

Thanks, most of all, to Robbie. You make me so happy.

Abstract

Homologous recombination is a highly regulated and complex biological process required for sexual reproduction in humans and many other species. The mechanisms of initiation and control of this process, however, are only partially understood. The aim of this work is to tease apart and utilize the differences in recombination localization between human populations to understand the underlying biological processes.

Prior to this work, recombination had been extensively studied in European populations, revealing that recombination events cluster in thousands of narrow segments of the genome, known as hotspots. However, much less was known about other human populations. I have developed an approach that leverages the recent mixture of peoples of West African and European ancestry in the Americas to build a map specifying the location of recombination events in a large number of African Americans. I showed that this map is significantly different from the European map at the scale of hotspots, demonstrating evolution of recombination patterns within the human lineage. I find that the African-American map is more diverse than the European map, with thousands of hotspots active in African Americans but not in Europeans.

I performed genome-wide association analysis, which shows that a single gene, *PRDM9*, largely controls the switch between different hotspot landscapes. Prior research had implicated PRDM9, a zinc-finger histone methyltransferase, as an important factor in determining the location of recombination events. This work shows that the switch is functionally due to variation in the zinc finger array, and that PRDM9 is likely to be the dominant factor determining the genome-wide distribution of hotspots. Further, there is no evidence that hotspots are shared between individuals with PRDM9 variants containing significantly different zinc finger arrays.

I find that, despite profound differences between hotspot locations, stochasticity in individual meioses, and large differences between the sexes, rates at mega-base scales are highly conserved between populations.

Finally, I investigate recombination in the pseudoautosomal region PAR1, which undergoes an obligatory recombination event in males, resulting in recombination rates several times greater than the genome-average. I build maps of recombination in PAR1 and investigate the controversy of whether *PRDM9* in responsible for localizing hotspots in this region.

Contents

List of Figures

Chapter 1

Introduction

Recombination or 'crossing over' is the process by which genetic material is shuffled between homologous chromosomes during sexual reproduction in eukaryotes, resulting in offspring that are genetically distinct from their parents. Recombination is responsible for the explosion of genetic diversity that accompanied the rise of sexual reproduction and is therefore of fundamental interest in the study of natural history and evolution. It is also of interest to cell biologists because it plays, with few known exceptions, an essential mechanistic role in the production of gametes (eggs and sperm, or spores) in eukaryotes. Research into recombination from these quite distinct fields of study has converged in recent years and led to many exciting discoveries. In this chapter, I give an overview of some of these discoveries as well as the tools used in making them. I start with an introduction to meiosis, the process of production of gametes during the course of which recombination takes place.

1.1 Meiosis

Sexual reproduction occurs in the vast majority of eukaryotes. An enduring mystery in evolutionary biology is why it is so widespread. For example, of the more than 41,000 known vertebrate species [Wilson, 1988], only 22 fish, 23 amphibians and 29

reptiles have been identified to reproduce asexually [Dawley et al., 1989]. Further, it is rare for any genus in the eukaryotic tree of life to be composed entirely of asexual species. This is puzzling because most theoretical arguments suggest that sexual reproduction should be costly [Otto and Lenormand, 2002]. First, sexual reproduction is highly complex, it involves the cost of finding a mate, it is often slower, and carries risks such as chromosomal aberrations and the spread of parasitic genetic elements. Second, the theory of the 'two-fold cost of sex' states that unless a sexually reproducing couple can produce and nurture twice as many offspring as an asexual individual, sexual reproduction will have lower reproductive success per individual. In the setting where one sexual partner does not contribute resources to the offspring (paternal care is relatively uncommon in nature), a sexual couple should produce the same number of offspring as an asexual couple can. This means that sexual species will produce only half as many offspring per capita, resulting in a 'two-fold cost' for sexual reproduction. However, the two-fold cost may not hold if the offspring in sexually reproducing species are more fit, and have a greater chance of reaching reproductive age. Most theoretical arguments that suggest such an evolutionary benefit from sexual reproduction are centred on the idea that sex generates greater genetic, and consequently phenotypic, variability in populations because recombination breaks down genetic associations between segments of chromosomes. These arguments generally work only in a limited range of settings, however. For instance, recombination should be disfavoured once a beneficial combination of alleles has arisen in the population. Nevertheless, sexual reproduction may be favoured if the fitness of such allelic interactions fluctuates rapidly, for example, due to the co-adaptation of parasites with their hosts. Changes in the environment or ecological niches (for example, due to migration) could also lead to rapid changes in selective pressures, and favour the ability of sexually reproducing organisms to retain variability. Despite the intuitive appeal of this arguments, however, evidence for these hypotheses is still scarce.

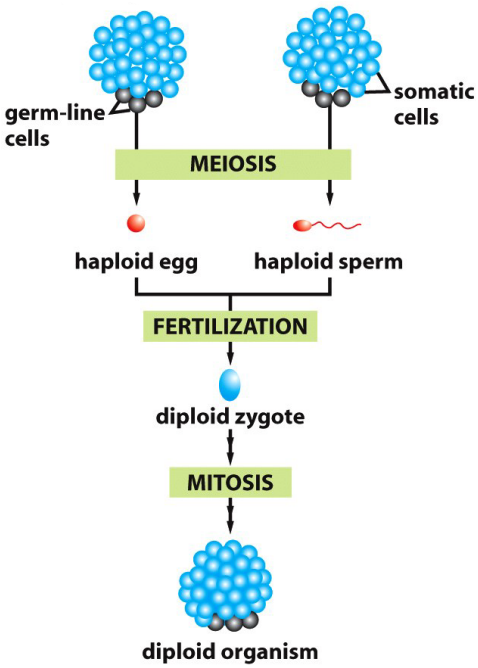


Figure 1.1: Meiosis and Mitosis. Most eukaryotes consist of diploid cells (blue). Haploid gametes (red) are produced by meiosis and fuse to produce the diploid zygote. The zygote develops into the offspring by cell proliferation via mitosis (reproduced from Alberts et al. [2007]).

Eukaryotes are usually diploid, i.e., they are composed of cells containing two sets of chromosomes. Sexual reproduction is carried out by creating specialized cells called gametes which are haploid, i.e., they contain only one set of chromosomes. Correct chromosomal content for the offspring is restored in the final stage of reproduction during which two gametes fuse together to produce the first cell of the offspring, known as the zygote. Corresponding chromosomes from each set – one maternal and one paternal – are called homologous chromosomes or homologs. Meiosis is the specialized cell division in which gametes are produced from diploid parental cells and is restricted to specialized cells called germ-line cells in higher eukaryotes. Mitosis is the far more common and simpler type of cell division in which a cell divides into two genetically identical daughter cells and is the primary means of cell proliferation (Figure 1.1).

The stages of meiosis are illustrated in Figure 1.2. It involves one round of chromo-

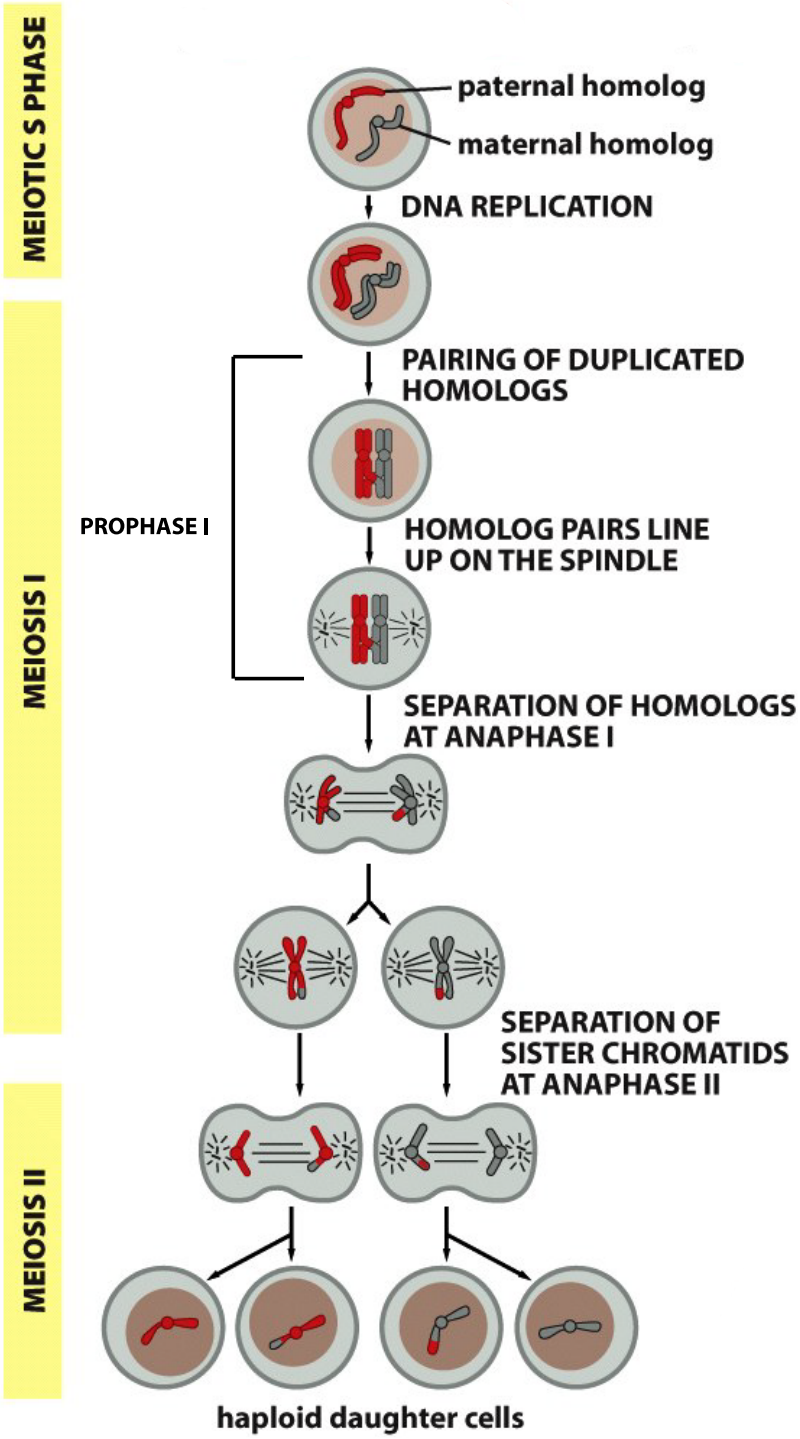


Figure 1.2: Meiosis consists of one round of replication followed by two rounds of division. During Meiosis I, duplicated homologs consist of two pairs of tightly held sister chromatids. Recombination or crossing over takes place during Prophase I eventually resulting in gametes with both maternal and paternal genetic material – as shown by the partly grey and partly red gametes (adapted from Alberts et al. [2007]).

some replication, called the Meiotic S phase, resulting in a cell with four chromatids of each chromosome – two sister chromatids from the maternal homolog and two from the paternal homolog[1]. Each pair of sister chromatids is tightly held together throughout their length by protein complexes called cohesin complexes [Petronczki et al., 2003]. These four chromatids have to be distributed to four different nuclei. This occurs simultaneously for all chromosomes and is achieved by two rounds of chromosome segregation and cell division. In the first division of meiosis called Meiosis I, the duplicated and conjoined homologous chromosomes pair up in a process called synapsis and exchange genetic material to form X-shaped structures called chiasmata (Figure 1.3). For segregation to take place correctly, homologous chromosomes participate in a complex dance during which they are stabilized by opposing forces – connections between the homologs that keep them together and forces that pull them apart towards the poles. Chiasmata create the first force [Petronczki et al., 2003]: the linkage between the sister chromatids due to cohesion gets transmitted to homologs resulting in them being held together (Figure 1.3). Once all the chromosomes are correctly oriented and stabilized, migration of the homologs towards the poles is triggered and they segregate into two daughter cells. In the second meiotic division (Meiosis II), the sister chromatids further segregate to produce four haploid cells.

1.2 Recombination and Crossovers

There are two main aspects to the redistribution of genetic material during meiosis, both of which are important for increasing genetic diversity and natural selection. First, the parental chromosomes are independently re-assorted: the maternal and paternal chromosomes in each homologous pair separate and migrate to opposite poles in the first meiotic division. The decision for each homologous pair is independent,

[1]This is true for the 'autosomes', i.e., chromosomes which are present in two copies for all members of a species. Sex-determining chromosomes may be present in different numbers in different sexes. Each sex chromosome will replicate and form a pair of sister chromatids.

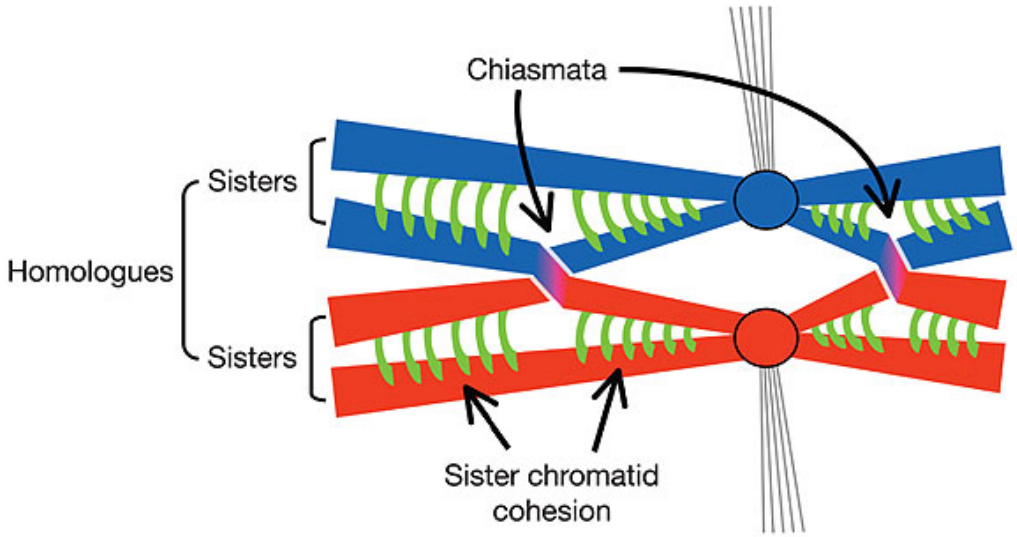


Figure 1.3: After the Meiosis S phase, sister chromatids from each homolog (red and blue) are held together by cohesin complexes (green). An exchange of genetic material between non-sister chromatids takes place yielding cross-shaped structures called chiasmata which hold the homologs together. Dissolution of sister chromatid cohesion along the arms allows the homologs to separate at the end of Meiosis I. (reproduced from Neale and Keeney [2006]).

and the maternal homolog in each pair is equally likely to end up in one pole as the other. In other words, each of the final haploid gametes is likely to contain some chromosomes (as defined by their centromeres) from the mother and the rest from the father, as opposed to all from one or the other. Second, genetic content is exchanged at points where the homologous chromosomes 'cross over' as shown in Figure 1.3. Crossovers, as these exchanges are called, are one of the outcomes of recombination initiation, and are particularly important as they are the only outcome which produce chiasmata. Mutations that result in the failure to form chiasmata lead to chromosomes that cannot be held together and are segregated at random in the first meiotic division in *S. cerevisiae* and *C. elegans* [Cao et al., 1990; Dernburg et al., 1998; Klein et al., 1999]. This can lead to massive aneuploidy and inviability of progeny. Less severe aneuploidies are compatible with life, but cause congenital birth defects such as Down's syndrome [Hassold et al., 2007]. Further, mutations

that interfere with the initiation of recombination, for instance in mice, result in the failure of both males and females in producing functional gametes [Baudat et al., 2000; Romanienko and Camerini-Otero, 2000], and hence infertility.

There are only a few exceptions to the near-universal requirement of recombination for meiosis. In some species, recombination is absent in one sex while being necessary in the other, e.g., fruit fly *Drosophila Melanogaster* males and silkworm *Bombyx Mori* females. Despite the profound importance of recombination, its initiation, mechanism and control are only partially understood and remain active fields of research.

1.2.1 Initiation of recombination

Recombination is initiated by formation of programmed double-strand breaks (DSBs) [Sun et al., 1989] during prophase I. The formation of DSBs and their carefully regulated repair lie at the heart of meiotic recombination. DSB formation is catalyzed by an enzyme called Spo11 [Cao et al., 1990; Keeney et al., 1997]. First discovered in *Saccharomyces cerevisiae*, Spo11 is evolutionarily conserved – Spo11 orthologs are present and required for recombination initiation in all species examined so far [Keeney, 2001]. In many organisms, the defects due to Spo11 mutants can be partially or fully rescued by the introduction of DSBs by an external agent, such as ionizing radiation [Dernburg et al., 1998; Storlazzi et al., 2003], proving that DSBs are required to initiate recombination.

Spo11 has sequence homology with, and is biochemically and structurally similar to, the archael topoisomerase VI [Bergerat et al., 1997; Diaz et al., 2002]. It functions as a dimer and remains covalently attached to the 5' strand ends of the DSB (Figure 1.5). Subsequently, it is removed from DSB ends by an endonuclease, releasing Spo11-bound oligonucleotides [Neale et al., 2005]. End-resection by an exonuclease then takes place on the 5' end, leading to the formation of 3' single-stranded DNA (ssDNA) tails on either side of the DSB [Sun et al., 1991; Bishop et al., 1992]. The 3' ssDNA

tails participate in homology search to initiate repair as described in the next section.

Spo11 acts in concert with several additional proteins. At least nine additional proteins are required in *S. cerevisiae*, though they vary widely across species and are not as well conserved [Keeney, 2008]. It is worth pointing out that Spo11 does not make DSBs uniformly at random along the genome, and this will be discussed in detail in Section 1.4.4.

1.2.2 Search for homology

In non-meiotic cells, DSBs may be repaired by non-homologous end-joining (NHEJ) or by homologous recombination (HR). NHEJ simply ligates trimmed DSB ends irrespective of sequence, generally with the loss of one or more nucleotides at the site of joining. It is quick but much less faithful than the HR repair mechanism and is repressed during meiosis [Goedecke et al., 1999]. HR is utilized, for example, in mitotic cells with newly replicated DNA where the sister chromatid can be utilized as a template. The HR mechanism uses single stranded DNA near the DSB site[2] to search for matching double-stranded DNA with "sufficient" homology to use as a template for repair.

In certain stages of the (mitotic) cell cycle with newly replicated DNA, the sister chromatids are intimately connected by cohesins. In those cells, HR mediated repair using the sister chromatid as the template is strongly favoured. In meiosis, however, repairing the meiotic DSBs using the sister chromatid is pointless as it would lead neither to chiasmata nor any genetic change. Meiotic recombination using the sister chromatid appears to be actively suppressed using mechanisms that are not fully understood (Section 1.4.3). Nevertheless, it appears that some fraction of programmed meiotic DSBs are repaired using the sister chromatid [Hyppa and Smith, 2010].

Enzymes that assess the degree of homology between the ssDNA overhang and

[2]Single-stranded DNA is produced by end-resection at the DSB site by exonucleases, for example, as described in Section 1.2.1

potential double-stranded DNA (dsDNA) templates, Rad51 and Dmc1, are orthologs of the *Escherichia Coli* protein RecA. Orthologs of RecA appear to be universally involved in DSB repair, from prokaryotes to humans. Rad51 is common to mitosis and meiosis in eukaryotes while its paralog Dmc1 is specific to meiosis. These proteins form long helical filaments hundreds of bases long on single-stranded DNA (ssDNA) flanking the DSB site. ssDNA in this form can intertwine with double-stranded DNA (dsDNA) regardless of sequence (Figure 1.4) and search it for a homologous sequence. The exact reaction is not fully understood, but appears to involve transient Watson-Crick base pairing exchanges between the ssDNA and the complementary strand of the dsDNA and bases that flip out when mismatched [Folta-Stogniew et al., 2004]. Once a homologous sequence is found, "strand invasion" happens: the ssDNA displaces one strand of dsDNA, which spools out of the DNA-protein complex (Figure 1.4). The result is a "heteroduplex" – a region of dsDNA formed by the pairing of two DNA strands that were previously part of different DNA molecules. How the heteroduplex is processed is the topic of the next section.

There is considerable evidence that homology search plays an important role in enabling the homologs to pair up in many organisms (Section 1.2.4). In vivo, the process of homology search is very rapid despite the very large number of potential dsDNA targets even in organisms with relatively small genomes such as *E. coli.* Understanding how the search is accelerated continues to be an active area of research and a model has been proposed involving the use of short DNA sequences as initial recognition fragments during exploratory searches [Dorfman et al., 2004]. Early homolog interactions prior to the formation of DSBs, such as telomere clustering [Scherthan, 2007], likely also play a role in speeding up the search.

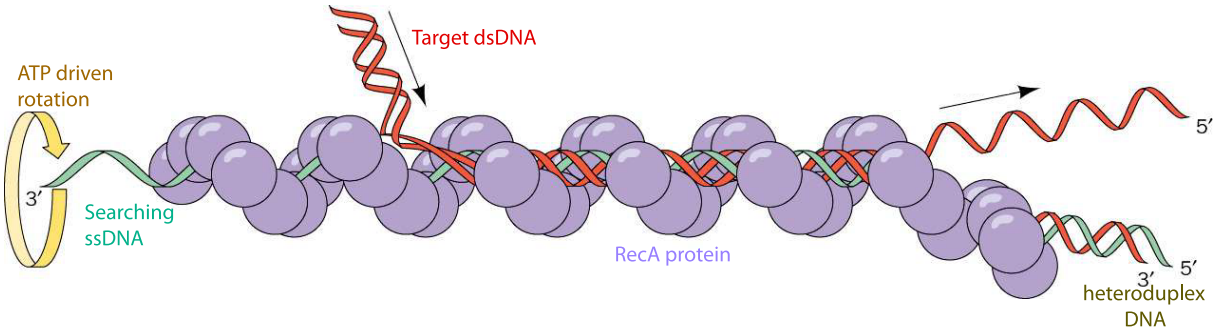


Figure 1.4: Homology testing and strand exchange by ssDNA (green) coated with RecA (purple). The ATP driven rotation of the RecA filament spools in dsDNA, causing it to be stretched and partially unwound (Figure adapted from West [1992]).

1.2.3 Double strand break repair via homologous recombination

Crossovers are *reciprocal* exchanges between chromosomes and, generally speaking, Mendelian segregation of parental alleles on either side of a crossover does take place. However, it was observed over fifty years ago in the study of *S. cerevisiae* and filamentous fungi such as *Neurospora* that this is not always the case. Certain filamentous fungi make spores containing all four chromatids from a single meiosis which are found together in a particular order (the order reflects the position of the centromeres of the homologous chromosomes during Meiosis I). It was found in these fungi that non-Mendelian segregation was common – and a range of different types of non-Mendelian transmissions with different frequencies were observed. For example, if one of the parental alleles is A and the other is a, we might expect, without loss of generality, the pattern $AAAAaaaa$ (4:4) in an ordered octad in the absence of a crossover, and $AAaaAAaa$ (4:4) if there has been a single crossover. However, $AAaaaaaa$ (6:2) and $AAAAAAaa$ (2:6) were also observed [Mitchell, 1955]. Such non-Mendelian transmissions are collectively known as 'gene conversions' and are *non-reciprocal* transfers of genetic information from chromosome to another. Finally, $AAAaAaaa$ (4:4), $AAAAAaaa$ (5:3) and $AAAaaaaa$ (3:5) were also found [Kitani

et al., 1962]. Observations like these in different model organisms, in mitotic and meiotic recombination, and increasing knowledge of the biochemistry of meiosis have led to an evolution of models of recombination in the last half century. It is now appreciated that the unexpected patterns of segregation are due to heteroduplex DNA and the different ways it can be resolved (or left unresolved). The current view of how crossovers happen, called the *Double Holliday Junction* (dHJ) model, is shown in Figure 1.5 [Szostak et al., 1983; Sun et al., 1991].

As discussed in Sections 1.2.1 and 1.2.2, a 3' ssDNA tail is produced by nucleolytic activity after the formation of a DSB, which then proceeds to find homologous dsDNA and performs strand invasion. The strand invasion results in the invading ssDNA (orange in Figure 1.5) to be paired with the corresponding strand of a section of the matching dsDNA (red). This causes the now unpaired strand of the dsDNA (red) to flip out in a "D-loop". The other resected DSB end (orange) then proceeds to anneal to the D-loop (red). This results in two cross shaped structures called Holliday junctions both of which have heteroduplex DNA on one side (orange/red pairings in Figure 1.5) and the original double-stranded chromosomes on the other. This structure is called a Double Holliday Junction, as illustrated in Figure 1.6. Each of the two Holliday junctions has two axes of cleavage. Depending on each of the axes chosen, the resolution could be crossover or gene conversion without crossover. Theoretically, each Holliday junction could be cleaved independently, leading to equal numbers of crossovers and non-crossovers, however, in most organisms the decision is highly regulated (Section 1.4.2).

It is worth noting that newly synthesized DNA (green in Figure 1.5) which fills the gaps created by the 3' to 5' resection on the chromatid experiencing the DSB (orange), is filled using the homologous chromatid (red) as template. Crossovers, therefore, are accompanied by segments of DNA that undergo gene conversion, and those gene conversions cause loss of the allele on the 'initiating' chromatid. A similar loss happens

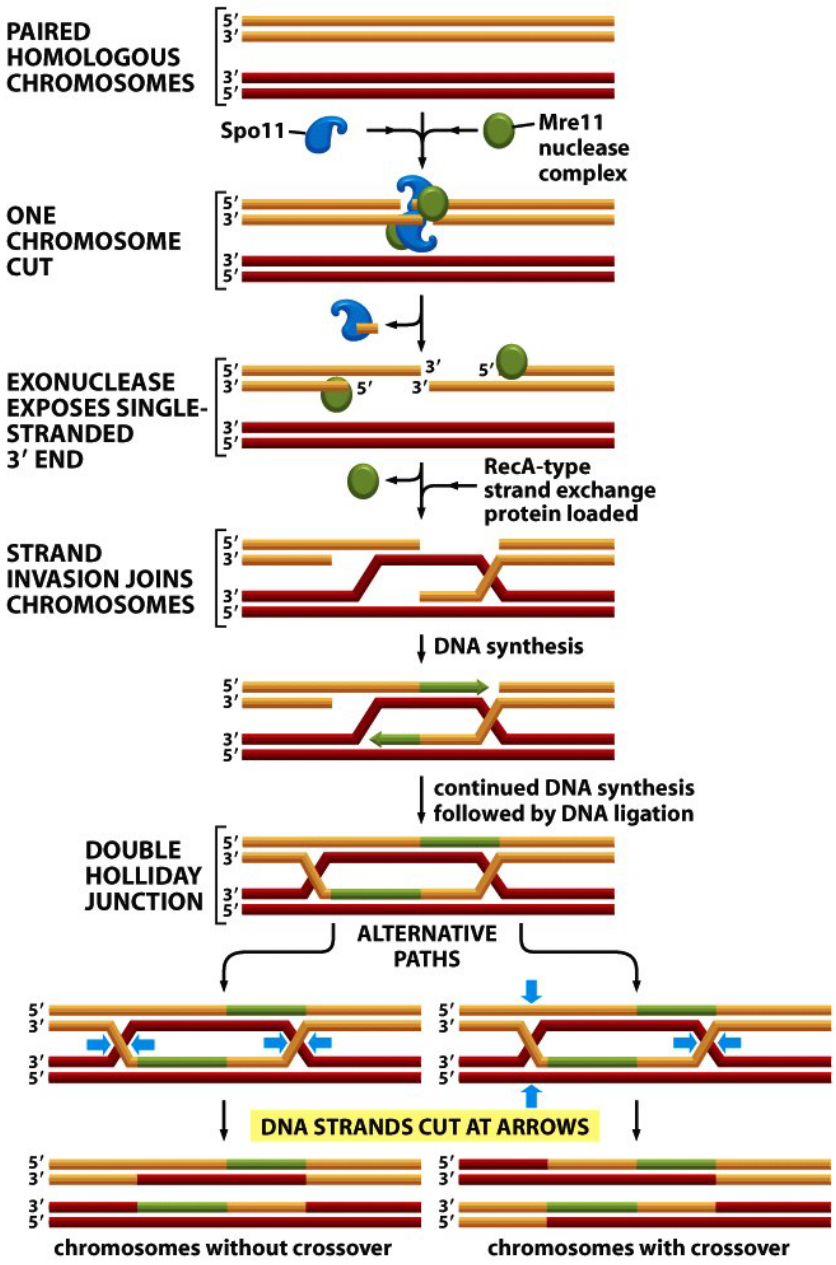


Figure 1.5: Double Holliday Junction Model for homologous recombination can lead to crossover and non-crossover outcomes (Figure is reproduced from Alberts et al. [2007]). Out of the four chromatids in a bivalent, only two homologous chromatids are shown (orange and red) for clarity. Each line represents one of the two strands of the double helix of a chromatid.

The model is described in detail in Section 1.2.3. Briefly, a double-strand break (DSB) is made by Spo11 on a chromatid (orange), called the initiating chromatid. The DSB is recognized by the MRE11 complex which triggers end resection. A 3' overhang of the initiating chromatid, loaded with the RecA protein, invades the template chromatid (red), resulting in the formation of a 'D-loop' (red). DNA is newly synthesized (green) on the initiating strand off the template chromatid, followed by annealing of the resected ends, resulting in the formation of a Double Holliday Junction (dHJ). Resolution of the dHJ can produce either a crossover or gene conversion without crossover, as illustrated in Figure 1.6.

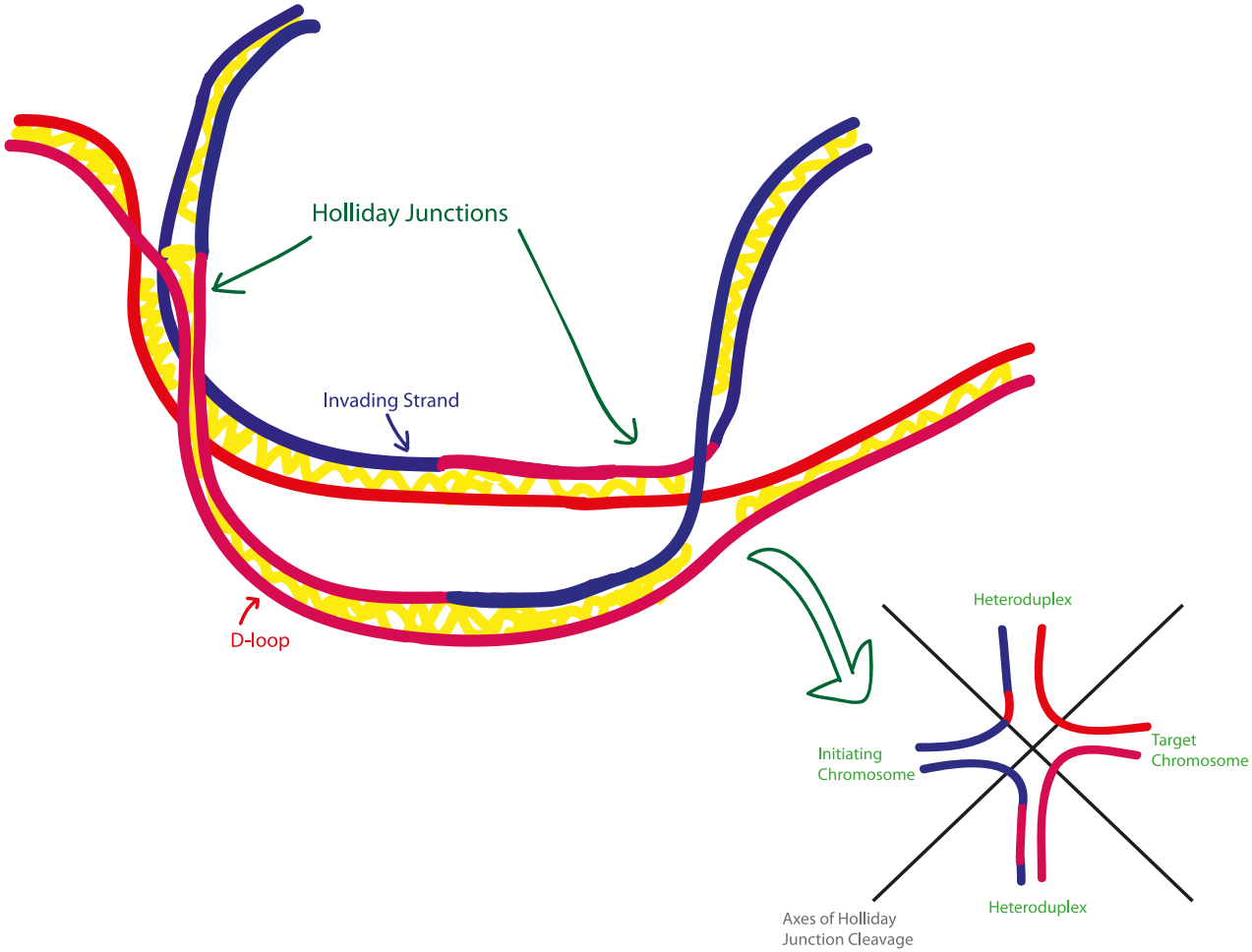


Figure 1.6: Double Holliday Junction cartoon for homologous recombination. The invading strand (blue) forces the formation of a D-loop (red) and undergoes gene conversion to fill the gaps created by the endonuclease (red segments within the blue strands). The yellow squiggles indicate base pairing interactions.
The bottom right picture shows a 'flattened out' version of the second Holliday Junction (HJ). Each HJ can be resolved in one of two ways – by cleaving along one of two axes shown and ligating the ends. Depending on the choice of axis for each HJ, the structure can be resolved with the same or different chromosomes on opposite ends of the two HJs with heteroduplex DNA in the middle, resulting in non-crossover and crossover outcomes, respectively.

also in the case of the non-crossover resolution. How the mismatched heteroduplex region (orange/red pairing in Figure 1.5 and blue/red pairing in Figure 1.6) between the two Holliday junctions is repaired is not fully understood, though in the wild-type it is usually resolved in favour of the template strand [Alani et al., 1994; Hoffmann, 2004; Mancera et al., 2008; Getz et al., 2008; Stahl and Foss, 2010].

As noted above, the initiating chromatid always undergoes gene conversion in the dHJ model. At the same time, the target (also referred to as template) chromatid is not left unaltered – it contains newly synthesized DNA (albeit to its own template) and may occasionally undergo gene conversion depending on the outcome of mismatch repair (Figure 1.5). It has been known for some time that many DSBs are not repaired in this way – they are non-crossovers and *directional* in the sense that the template chromatid is left completely unaltered while the initiating chromatid gets gene-converted [Gloor et al., 1991]. The models proposed to explain such outcomes are collectively known as *Synthesis Dependent Strand Annealing* (SDSA). The current view on how SDSA takes place involves one end of the initiating chromatid invading the target chromatid, followed by DNA synthesis using the target (Figure 1.7) [Ferguson and Holloman, 1996]. Crucially, however, the D-loop is not static, but small and moving, so that the newly synthesized strand gets displaced and eventually anneals with the second end of the initiating chromatid. Some data, however, support the idea of both ends performing strand invasion [Pâques et al., 1998]. Whatever the case, there is now considerable evidence that SDSA or an SDSA-like non-crossover gene conversion process plays a major role in the repair of DSBs in meiosis (Section 1.4.2).

Regardless of the choice of repair pathway – dHJ or SDSA – the alleles near the DSB site on the initiating chromatid are far more likely to be lost than the template chromatid. Therefore, if there is a systematic difference in the probability of one allele being subject to a programmed DSB than another due to its molecular properties, it follows that that allele is more likely to be lost from the population. This has

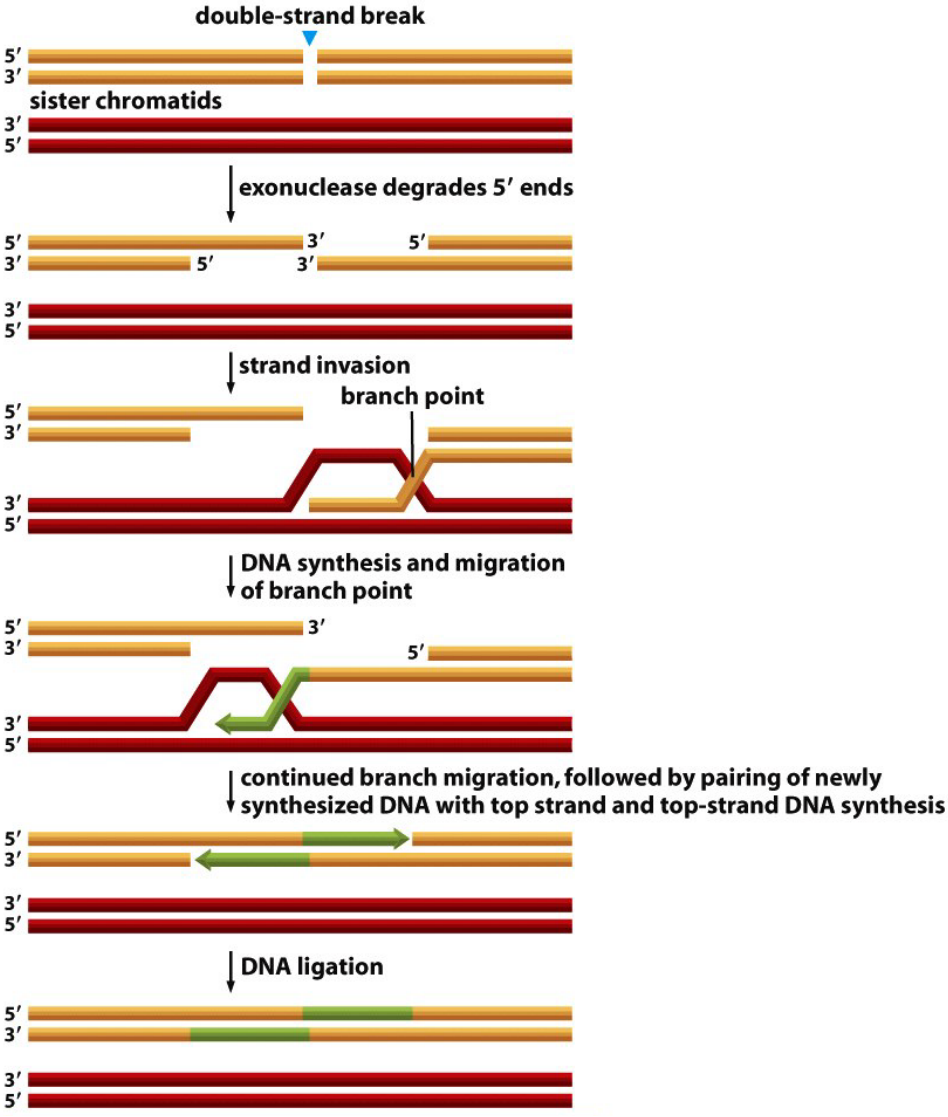


Figure 1.7: Model for Synthesis Dependent Strand Annealing

been observed *in vivo* [Jeffreys and Neumann, 2002] and is known as *biased gene conversion*.

Finally, there is evidence that some fraction of DSBs in *S. cerevisiae* and mice and nearly all DSBs in *S. pombe* which are resolved as crossovers may not result from dHJ formation, but from a different pathway involving single HJ intermediates [Osman et al., 2003; Argueso et al., 2004; McPherson et al., 2004].

1.2.4 The Synaptonemal Complex

The repair of meiotic DSBs in nearly all eukaryotes is accompanied by the formation of a meiosis-specific structure called the Synaptonemal Complex (SC) that mediates full pairing (or synapsis) of the homologous chromosomes [Handel and Schimenti, 2010]. The SC functions both to stabilize the pairing of the homologous chromosomes [Mlynarczyk-Evans and Villeneuve, 2010] as well as promoting the maturation of recombination intermediates into crossovers [de Boer and Heyting, 2006].

Figure 1.8 shows the formation and progression of the SC in the timeline of meiotic prophase I. It forms along the entire length of the pairing surfaces between the homologous chromosomes and has a tri-partite zipper-like structure by pachytene. The tri-partite structure consists of two "lateral elements" (LE) and a latticework of numerous "transverse filaments" (TFs) that comprise the central region of the SC. The lateral elements are proteinaceous axes and each lateral element supports the two sister chromatids of one homolog while the transverse filaments connect the axes of the two lateral elements into a single structure. To put this in the context of the timeline of recombination, in *S. cerevisiae*, DSBs occur prior to the appearance of SC and their interactions with matching sequences on the homolog (Section 1.2.2) bring the axial elements of homologous chromosomes into alignment during early to mid leptotene (Figure 1.8). Strand invasion is concomitant with the initiation of SC formation in late leptotene and double Holliday Junctions (dHJs) form during

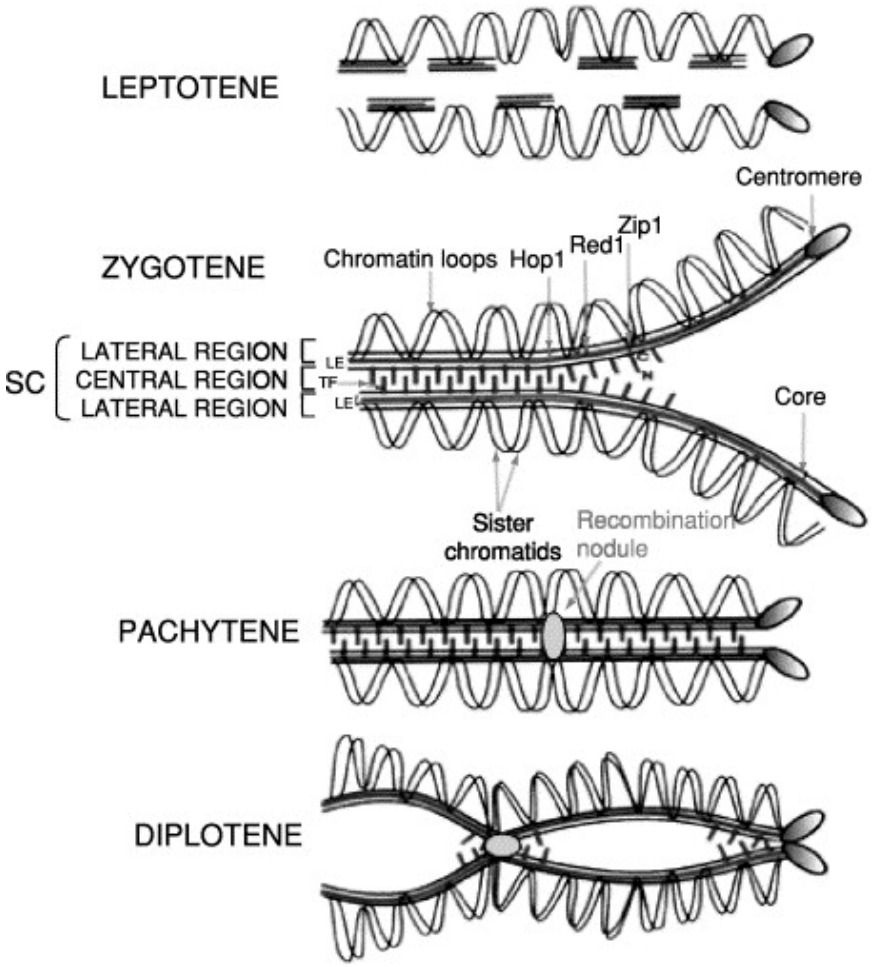


Figure 1.8: Synaptonemal Complex (SC) formation during Prophase I. During leptotene, sister chromatids start to get organized into chromatin loops along structures called axial elements which appear as discontinuous structures on chromatin. By zygotene, they become continuous to form lateral elements (LEs). Transverse filament proteins get assembled in zygotene resulting in synapsis and the SC is mature by pachytene. Sites of recombination are cytologically visible due to the formation of protein complexes, which are called recombination nodules. Chiasmata become fully visible only after the SC disintegrates in diplotene (reproduced from Anuradha and Muniyappa [2005]) .

pachytene. The resolution of dHJs into mature crossovers takes place at the end of pachytene. The SC disassembles in diplotene and the homologs often separate except at chiasmata. In most organisms studied, including *S. cerevisiae*, mice and humans the formation of DSBs is necessary for SC formation and complete synapsis; and evidence suggests that SC formation is initiated at sites that eventually become resolved as crossovers [Page and Hawley, 2004]. However, in some organisms, such as *C. elegans* and *Drosophila* females, other mechanisms have evolved for synapsis (nevertheless, they still require crossovers for faithful segregation).

Multiple lines of evidence suggest that the SC plays, either directly or indirectly, an important role in the formation of crossovers. Several proteins which have critical roles in the formation of the SC and, in some cases, are structurally part of the SC, also appear to influence the choice of directing the meiotic DSB repair pathway towards crossovers [Agarwal and Roeder, 2000; Manheim and McKim, 2003; Rockmill et al., 2003; Börner et al., 2004; Webber et al., 2004]. This is deduced from observations that defects in these proteins, for instance, the *ZMM* proteins (collection of genes acting in concert – *Zip1, Zip2, Zip3, Msh4, Msh5*, amongst others) and the *Sgs1* helicase, result in loss of crossovers, but do not prevent non-crossover events and the total number of gene conversion events does not appear to be changed significantly. Roles in the stabilization and resolution of Holliday junctions for some of these proteins have been proposed [Snowden et al., 2004; Jessop et al., 2006]. The direction of causality, however, is not entirely clear, and it is difficult to separate whether the primary goal of these proteins in promoting crossovers is to ensure the formation of a stable SC or the properties of the SC itself place constraints on the resolution of DSBs (or finally, neither). Kinetics of non wild-type recombination suggest the former may be more likely in *S. cerevisiae* (Section 1.4.2).

Another set of proteins which are important for ensuring correct structural properties of chromosomes during meiosis and also form part of the SC are cohesins and

condensins. There is evidence that the cohesin Rec8 influences the distribution of Spo11 along the chromosomes in S. cerevisiae [Kugou et al., 2009] and has a role in controlling whether the sister chromatid or the homolog are used as template for DSB repair. It has been found to influence the number of crossovers in cattle, which could be explained by either of these two roles [Sandor et al., 2012]. The condensin complex influences higher-order chromatin structure, and appears to have a role in influencing the number of DSBs, repressing repair on the sister chromatid and resolving recombination-dependent links between homologous chromosomes [Mets and Meyer, 2009; Yu and Koshland, 2003].

Finally, a strong association has been found between the physical length of the SC (in μm) and the number of crossovers in plants, grasshoppers, mice and humans [Kleckner et al., 2003]. As discussed above, each homologous chromosome organizes into chromatin loops tethered to its respective axis (Figure 1.8). These loops occur at an evolutionary conserved density of approximately 20 per μm of axis length [Zickler and Kleckner, 1999]. For DNA segments of the same length, a longer axis therefore implies a greater number of small loops and vice versa. Recombination is proposed to occur in DNA segments residing in chromatin loops. That would correspond, as observed, with greater recombination in regions with smaller loops and longer axes [Blat et al., 2002; Kleckner et al., 2003] (Figure 1.9).

1.3 Detecting and Measuring Recombination

Understanding the process and patterns of recombination is obviously of great biological interest. Recombination in humans, specifically, is of additional interest because of its importance in medical genetics. Knowledge of recombination rates is important in linkage mapping of diseases [Lathrop et al., 1984] and in the design and imputation-aided analysis of genetic association studies in unrelated individuals [Jorde, 2000;

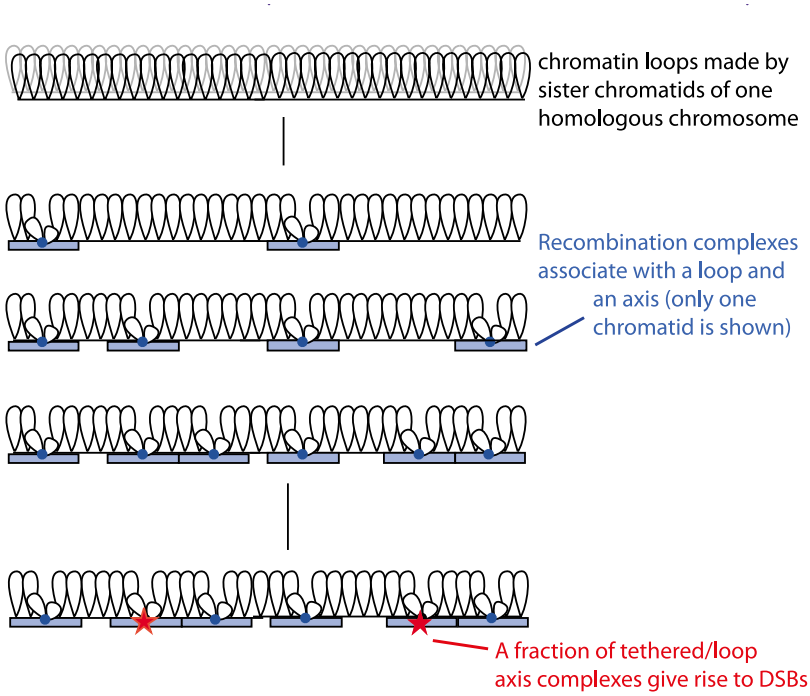


Figure 1.9: A model explaining the association of the number of crossovers with axis length given fixed loop density (adapted from Kleckner et al. [2003]).

Abecasis et al., 2005; Marchini et al., 2007]. Furthermore, the study of flawed recombination events, such as unequal crossover, has provided insights into pathological genome deletions, duplications and rearrangements in certain congenital conditions and cancers [Higgs et al., 1989; Pentao et al., 1992; Feunteun, 1998; Lupski and Stankiewicz, 2005; Lindsay et al., 2006; Raedt et al., 2006; Myers et al., 2008]. Insufficient or poorly localized recombination has been implicated in aneuploidies leading to conditions such as Down's syndrome [Oliver et al., 2008]. More generally, knowing how recombination influences DNA sequence diversity is essential for understanding the history of human populations and the dynamics of genome evolution [Pääbo, 2003; Lohmueller et al., 2009]. These applications require high resolution mapping of recombination events and genome-wide maps of recombination frequencies, known as *genetic maps*.

A variety of techniques have been used to detect recombination. Apart from

cytogenic approaches, they generally involve detecting polymorphisms (also called markers) that would signal a change in the DNA sequence inherited by a gamete or offspring from the parental sequences. Crossovers result in reciprocal exchange of extended DNA sequences and are relatively easy to detect. Non-crossovers, on the other hand, make highly localized changes and their detection requires greater density of markers and resolution of detection. More recently, sites initiating recombination have been identified genome-wide by targeting the binding locations of proteins involved in the early stages of recombination, such as *Rad51* and *Dmc1* (Section 1.2.2).

1.3.1 Sperm Assays

The first high-resolution study of recombination in humans was performed using an experimental assay called *sperm typing* that specifically targets recombinant molecules in sperm. In this approach, a specific locus in the genome is selected for analysis (this choice is often guided by patterns of breakdown of linkage disequilibrium, as discussed in Section 1.3.3). This region is subjected to high density discovery of polymorphisms, usually SNPs, because of their high abundance and low rate of recurrent mutation. Semen donors are typed in this region, and men with sufficiently many informative polymorphisms are selected for further analysis. The probability of a randomly chosen sperm to have undergone recombination at a particular locus is low, therefore this technique selectively amplifies recombinant molecules using *allele-specific* PCR (polymerase chain reaction, a method of amplifying a DNA molecule segment). Precise crossover breakpoints can then be mapped by typing internal SNPs (Figure 1.10). This strategy led to the identification of the first human recombination 'hotspot' at the molecular level [Jeffreys et al., 1998], and was an early indication that recombination may be concentrated in narrow segments 1-2 kb wide in humans as had been observed in *S. cerevisiae* (Section 1.4.4). The distribution of observed switch points has been very informative – in identifying the gene conversion tracts associated with

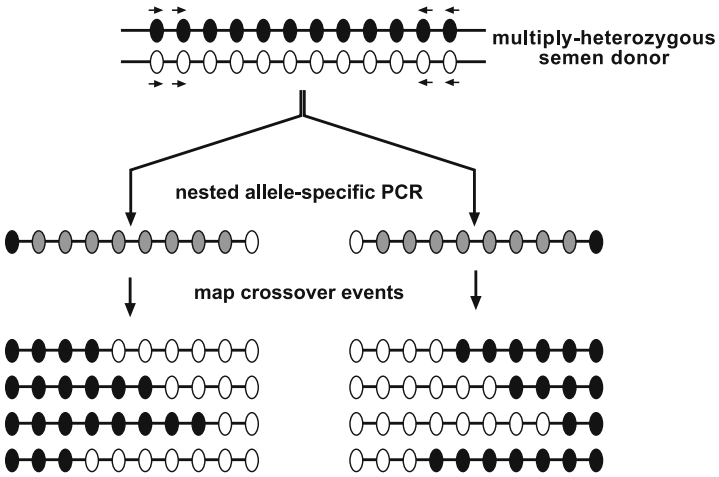


Figure 1.10: A candidate region likely to undergo recombination is selected for analysis and densely typed for SNPs. A man with heterozygous polymorphisms flanking and within the candiate region (grey) is selected and batches of his sperm are analyzed. Allele-specific PCR is performed to selectively amplify recombinant molecules (black → white or white → black). The crossover switch point can be localized by typing the internal SNPs (Figure adapted from May et al. [2007]).

crossover, which also led to the discovery that some haplotypes are more likely to initiate recombination than other [Jeffreys and Neumann, 2002] (Section 1.4.5). More recently, this strategy has led to the most detailed survey of non-crossovers in humans to date [Sarbajna et al., 2012] at a hotspot in the pseudo-autosomal region PAR2. More than a decade of findings using this technique are discussed later in this chapter. High-resolution sperm typing is considered to be the 'gold standard' in crossover detection, as the resolution is exquisite and does not suffer from biases that may influence other high resolution methods (Section 1.3.3). However, the method does not scale genome-wide (only about two dozen hotspots have been characterized using this method so far), and is obviously only informative about male recombination.

Another technique leveraging sperm for recombination analysis relies on the typing or sequencing of single sperm. Single sperm are isolated by dilution, microfluidic devices or cell sorting and a semen donor with sufficient genome-wide heterozygosity is selected. Whole-genome amplification is performed. In the past, this has been

followed by genotyping of regions of interest [Lien et al., 2000], which has now been superseded by the use of whole-genome next-generation sequencing [Wang et al., 2012]. This technique holds the promise of identifying the extent of intra-individual variation in recombination, and of finding the differences between sperm that lead to viable progeny and those that do not. However, this technique has been limited by the small number of sperm analyzed and the noisy nature of sequencing data.

1.3.2 Linkage of alleles and detection of recombination in pedigrees

In any given meiosis, one or more crossovers can occur on homologous chromosomes, resulting in gametes with alternating segments of paternal and maternal chromosomes (or grandpaternal and grandmaternal chromosomes from the gamete's point of view). We can think of this in terms of changing *inheritance* of the gamete as we go along the chromosome. Specifically, we can define the inheritance I at locus j as having two possibilities: the maternal or paternal chromosome. For two distinct loci on a chromosome, the probability of *inheritance change*[3] is the probability that the alleles descend from different parental chromosomes (or equally, they have different grandparents). Between two loci, j and j', the inheritance change probability is, therefore,

$$\theta = \Pr(I_j \neq I_{j'})$$

An inheritance change will occur when there is an odd number of crossovers between the loci j and j'. For loci that are very close together on a chromosome, the chance of crossing over between them in a typical meiosis will generally be low. Consequently, θ is close to 0 and alleles at those loci will be strongly dependent on their grandparental

[3]Probability of inheritance change is my term. The term used in the literature is recombination frequency, however, I find it confusing since I use the term recombination here to refer to the process, as opposed to a specific outcome.

origins. In general, θ tends to increase as the distance between loci increases. For loci that are far apart on a chromosome or are on different chromosomes, $\theta \to \frac{1}{2}$, indicating independent assortment of the alleles during meiosis and no information about the grandparental origin of one locus given the other. Loci for which $\theta < \frac{1}{2}$ are said to be *linked*.

The idea of linkage is utilized to work out the transmission of alleles in pedigrees, which can then be leveraged for identifying recombination or mapping of inherited diseases. In a general sense, calculating the likelihood of a pedigree's genotypes requires two distinct calculations:

- The likelihood of the genotypes of individuals at the top of the pedigree called *founders* (i.e., those without any parents in the pedigree).
- The transmission of alleles from the founders to the rest of the pedigree. In mammals, at least, every observed meiosis can be thought of as independent. Therefore, the genotypes of individuals are independent conditional on the genotypes of their parents.

This can be expressed as

$$\Pr(\mathbf{G}) = \prod_{founders\ i} \Pr(G_i) \prod_{nonfounders\ j} \Pr(G_j | G_{M_j}, G_{F_j})$$

where $\mathbf{G}$ is the set of all genotypes in the entire pedigree, G_i is the set of genotypes of individual i and M_j and F_j are the mother and father respectively of individual j.

This is essentially a latent variable problem. Individuals' genotypes may be partially or fully unknown or prone to measurement error and the allele transmitted by a parent to a child is unknown *a priori*. $\Pr(G_i)$ for the founders can be calculated using population allelic or haplotypic frequencies. $\Pr(G_j | G_{M_j}, G_{F_j})$, or the transmission probabilities from parents to child, depend on Mendelian segregation

and the probabilities of inheritance switches between loci. In the context of a latent variable problem, the probabilities of inheritance switches θ between any two loci are the unknown parameters, and lend themselves to estimation by the Expectation-Maximization (EM) algorithm or Monte Carlo methods. θ can, in turn, be used to model the probability of crossovers between the corresponding loci. In Chapter 3, I discuss a range of algorithms which have used this idea to detect recombination and build genetic maps.

The biggest challenge in building genetic maps using pedigrees is scale – the small size of human families and the low frequency of crossovers means that the number of meioses required to identify one crossover, on average, per kilobase (kb) of the genome is about 100,000. The richest pedigree-based maps in the last ten years have had just over 1,250 meioses typed on ∼5000 markers [Kong et al., 2002] and 728 meioses typed on ∼500,000 markers. In a big leap in scale, Kong et al. [2010] used over 15,000 parent-offspring pairs from Icelandic pedigrees and genotyped them at ∼300,000 markers. Their approach resulted in a genetic map with a resolution of a few tens of kilobases. The problem of resolution can be partially ameliorated in mice wherein crossovers and non-crossovers have been studied in F1 hybrids derived from suitably diverse and diverged inbred mouse strains [Cole et al., 2010].

Powerful features of pedigree-based analyses are that individual-level differences in recombination can be studied, and that they are informative about both male and female recombination. However, no high resolution maps exist for non-European populations. A linkage map previously made in Mongolians measured recombination between only 1,039 polymorphic microsatellite markers genome-wide [Ju et al., 2008], while a map built with African American, Mexican American and East Asian families uses only 353 microsatellite markers [Jorgenson et al., 2005].

1.3.3 Genetic maps using linkage disequilibrium

The small number of crossovers per meiosis limits the scale and the resolution of genetic maps that can be made using pedigrees. Considerably greater resolution can be obtained by using similar ideas at the population scale as discussed below.

In an infinitely large randomly mating population, in the absence of selection and recurrent mutation, allele frequencies of loci remain constant over time (Hardy Weinberg equilibrium). For the joint frequencies of two loci, however, this is not the case as crossovers can alter their transmission relative to each other. Let us denote the alleles at two biallelic loci as A and a, and B and b (Figure 1.11), and their frequencies by p_A, p_a, p_B and p_b. Two loci are said to be in *linkage equilibrium* (LE) if the observation of an allele at one locus has no information about the allelic state of the other locus, and vice versa. The frequency of the haplotype AB expected in a state of linkage equilibrium is therefore $p_A p_B$. A standard measure of *linkage disequilibrium* (LD) is the deviation from the expected LE frequency:

$$D = p_{AB} - p_A p_B$$

After one generation of random mating, in an infinite population, it can be shown that D decreases to

$$D^* = D(1 - \theta)$$

where θ is the inheritance change probability between these two loci in a single meiosis (defined in Section 1.3.2). Therefore, equilibrium haplotype frequencies are not obtained in a single generation even for *unlinked loci* ($\theta = \frac{1}{2}$). If linkage is tight ($\theta \approx 0$), LD can persist for thousands of generations. Measuring LD can, therefore, be used as an indirect way of measuring crossovers that have accumulated over thousands of generations. Figure 1.11 shows the popularly used measures of LD, called D' and r^2, both of which are transformations of D and have more favourable statistical prop-

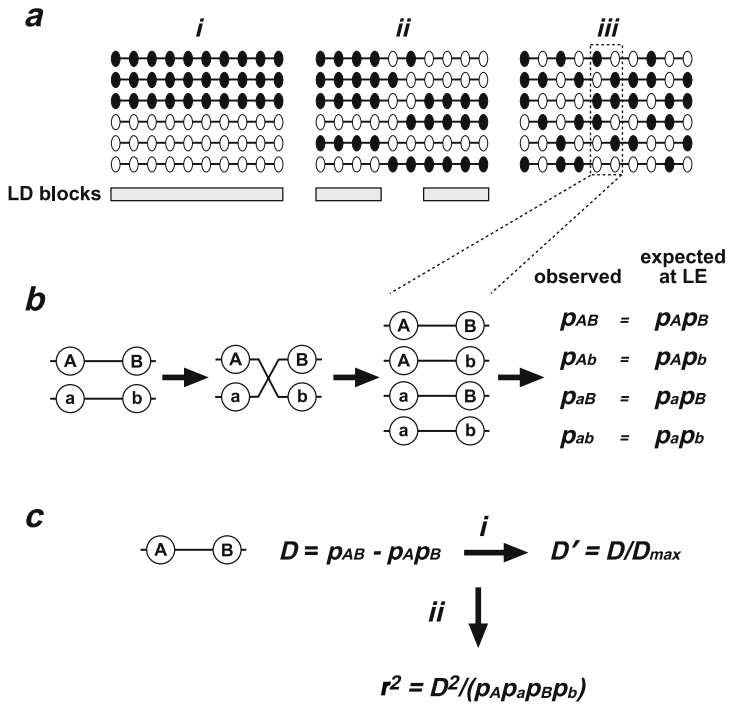


Figure 1.11: Measures of Linkage Disequilibrium (LD) (Figure adapted from May et al. [2007]).

(a) LD measures the degree of dependence in the population allele frequencies of two loci (details in text). (i) Strong association between all markers in a region makes them part of an *LD block*, and reflects lack of observable recombination in this region in the history of the population. (ii) LD blocks may be separated from each other by regions of high crossover rate and appear independently assorted (iii) All markers in a region may be in linkage equilibrium, suggesting a high crossover rate throughout the region.

(b) Two loci with alleles A and a, and B and b, respectively, can have four possible haplotypes. At linkage equilibrium (LE), the allele frequencies are expected to be independent.

(c) D is a measure of deviation from the expected equilibrium allele frequencies ($D = p_{AB} - p_A p_B$). Its value depends on the allele frequencies, making it a difficult measure to work with. (i) D' defined as

$$|D'| = \begin{cases} \frac{-D}{\min(p_A p_B, p_a p_b)}, & \text{if } D < 0 \\ \frac{D}{\min(p_A p_b, p_a p_B)}, & \text{if } D > 0 \end{cases}$$

rectifies this problem to some extent. Specifically, $|D'| = 1$ is complete LD. This happens if and only if no more than three of the four possible haplotypes are found in the population, and is compatible with a history of no recombination between the loci under the infinite sites model (i.e., under the assumption that a site may undergo mutation at most once). (ii) r^2 reflects the statistical correlation between the alleles.

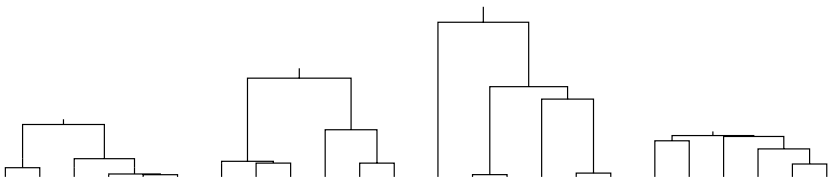


Figure 1.12: Coalescent trees reflecting four possible genealogical histories for a sample of six chromosomes at a single locus. The topology of each tree indicates the relatedness of the chromosomes in a specific history. The vertical axis represents the time it takes to find a common ancestor (backwards in time).

erties. In humans, the extent of LD can be tens to a few hundred kilobases, which means there is information to permit estimation of recombination rates at these scales using patterns of LD.

Early works in the sixties and seventies [Ohta and Kimura, 1969, 1971; Hill, 1975] attempted to provide estimators of quantities related to r^2 under a variety of simple population models. However, the estimators were extremely noisy and not especially informative about recombination between pairs of loci. These approaches modelled populations as a whole and then related the properties of the population to an observable sample (for example, the polymorphisms on a set of chromosomes). The use of *coalescent modelling* [Kingman, 1982a,b] changed the way genetic variation is analyzed. It focusses on the properties of the sample itself by considering the genealogical history that relate the samples to each other. This explicit modelling of the complex correlation structure imposed by family relationships has facilitated both simulation and inference and led to rapid advances in population genetics.

It is intuitive to think of the history of a specific locus on a set of chromosomes in the form of a genealogical tree. Looking back in time, lineages may find a common ancestor and *coalesce*. Figure 1.12 shows four possible genealogical histories for a sample of six chromosomes at a single locus. Loci adjacent to each other on a chromosome are likely to have similar histories. However, because of recombination, the histories may not be exactly alike. In fact, the key idea is that changes in genealogy

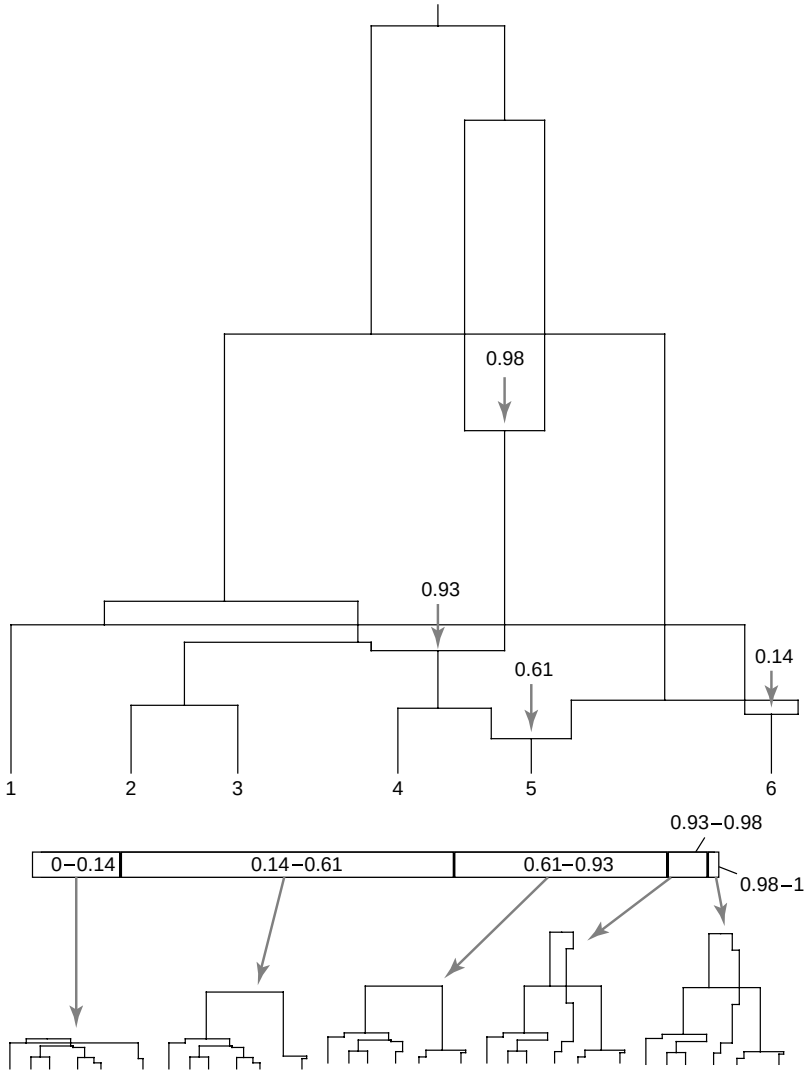


Figure 1.13: An example of an Ancestral Recombination Graph (ARG) (Figure from Nordborg [2007]). Loci on a chromosomes are represented from left to right as points in [0,1]. Four recombination events have occurred in this history (vertical grey arrows) at loci marked 0.14, 0.61, 0.91 and 0.98. The four recombination events imply that there are five embedded coalescent trees, shown underneath. It is worth noting that the final two trees are identical, in spite of the recombination event. In general, depending on the genealogical history and the observed mutations (mutations not shown), it may or may not be possible to detect a historical recombination event in a present-day sample.

happen only at recombination sites, and therefore recombination fundamentally alters the genealogical relationship between sequences. Therefore learning genealogical trees from the data helps learn about sites of recombination.

The histories of multiple loci can collectively be represented in an *ancestral recombination graph* (ARG). A lineage can branch into two lineages backwards in time reflecting different ancestries of loci prior to a recombination event (Figure 1.13). By relating genealogical history to the observed sample in this way, it is possible to calculate the likelihood of the sample conditional on a history. To calculate the full likelihood of a sample without knowing the genealogy (as is almost always the case), one has to integrate over all compatible genealogies. However, a very large number of histories may be compatible with an observed sample (in the presence of recombination, the number of possibilities is infinite). Fearnhead et al. [2004] showed that this *full likelihood* approach was promising; however, the computation becomes intractable for all but the smallest datasets and restricting assumptions. McVean et al. [2002] showed that a particular approximation to the full likelihood, called the *composite likelihood* [Hudson, 2001] was computationally scalable to large numbers of loci and samples. In this approach, the likelihood of pairs of loci are calculated separately and then combined as if independent. While intuitively surprising as an effective approximation, McVean et al. [2002] showed that the maximum composite-likelihood estimator was strongly correlated with the full likelihood MLE. Wiuf [2006] have shown the consistency of this estimator, if the underlying recombination rate is uniform, as the number of loci tends to infinity. However, whether the estimator is consistent when recombination rates are not uniform (Section 1.4.4), or as the number of samples increases, is less clear. In any case, this approach has proved exceptionally useful, as discussed further below. A different approximation, called the product of approximate conditionals (PAC) likelihood, is discussed in Chapter 2.

The composite likelihood approach turned out to be extremely powerful and pro-

duced the first exquisitely fine-scaled genetic maps [McVean et al., 2004; Myers et al., 2005]. This work leveraged the genetic diversity characterized in the HapMap project [International HapMap Consortium, 2005, 2007]. By HapMap's Phase 2 release, 60 individuals from Yoruba in Ibadan, Nigeria (YRI), 60 European ancestry individuals from the Centre d'Etude du Polymorphisme Humain collection (CEU), 45 Han Chinese from Beijing (CHB) and 45 Japanese from Tokyo (JPT) were genotyped at over 3 million SNPs and this data were utilized in a sophisticated statistical package called *LDhat* [McVean et al., 2004]. These maps were the first to demonstrate that recombination was concentrated in 'hotspots' throughout the genome (Section 1.4.4), with about 80% of recombination events occurring in no more than 20% of the genome [Myers et al., 2005]. They also showed that regions that were hot at a broad-scale were due to greater density and intensity of hotspots. Published maps for YRI, CEU and COMBINED (YRI, CEU, CHB, JPT all pooled together) populations are used extensively throughout this work.

These maps are the most fine-scale maps available today and have a large number of crossovers. Population genetic theory suggests that approximately 600,000 meioses have occurred in the ARG represented by the COMBINED map (corresponding numbers for YRI and CEU are around 300,000 and 200,000 respectively). However, not all crossovers that happen in the ARG leave effects that are visible in the present-day samples. Schierup et al. have simulated the effect of a *single* recombination event on Hudson's composite likelihood estimate of recombination [Hudson, 2001] under a model of human demographic history incorporating an out-of-Africa split, European bottleneck and post-split migration [Schaffner et al., 2005] over thousands of genealogies. The fraction of crossovers showing statistically significant LD breakdown was approximately 12% in both Africans and Europeans, while the estimated scaled recombination rate (measured as Hudson's composite likelihood estimator of the population-scaled recombination rate) depended strongly on the number of SNPs,

and samples used and the underlying demographic model. Another question concerns the age of recombination events incorporated in the maps. As we will see in Section 1.4.5, recombination patterns are evolving rapidly and it is useful to compare present-day recombination patterns with ancestral patterns. Based on the analysis by Schierup et al. on the impact of a single crossover on the estimated recombination rate going back in time (reproduced in Appendix A), I have calculated[4] that approximately 60% of the recombination rate estimate is contributed by crossovers after the out-of-Africa split (under a model of constant population size), which are private to each of the two populations. The calculation also indicates that approximately 50% of the estimate is based on crossovers that happened more than 2000 generations ago.

The major challenge in using LD-based maps as reliable measurements of past recombination is that recombination is only one of many factors that affect LD. LD is influenced by recurrent mutation and demographic processes such as natural selection, genetic drift, population bottlenecks and admixture. The impact of recent severe bottlenecks and selective sweeps can be particularly strong. For example, consider the ARG in Figure 1.13, and assume that a population bottleneck affects lineages numbered 1-6 forward in time. If only lineages 2 and 3 survive the bottleneck, all history of recombination in this sample has been lost, and no hotspots can be inferred anywhere along the sequence. If lineages 2 and 3 had been long-lived at the time of the bottleneck and therefore accumulated mutations, then the branch-specific mutations would be perfectly correlated and lead to highly inflated estimates of LD. If, on the other hand, lineages 4, 5, and 6 survive, the inference will be quite different. The recombination event at position 0.61 changes the topology of the

[4]I must emphasize that Schierup et al.'s analysis specifically examines the effect of a *single* recombination event on rate estimates. How recombination rates are influenced in realistic scenarios of large numbers of crossovers in a history is not known. The calculations I present here are based on extrapolating from the single crossover case and assuming a constant-sized population, by weighting each event using the results in Appendix A and in proportion to the probability of observing an event when different numbers of lineages remain in the history. These calculations are presented only to give an approximate sense of the decay of visibility of recombination back in time in LD-based maps.

tree, and since all three lineages are long-lived, enough mutations are likely to have accumulated to result in strong evidence of LD breakdown and inference of a hotspot, even though only event is observed in the sample. Thus, bottlenecks can increases the variance in estimating the recombination rate dramatically. Selective sweeps also lead to rapid coalescence near the tips of a genealogical tree, with long branches prior to the selective sweep. These branches are likely to contain neutral mutations which increase in frequency together will the selected allele (hitchhiking), and which will be in high LD with it. In principle, therefore, selective sweeps can also produce patterns similar to population bottlenecks due to their shared characteristic of a small number of long branches in the genealogy. Genetic drift and background selection also influence LD, however, their impact is much more modest. These effects lead to the loss of one or a few lineages repeatedly over extended periods of time, resulting in apparently accelerated coalescence of the remaining lineages in the genealogy. The result is an increase in linkage disequilibrium, which can be summarised adequately by a change in the effective population size [Zeng and Charlesworth, 2011].

A further caveat is that LD-based maps are sex-averaged and cannot be used to study individual differences in recombination.

1.3.4 Immunoprecipitation followed by sequencing

This technique identifies the binding sites of DNA-binding proteins that play critical and highly localized roles in recombination. Thus far, proteins that have been targeted for analysis include *Spo11*, *Rad51* and *Dmc*1 (their roles are discussed in Section 1.2). This technique involves capturing a DNA-bound protein using a specific antibody, separating the protein from the DNA to which is was bound, and sequencing the DNA. In species where a highly-quality map of the genome exists, this DNA sequence (if sufficiently long) can then be mapped back to a specific location on the genome, thereby localizing a site where recombination was initiated. A schematic of this

technique is shown in Figure 1.14.

As discussed in Section 1.2.1, early processing of programmed DSBs involves endonucleolytic cleavage of DNA adjacent to the Spo11-DNA complex. This liberates Spo11 bound to a short oligonucleotide [Neale et al., 2005]. Pan et al. [2011] immunoprecipitate these Spo11-bound oligonucleotides in *S. cerevisiae* meiotic cells, followed by amplification and deep sequencing. They used this data to localize *S. cerevisiae* crossovers to within a few nucleotides. ChIP-seq of *Rad51* and *Dmc1* in male mouse testes have now been used to generate the finest-scale mammalian maps of recombination initiation [Smagulova et al., 2011; Brick et al., 2012]. They have mapped hotspot centres within 200bp and used them to find novel molecular features of mammalian recombination. The findings of these experiments are discussed in the rest of this chapter.

Advantages of this technique are the very high resolution of mapping and the ability to identify recombination initiation (most other technique identify the outcome). These ideas can, in principle, be extended to humans though spermatocytes are required (not simply sperm). They are at present informative only about male recombination in mammals and need to be used in conjunction with crossover maps to track the control of recombination from initiation to outcome. Individual-level differences are, at present, not discernible. Further, it is not clear if this method can be used to *count* the number of DSBs per locus, since the ChIP procedure is, in general, affected by sequencing and mapping artefacts.

1.4 Localization, control and evolution of recombination

Study of the distribution of crossovers as well as extensive study of mutants in various model organisms has shown that the number of crossovers is under many levels of

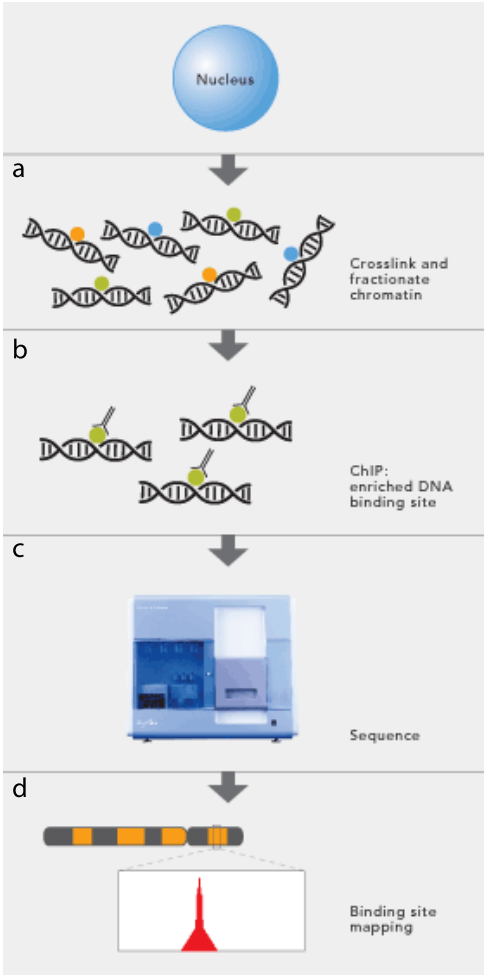


Figure 1.14: Chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq). **a.** Proteins are cross-linked to DNA using formaldehyde and DNA molecule is fractionated into smaller pieces. **b.** The target protein (green) is recovered using an antibody. **c.** DNA bound to the target protein is recovered and subjected to sequencing. **d.** The sequence is mapped back to a reference sequence for the organism, to localize the original binding site (Figure from http://www.illumina.com).

control. One way in which these controls manifest themselves is in achieving a balance in the number of crossovers – neither too many, nor too few. This is called "Crossover Homeostasis" and describes the fact that, in most organisms studied, the number of crossovers is more strongly clustered around a mean value than would be expected by chance There are two aspects to crossover homeostasis. The first is 'crossover assurance' and ensures a minimum number of crossovers (Section 1.4.1). The second reduces the probability of too many events above the mean value, and is called 'crossover interference'. It applies both locally – occurrence of a crossover reduces the likelihood of crossovers nearby, and globally – the total number of crossovers per homologous pair. Crossover interference is likely mediated by the choice of repair pathway, discussed in Section 1.4.2. Following that, I discuss the intense clustering of crossovers in narrow segments of the genome, known as recombination hotspots, and other known deviations from randomness in hotspot number and location.

1.4.1 Crossover assurance

The number of crossovers, especially in the shorter chromosomes, of many organisms are 1-2 per homologous chromosome pair [Coop and Przeworski, 2007; Dumas and Britton-Davidian, 2002]. If the distribution of the number of crossovers was from a random memoryless distribution (Poisson), there would frequently be no crossover at all, presumably with the associated risk of non-disjunction. This is rarely seen in wild type organisms[5], in whom at least one *obligate crossover* is observed per homologous pair. This effect can be quite robust – in *C. elegans*, a single DSB is generally sufficient to form the obligate crossover [Rosu et al., 2011]. In yeast and mammals, the number of crossovers is generally well explained by a linear model with one obligate crossover per chromosome or chromosome arm, plus a term dependent

[5]Exceptions include the tiny 4^{th} chromosome and occasionally the X chromosome in Drosophila females [Orr-Weaver, 1995; Koehler and Hassold, 1998] and possibly, occasionally, a small subset of chromosomes in humans [Fledel-Alon et al., 2009]

on the physical (base-pair) size [Mancera et al., 2008; Coop and Przeworski, 2007].

Crossover assurance between and segregation of sex chromosomes in the heterogametic sex of higher eukaryotes is discussed in Section 1.4.8.

1.4.2 Crossover interference and the crossover/non-crossover decision

Non-crossovers (NCOs) can form via at least two DSB repair pathways – double Holliday Junction (dHJ) resolution and Synthesis Dependent Strand Annealing (SDSA), as discussed in Section 1.2.3. Studies of the kinetics of recombination as well as mutants in *S. cerevisiae* suggest that most NCOs precede crossovers (COs) by about an hour and few, if any, are the outcome of dHJ resolution [Allers and Lichten, 2001; Hunter and Kleckner, 2001]. SDSA is now thought to cause the majority of NCOs (the total number of COs and NCOs in *S. cerevisiae* are comparable [Mancera et al., 2008]). Börner et al. [2004] analyzed *S. cerevisiae* recombination at high temperature and observed that DSBs and NCOs formed normally while strand invasion intermediates, dHJs, COs and the SC were all defective. They deduced that the choice of CO/NCO pathway is made very early – at or around the time of DSBs and prior to stable strand exchange and before and independently of the SC. They also noted that mutations in the *ZMM* proteins (Section 1.2.4) specifically block the CO pathway. Finally, it was noted that interference is specific to COs[6] and that NCOs did not interfere with each other (CO/NCO interference with each is touched upon below).

Crossover interference has also been observed in humans and mice [Broman and Weber, 2000; Lawrie et al., 1995]. However, there are significant differences relative to *S. cerevisiae*. In mice and humans, only a small fraction of DSBs, estimated at about

[6]The phenomenon of crossover interference appears to be specific to crossovers resulting from the dHJ pathway. The second pathway (Section 1.2.3), which involves single Holliday Junction intermediates and is responsible for most crossovers in *S. pombe* and a fraction of the crossovers in *S. cerevisiae* and mice, does not exhibit interference.

10% for mice and between $10 - 30\%$ for humans, get resolved as COs [Baudat and de Massy, 2007]. Further, NCOs show the same kinetics as crossovers in mice [Guillon et al., 2005] (they do not appear early, in contrast with *S. cerevisiae*). There is, nevertheless, convincing molecular evidence that different pathways are involved in CO and NCO outcomes. First, the size of the gene conversion tract associated with CO is estimated to be about 500 base pairs [Jeffreys and Neumann, 2002; Cole et al., 2010] in sharp contrast with the maximal possible extent of NCO gene conversion tracts observed in one study to range between 16-117 bp [Cole et al., 2010]. Second, only the CO pathway was found to require the mismatch repair *Mlh1* protein [Guillon et al., 2005], which is specifically implicated in the resolution of dHJs [Wang and Kung, 2002] and has traditionally been used as a proxy for crossovers because *Mlh1* foci are cytologically visible [Anderson et al., 1999].

In contrast with *S. cerevisiae*, however, the *Msh4* protein (member of the *ZMM* group discussed in Section 1.2.4 and thought to stabilize strand invasion intermediates as well as dHJs) appears to be necessary for the resolution of NCOs in mice [Baudat and de Massy, 2007]. Correspondingly, NCOs in mice show a low degree of interference not observed in *S. cerevisiae* NCOs.[7] *Msh4*-dependent NCOs cannot, however, be the whole story as the number of DSBs in mice is estimated to be larger[8] [de Boer and Heyting, 2006]. The differences could be explained by an alternative NCO pathway, repair using the sister chromatid (Section 1.4.3), or simply an insufficient ability to detect all NCOs.

Two, not necessarily mutually exclusive, roles have been proposed for crossover interference: it may be necessary for the correct formation of the SC [Börner et al.,

[7] In contrast with the traditional view, Mancera et al. [2008] observed interference between COs and NCOs in *S. cerevisiae,* albeit smaller than the interference shown between COs. Both CO/CO and CO/NCO interference, however, disappeared in *msh4* null mutants. Their interpretation of this data is that there are two types of CO. An alternative explanation that might be consistent with the mouse data, however, is that there are two types of NCO: those that do or do not depend on *Msh4*, and which have different relative frequencies in *S. cerevisiae* and mice.

[8] It appears to be an open question whether DSB locations show interference

2004] and/or to ensure proper disjunction by causing chiasmata to be placed such that the homologs are adequately held together prior to anaphase I, i.e., enough cohesive force between sister chromatids is transmitted through to the homologs [van Veen and Hawley, 2003]. It is notable that species that do not form SC, such as *Aspergillus nidulans* and *S. pombe*, also do not exhibit crossover interference [Egel, 1995].

1.4.3 Choice of sister chromatid or homolog for DSB repair

Despite considerable overlap in the mechanisms and enzymes involved in repairing DSBs in meiosis and mitosis, the choice of repair pathway is regulated very differently in them [Andersen and Sekelsky, 2010]. In mitosis, the repair is strongly biased towards using the sister chromatid as template, presumably because replication is the intent of mitosis and repair using the homolog is more likely to cause chromosomal rearrangements and increased homozygosity (of potentially deleterious mutations). In meiosis, where chiasmata are important, the requirements are different and inter-homolog repair is preferred. The extent of this preference is not well understood: recent estimates in *S. cerevisiae* suggest anywhere between two-thirds of DSBs being repaired on the homolog [Goldfarb and Lichten, 2010] (i.e., close to what might be expected in an unbiased system since there is one sister and two homologous chromatids) to nearly all DSBs repaired by inter-homologous recombination [Pan et al., 2011; Mancera et al., 2008].

Homology search in meiosis requires two proteins, *Rad51* and *Dmc1*, and both are orthologs of bacterial *RecA* (Section 1.2.2). *Rad51* is widely conserved and is critical for recombination in both mitotic and meiotic cells. An enduring mystery has been to understand the role of *Dmc1*, which is meiosis-specific and *Dmc1* mutants display reduced or absent meiotic recombination in many organisms [Shinohara and Shinohara, 2004; Hyppa and Smith, 2010]. It has been speculated that *Dmc1* may play a role in suppressing repair on the sister chromatid (as preferred in mitosis)

and promoting inter-homolog repair. Studies in *S. pombe* suggest that, in *Dmc1* null mutants, DSBs are repaired rapidly and faithfully. However, the sister chromatid is used as the template several times more often than in the wild type. In *S. cerevisiae*, *Dmc1* is required for virtually all inter-homolog recombination and *Dmc1* mutants have a severe meiotic defect, suggesting models involving a 'barrier to sister repair' complex of proteins [Niu et al., 2005]. A variety of other proteins, such as the *Rec8* cohesin, condensins, and proteins involved in the formation of the SC have also been implicated in template choice [Couteau et al., 2004; Kim et al., 2010; Pradillo and Santos, 2011].

1.4.4 Hotspots of Recombination

In all eukaryotes studied thus far at high resolution and in large numbers of meioses, crossovers have been found to be distributed highly non-uniformly in the genome. They cluster in narrow segments 1-2 kb wide called *hotspots,* both at the individual level and at the population level. Hotspots have so far been demonstrated in S. cerevisiae [Pan et al., 2011], S. pombe [Steiner and Smith, 2005], mice [Paigen and Petkov, 2010], chimpanzees [Winckler et al., 2005; Auton et al., 2012] and humans [Jeffreys et al., 2001; Myers et al., 2005].

In humans, Myers et al. [2005] used LD-based methods (Section 1.3.3) to identify around 33,000 hotspots that are or have been active in at least two human populations. Further, approximately twenty human hotspots have been characterized using high-resolution sperm typing (Section 1.3.1). These approaches suggest that the vast majority of crossovers in humans occur in hotspots. Estimate of the mean fraction of crossovers in LD-based hotspots among the Hutterite population is approximately 60% [Coop et al., 2008], however this may be an underestimate due to insufficient power to call all hotspots. Sperm-typing studies [May et al., 2007] suggest that as many as 95% of all crossovers occur in hotspots. That said, apart from the centromeric

regions (which are depleted in recombination), no autosomal region greater than 200 kb in size has been completely lacking in historical recombination [Myers et al., 2005].

Hotspots vary significantly in intensity (several orders of magnitude) from each other [Myers et al., 2005], and also between individuals [Sarbajna et al., 2012]. However, they also share similar general properties – crossover resolution points are smoothly and symmetrically distributed across the hotspot centre, which can usually be localized within tens of bases. 95% of crossover resolution points fall within a $1-2$ kb interval, and this distance may reflect the stochastic nature of branch migration (Section 1.2.3). Only one hotspot characterized to date shows a skewed crossover distribution [Jeffreys and Neumann, 2005], and it is hypothesized that this is due to a palindromic minisatellite located within the hotspot that might perturb the processing of recombination intermediates.

1.4.4.1 Biochemistry of hotspot localization

A major step in understanding how hotspot locations are specified came with the observation that an epigenetic modification, H3K4me3 (trimethylation of histone H3 on lysine 4), is an important mark for the initiation of recombination in *S. cerevisiae* [Borde et al., 2009] and mice [Buard et al., 2009; Smagulova et al., 2011]. In addition to the H3K4me3 mark, Buard et al. [2009] found that H3K4me2 (dimethylation of histone H3 on lysine 4) and H3K9ac (acetylation of histone H3 on lysine 9) epigenetic marks are also enriched on the chromatid initiating recombination at a particularly hot mouse hotspot, and that these marks are not due to post-DSB processing. Smagulova et al. [2011] found that H3K4me3 marks associated with hotspots were distinct from other H3K4me3 marks and specific to the testes in male mice.

Myers et al. [2005] and Myers et al. [2008] showed that several DNA sequence motifs and repeat features are enriched in hotspots, which may be bound by a trans-acting factor that activates recombination. In the last few years, multiple lines of

evidence have confirmed the role of the gene *PRDM9* in specifying hotspot locations in several mammals including humans, mice and cattle. I review this in detail in Chapter 5.

Smagulova et al. [2011] showed that recombination-initiating DSB locations are also highly clustered. Further, they found evidence that local nucleosome occupancy profiles are co-centred with DSB hotspots, suggesting that hotspot locations may have an intrinsic preference to be occupied by a nucleosome. They also observed a purine-pyrimidine nucleotide skew, which changes polarity at the centre of the hotspot, and is hypothesized to be due to biases in the DSB repair process.

Finally, recombination rates are elevated in 5' and 3' UTRs (untranslated regions adjoining genes) and in CpG islands in humans and chimpanzees [Kong et al., 2002; Auton et al., 2012]. A role for local chromatin state has been proposed to explain these observations [Pan et al., 2011].

1.4.4.2 Recombination and transcription

The epigenetic modifications associated with recombination hotspots, namely H3K4me3, H3K4me2 and H3K9ac, are all marks that are associated with actively transcribed regions in the genome. In mice, DSB hotspots are significantly enriched in genes (defined from start to stop codons, including introns). The opposite pattern has been observed in humans, where crossover rates are lower in genes on average [Myers et al., 2005; Kong et al., 2010] and increase with distance from the gene in both 5' and 3' directions before decreasing again [Myers et al., 2005]. Sex-specific differences have been suggested in exonic and intronic rates of recombination [Kong et al., 2010]. Brick et al. [2012] report that PRDM9$^{-/-}$ mice, which are infertile, have the majority of their DSBs near promoters. Finally, a negative correlation has been observed between the rate of transcription of genes in meiotic tissues and crossover activity in their vicinity [McVicker and Green, 2010]. However a causal relationship cannot be

inferred. These studies together suggest an intriguing but still mysterious relationship between transcription and recombination.

The picture in strikingly different in *S. cerevisiae*, where the vast majority of hotspots (82%) overlap promoters [Pan et al., 2011], and map to promoter associated nucleosome depleted regions (unlike mice where there appears to a preference for very local nucleosome occupancy). This study found the nucleosome occupancy profile itself to be largely unchanged during early meiosis, and that the preference for promoters is largely explained by DNA accessibility. Unsurprisingly, DSBs were found to be depleted in *Rec8* binding sites.

1.4.5 Evolution of recombination hotspots

Wall et al. [2003] made an intriguing observation that that an intense human hotspot in the β-globin locus was absent in rhesus macaques. Two independent studies in 2005 further showed that fine-scale recombination patterns across multi-megabase regions are completely different in humans and chimpanzees despite high sequence identity [Winckler et al., 2005; Ptak et al., 2005]. Since then, further studies, Myers et al. [2010], Auton et al. [2012] have confirmed that there is no evidence of shared hotspots between humans and chimpanzees genome-wide.

Myers et al. [2010] also showed that a 13-mer motif, which has been found to be recombinogenic in humans (Chapter 5) is not correlated with increased recombination rates in chimpanzees. Furthermore, it is being removed by self-destructive drive in the human lineage, but not in the chimpanzee lineage. Self-destruction is a property we expect to see in a motif that triggers or stimulates DSB formation in close proximity in *cis* due to biased gene conversion, as discussed in Section 1.2.3. Motifs that enhance hotspot activity are likely to be replaced by less active alleles during DSB repair, resulting in the loss of hotspot locations over time [Coop and Myers, 2007]. Humans and other species nevertheless have intense hotspots, a fact which is now referred to

as the hotspot paradox. A possible resolution of this paradox is the replacement of hotspots genome-wide by the rapid evolution of hotspot alleles in the populations, potentially via the evolution of the gene *PRDM9*. Even though selective forces on PRDM9 are not well understood, phylogenetic analysis has shown that *PRDM9* is undergoing rapid, continuous evolution in many metazoans including rodents and primates [Oliver et al., 2009].

Although evolution of hotspots between species had been shown in previous work, the extent of differences in the recombination landscape between human populations, if any, was not known prior to the work described in this thesis.

1.4.6 The role of genetic variation

The role of the gene *PRDM9* in positioning essentially the entire landscape of an individual's hotspots is discussed in detail in Chapter 5.

The effect of genetic variation on the number of crossovers is also significant in humans: Fledel-Alon et al. [2011] have estimated a narrow-sense heritability of 0.25 for females, i.e., mother to daughter heritability, and 0.14 for males, i.e., father to son heritability. They found, however, no detectable heritability in this phenotype from mother to son or from father to daughter. Comparisons between several inbred mouse lines have also revealed significant differences in the number of crossovers in males [Koehler et al., 2002a; Dumont and Payseur, 2011].

Variants currently known to influence the number of crossovers, many of which are sex-specific, are discussed in Chapter 5.

1.4.7 Differences between males and females

In humans, females have a higher overall crossover rate than males, by a factor of approximately 1.6 in both African-American and European populations (Table 3.1). Mice also show somewhat greater recombination rate in females [Baudat and de Massy,

2007]. This corresponds with differences in the physical length of the synaptonemal complex (Section 1.2.4) between males and females [Lynn et al., 2002; Tease and Hultén, 2004]. The reason for difference in SC length remains unknown. Cattle, however, do not appear to show such a sex difference [Ihara et al., 2004]. The strength of interference seems to be similar in both sexes in mice and humans [Broman and Weber, 2000; Broman et al., 2002; Tease and Hultén, 2004]. Despite having fewer crossovers, human sperm show far lower degrees of non-disjunction than human oocytes. Aneuploidy rates are about $1-2\%$ for males and estimated to be about 20% for females, with about 5% of clinically recognized pregnancies being aneuploid [Hassold and Hunt, 2001; Hassold et al., 2007][9]. In general, there appears to be much greater *within individual* variation in the number of crossovers per gamete in females [Lenzi et al., 2005]. An association between maternal age and the number of crossovers has been observed in humans (European populations), however, the findings are contradictory in terms of the direction of the effect [Hussin et al., 2011; Coop et al., 2008; Kong et al., 2004a].

In mammals, the distribution of crossovers is also very different between males and females at the mega-base scale, and this is discussed further in Chapter 6. Male recombination is highly concentrated in the telomeric and sub-telomeric regions, and they have significantly reduced recombination near centromeres relative to females [Broman et al., 1998; Kong et al., 2002].

Despite the differences in broad-scale rates, hotspot locations are shared between the sexes with little evidence of completely sex-specific hotspots [Kong et al., 2010].

[9]Some of this difference is likely due to sex-specific differences in the efficiency of meiotic checkpoints. Several meiotic mutations that result in defects in synapsis and crossover cause prophase arrest in mice spermatocytes while at least some proportion of oocytes progress through [Morelli and Cohen, 2005].

In general, there are major differences between male and female meiotic progression in mammals. Spermatocytes are produced throughout male life after puberty, while meiosis is initiated in the *foetal* ovary. Oocytes enter prophase I, crossovers are formed,and meiosis progresses to diplotene. At this stage, meiosis is suspended and is only completed once fertilization occurs, which is likely to be decades later in humans [Hunt and Hassold, 2008].

This is, once again, indicative of multi-layered control of recombination in mammals.

1.4.8 Broad scale rates and the pseudo-autosomal region

Auton et al. [2012] showed recently that, despite little or no overlap in hotspot locations between humans and chimpanzees, broad-scale rates between the species are relatively conserved. The Pearson correlation coefficient $\rho = 0.1$ at the 10kb scale, while $\rho = 0.6$ at the 1 Mb scale. A similar pattern was observed in different strains of mice whose broad-scale patterns were also highly similar despite differing hotspot locations [Paigen et al., 2008]. This suggests either a biologically important role for broad-scale rates in meiotic processes such as synapsis or segregation or a currently unknown factor pre-disposing genomic regions to different levels of recombination for unrelated reasons such as chromatin organisation. Despite our increasing understanding of the differences within species at hotspots, the slower evolution of patterns at broad scales is not well-understood. A possible exception is the effect of large structural variants, such as inversions, which are known to influence recombination rates in *cis*. Current knowledge of broad-scale rate variation is reviewed in detail in Chapter 6.

A particularly important case of broad-scale control of rates is evident in the Pseudo-Autosomal Regions (PAR) on the X and Y chromosomes (or Z and W chromosomes in organisms with female ZW heterogamy) in most vertebrates [Livernois et al., 2011]. These regions are typically small – PAR1 is less than 3 Mb long in humans, and only about 700 kb in laboratory mice [Perry et al., 2001]. Nevertheless, an obligate crossover occurs in this region in the heterogametic sex to ensure proper disjunction (Mohandas et al. [1992], Section 1.4.1). While much remains to be understood about ensuring recombination in the PAR, studies so far implicate both genetic and epigenetic mechanisms in this process, and are reviewed in Chapter 6.

1.5 This work

Prior to the start of this work in 2009, chimpanzee and human recombination hotspots were shown to be almost entirely non-overlapping in several large orthologous regions, despite the fact that the two species share ∼98% of their sequence [Ptak et al., 2005; Winckler et al., 2005]. This demonstrated extremely rapid evolution of recombination hotspots, and raised the possibility that detectable differences might exist between human populations also. Ptak et al. [2005] also examined rates at a somewhat greater scale (50 kb), and proposed that there may potentially be weak conservation of rates between the two species at that scale. Between human populations, it was not known if recombination varied, and if so, at what scale.

A major aim of this work was to examine if recombination has evolved within the human lineage by investigating if it differs systematically between human populations. This required comparing genetic maps between populations who have differentiated sufficiently to allow manifestation of subtle differences in recombination (if any). High-quality maps in modern Europeans had been built, as discussed previously in this introduction, using large pedigrees in the Icelandic and Hutterite populations. However, comparable pedigrees have not been available in African or other non-European populations. Sperm typing is laborious and not applicable for examining genome-wide changes[10]. That said, very fine-scale LD-based maps were available for a West African population (Yoruba) and East Asians (Chinese and Japanese), in addition to one built in individuals of northern and central European ancestry [International HapMap Consortium, 2007]. However, interpretation of differences between LD-based maps from different populations is fraught, as LD is influenced by their demographic histories and the role of natural selection in these populations. Therefore, answering this question required a new map – one that was independent of ancient

[10]Sperm-typing hotspots that are highly active in many African men but are not active in European men are now known [Berg et al., 2011], and were published at approximately the same time as the work described in this thesis.

recombination and demographic influences, and that represented a human population with appreciable differentiation from Europeans.

To build such a map in the absence of large pedigrees from African or other non-European populations, I have used a novel approach. African Americans are descended from the admixture of individuals of West African and European ancestries in the Americas in the last several generations. Since the mixture is recent, African Americans have large contiguous stretches of their chromosomes composed of one of these two ancestries, and these chunks can be identified statistically. Switching of ancestry within a chromosome implies the occurrence of a crossover in an ancestor subsequent to the mixing of the two populations. A particular challenge in using crossovers identified in this way is that the data can be quite noisy. In Chapter 2, I describe how we identified these crossovers, and overcame the challenges to build a fine-scale map, called the AA map, with resolution finer than the best available pedigree maps. It also has the largest number of meioses of any genetic map in a contemporary human population.

Since the approach of using ancestry detection to identify crossovers and build a genetic map is novel, it was important to validate the approach using an independent method. To do this, I have built the first detailed pedigree-based map in African Americans, which contains over 1,000 meioses. The data I used had been collected for independent medical research, and in most cases consisted of a nuclear family with only one parent genotyped. Since existing methods could not handle such data, a novel approach was implemented, which is described in Chapter 3.

This pedigree-based map and sperm typing hotspots were used to validate the AA map, and to demonstrate its accuracy and high resolution in Chapter 4. In this chapter, I also explore biologically interesting properties of the map, such as its similarities and differences with different population-specific LD-based maps, and the degree of concentration of recombination in a fraction of the genome.

I was then in a position to use the AA map to investigate differences between the recombination patterns in Africans and Europeans. In Chapters 4 and 5, I give specific examples of hotspots that are highly active in Africans and not in Europeans. Further, in Chapter 5, I show that this is true genome-wide, that Africans carry thousands of hotspots not active in Europeans. However, there is no evidence that the converse is true, i.e., there do not appear to be hotspots active only in Europeans.

Another major aim of this work was to understand the biological drivers influencing recombination. The complete change between the recombination landscapes of humans and chimpanzees suggested the role of a fast-evolving *trans*-acting factor in specifying recombination, as did large differences in the activity of sperm-typing hotspots independently of local DNA sequence variation [Neumann and Jeffreys, 2006]. Asymmetry in the activity of certain hotspots in individuals heterozygous for particular mutations within those hotspots [Jeffreys and Neumann, 2002, 2005] also pointed to the role of a *cis*-acting factor. In 2008, the first human motif was discovered that was responsible for activating ∼40% of human hotspots in *cis* [Myers et al., 2008]. Multiple lines of evidence showed that a major *trans*-acting factor was PRDM9, which interacted with the motif via direct binding of DNA [Baudat et al., 2010; Myers et al., 2010; Parvanov et al., 2010; Berg et al., 2010]. However, understanding of the extent of PRDM9's role and its relationship with the motif was complicated by the observation that not all hotspots had the motif in them, and that subtle changes in the amino acid sequence had large and unexpected effects on hotspot activity [Berg et al., 2010].

To investigate these questions further, I mapped genetic variants influencing the choice of recombination landscape between the use of hotspots active only in Africans as opposed to those active in both Europeans and Africans. In Chapter 5, I describe how I constructed a highly sensitive phenotype that reflected, for each individual, what fraction of their crossovers belonged to each of the two hotspot landscapes. I

next used these phenotypes to perform a genome-wide association study (GWAS) to search specifically for factors acting in *trans*. This is an unusual study design because the phenotype reflects activity in the subjects' ancestors, as opposed to the subjects themselves. Nevertheless, I show that the design had power to detect true effects. I confirm the role of PRDM9 in specifying hotspot locations, and show how even subtle changes leave a mark on the hotspot landscape. I also show that PRDM9 appears to be the dominant factor responsible for changes in hotspots in *trans*. In parallel, my colleagues showed that the main genetic variant I identified in the GWAS above tagged significant changes in the DNA binding zinc finger (ZF) domain of PRDM9. I showed further that there is no evidence for shared hotspots among individuals with substantially different PRDM9 ZF variants. My colleagues leveraged the African-specific hotspots identified above to identify a motif that was significantly enriched in those hotspots, and was a close match to predicted binding motif of a common African-specific variant of PRDM9.

Control of recombination at broad scales is another important question introduced previously in this chapter, and one about whose control very little is known. Aberrant patterns of maternal recombination, particularly pericentromeric and telomeric crossovers are implicated in Down's syndrome [Oliver et al., 2008]. Comparison of maps of different mouse strains with non-overlapping hotspots had previously suggested megabase scale conservation of rates [Paigen et al., 2008]. Other than this work, the difference between closely related species or between human populations was not known. In Chapter 6, I show that broad-scale (3 Mb) rates are highly similar for the African-specific and European maps. I also describe the work I did to search for genetic variants affecting recombination rates subtly at scales greater than the size of hotspots. With the exception of *PRDM9*, which has a measurable effect even at the 4 megabase scale, I did not identify any variants that significantly influenced broad-scale rates in *trans*.

In this chapter I also discuss my work on understanding the landscape of recombination in the Pseudoautosomal region PAR1, site of an obligate crossover in males and the most recombinogenic multi-megabase region in the human genome. Recent work in mice [Brick et al., 2012] has suggested that PRDM9 may not be responsible for hotspots in the PAR, which would imply the existence of some other mechanism to position DSBs there. I find that the PAR is full of hotspots, although the 'background' rate of recombination also appears to be significantly higher than the autosomes. I also find suggestive evidence that PRDM9 plays a role in PAR1 recombination, however, this question will be explored further in future work.

I summarize the implications of this work in the next and final chapter, Chapter 7. Lots of interesting questions remain to be answered to fully understand the control of recombination and its evolution in humans and other species, and some of them are discussed in this chapter.

Chapter 2

Genetic map from unrelated African-American samples

In 2005, two independent research groups [Ptak et al., 2005; Winckler et al., 2005] showed that chimpanzee and human recombination landscapes are almost entirely different despite ∼98% sequence identity between the two species [Ebersberger et al., 2002]. This demonstrated rapid evolution of recombination hotspots in one or both species, and raised the possibility that human populations may also differ systematically in their recombination patterns. A key aim of this work was to examine if this was the case, i.e., if patterns of recombination had evolved in the human lineage, and if so, at what scales.

This required knowledge of genetic maps in multiple populations. Acquiring pedigree data with enough meioses to build a fine-scale map is very difficult – it would require genotyping thousands of families, and only one pedigree-based map is currently available with enough resolution to approach the size of a recombination hotspot. This map [Kong et al., 2010] was built in the Icelandic population, who are a European population, and has over 15,000 parent-offspring pairs. Comparable pedigrees have not been available in African or other non-European populations. As discussed

in Chapter 1, fine-scale LD-based maps have been available for the Yoruba population from West African and also for East Asians [International HapMap Consortium, 2007]. Since it is difficult to ascertain to what extent LD-based maps are influenced by factors other than recombination (such as natural selection) at any particular locus, it is difficult to use them to assess differences between populations. Secondly, a considerable fraction of the crossovers represented in the LD-based maps are likely to have occurred prior to the out-of-Africa split (Section 1.3.3).

The availability of genotype data of large numbers of mostly unrelated African Americans presented an opportunity. These data were collected for medical research on a variety of conditions in African Americans in the United States of America (USA). The sources for these data are described in detail in Section 2.1.

To detect crossovers and build a genetic map using *unrelated* individuals, I have used a novel approach that is rooted in the particular history of African Americans. African Americans are descended from individuals of West African and European ancestries in the Americas in the last several generations, as outlined below in Section 2.2.

The key idea is as follows: offspring of one European and one African parent would have equal numbers of entirely European and African chromosomes (Figure 2.1). Recombination breaks the chromosomes up over several generations (Figure 2.1), resulting in genomes that are mosaics of chunks of contiguous European and African ancestry. Since the admixture is only a few generations old, only a few rounds of recombination have happened, and therefore chunks of ancestry are likely to be large (tens of megabases). By identifying these chunks and points where their ancestry switches, we can identify sites of recombination since admixture. Therefore, if ancestry chunks can be accurately and precisely inferred in large numbers of individuals, we can use this information to build a map of recent human recombination in a population with majority African ancestry. Models and methods used to perform this

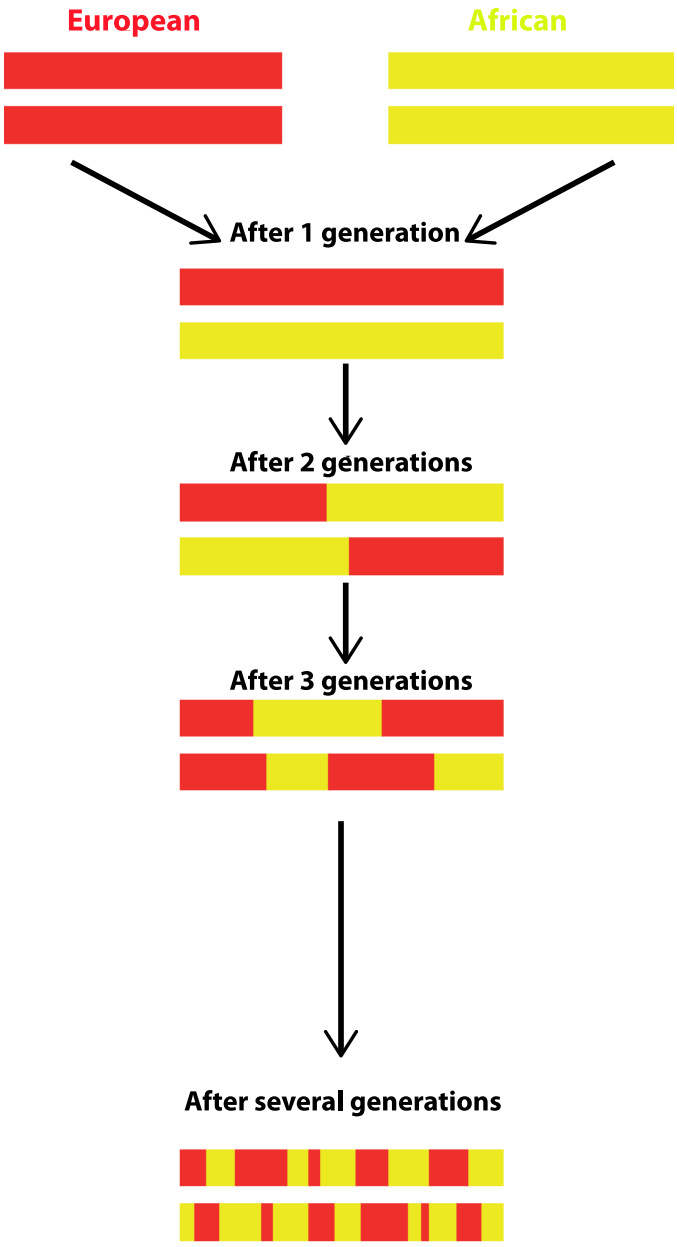


Figure 2.1: Admixed chromosomes are mosaics of blocks of contiguous ancestry. Blocks break down over time due to recombination. The switching points of blocks are therefore informative about the location of recent recombination, while the sizes of blocks can be used to estimate the number of generations since admixture. In this example, admixture has occurred in a single generation followed by panmictic mating.

inference are discussed in Section 2.3.

In Section 2.4, I describe how I detect crossovers using switches in ancestry segments inferred in nearly 30,000 African Americans. There is statistical uncertainty in the location of crossovers, and I discuss how we account for this uncertainty. I also describe the Bayesian approach we used to combine the information in approximately 2.1 million crossovers to produce a high resolution map in Section 2.5.

Finally, in Section 2.6, I assess the resolution of the map and the uncertainty in its rate estimates.

I validate the map, and use individual-level data described below, together with the map, to make biological inferences about human recombination in later chapters.

2.1 Details of Data

Genotype data from unrelated African Americans from the following five cohorts in the USA were used for this project. Informed consent was provided by all individuals participating in the study.

- *Candidate Gene Association Resource (CARe).* CARe is a consortium of cohorts including the Atherosclerosis Risk in Communities study (ARIC), the Cleveland Family Study (CFS), the Coronary Artery Risk Development in Young Adults study (CARDIA), the Jackson Heart Study (JHS) and the Multi- Ethnic Study of Atherosclerosis (MESA).

- *African American Breast Cancer Consortium (AABCC).* AABCC comprises several studies: Multiethnic Cohort study (MEC), the Los Angeles component of the Womens Contraceptive and Reproductive Experiences study (CARE), the Womens Circle of Health Study (WCHS), the San Francisco Bay Area Breast Cancer study (SFBC), the Carolina Breast Cancer Study (CBCS), the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Cohort (PLCO),

the Nashville Breast Health Study (NBHS) and the Wake Forest University Breast Cancer Study (WFBC).

- *African American Prostate Cancer Consortium (AAPCC).* This consortium is composed of the following studies: MEC, the Southern Community Cohort Study (SCCS), PLCO, the Cancer Prevention Study II Nutrition Cohort (CPS-II), the Prostate Cancer Case-Control Studies at MD Anderson (MDA), the Identifying Prostate Cancer Genes study (IPCG), the Los Angeles Study of Aggressive Prostate Cancer (LAAPC), the Prostate Cancer Genetics Study (CaP Genes), the Case- Control Study of Prostate Cancer among African Americans in Washington DC (DCPC), the Gene-Environment Interaction in Prostate Cancer Study (GECAP) and the Cancer Prevention Study II (CPS-II).

- *African American Lung Cancer Consortium (AALCC).* This study consists of individuals from the MEC, the SCCS, PLCO, the MD Anderson (MDA) African American Lung Cancer Study, the NCI-Maryland Lung Cancer Case-Control Study, the University of California at San Francisco African American Lung Cancer Study and the Wayne State African American Lung Cancer Study.

- *Childrens Hospital of Philadelphia (CHOP).* Individuals analysed from this study are drawn from a biobank for Philadelphia children developed by this paediatric hospital to facilitate diverse genotype-phenotype association analyses.

We sought to use only "unrelated" individuals to minimize the chance of finding ancestry segments (and therefore crossovers) that are identical by descent (IBD) in multiple individuals due to their shared ancestry. Therefore, I sought to remove closely related individuals. Relatedness was determined using *smartrel* package of the EIGENSOFT [Patterson et al., 2006] and individuals were filtered out from the study such that no pair of individuals in the study are second-degree relatives or

Consortium	Number of samples	Number of SNPs	Genotyping Array
CARe	6,209	580,000	Affymetrix 6.0
AABCC	5,203	894,717	Illumina 1M
AAPCC	6,540	896,036	Illumina 1M
AALCC	4,134	906,687	Illumina 1M
CHOP	7,503	491,572	Illumina 610-Quad / Illumina HumanHap550

Table 2.1: Summary of unrelated African-American samples (post-filtering) used in map building. Within each consortium, filtering was performed to remove related individuals and to include only those SNPs that showed good completeness across samples.

closer. The precise threshold on the proportion of genome-wide identity by descent (IBD) was 0.2 (relatedness of monozygotic twins is 1, and parent-child and sibling pairs is 0.5). The relatedness threshold therefore eliminated relationships as close as grandparent-child or uncle-nephew while retaining first cousins. After filtering, we had a total of 29,589 samples.

Table 2.1 summarizes the number of samples in each study and the number of SNPs available after data curation. Note that different samples are genotyped on different arrays, with differing numbers of overlapping and non-overlapping SNPs. We build our map on the SNP set obtained by the union of all these sets of SNP.

Cases and controls were included from all consortia, however, no phenotype information was available to us. The initial data curation described above was performed by Arti Tandon from the Harvard Medical School in association with members of each consortium.

2.2 African American populations

African Americans are populations in the USA who derive a substantial fraction of their ancestry from populations of sub-Saharan Africa. The African portion of their

ancestry traces back to individuals displaced from Africa due to the trans-atlantic slave trade. According to records compiled from slaving voyages [Lovejoy, 2012], expeditions had started as early as the early 15^{th} century, however, over 95% of enslaved individuals left Africa between 1601 and 1867. The scale of this forced migration was huge, with an estimated 12.8 million people forced out of western and west-central Africa [Lovejoy, 2012]. Of these, an approximately $4-5\%$ arrived in what is now the USA, with the majority taken to the Caribbean, Brazil, and the rest of South America [Rawley and Behrendt, 2005].

Previous genetic studies in African Americans (AA) have agreed with the historical record: AA populations from Chicago, Baltimore, Pittsburgh and North Carolina show a substantial portion of their ancestry (an average of $\sim 71\%$) from groups closely related to peoples of the Niger-Kordofanian language group and to other African populations ($\sim 8\%$) [Tishkoff et al., 2009]. Niger-Kordofanian languages are spoken by peoples throughout sub-Saharan West Africa and large parts of central and southern Africa. This language group includes the Yoruba language, which is spoken in Nigeria, Benin and Togo [Heine and Nurse, 2000]. Yoruba speakers from Ibadan, Nigeria, were genotyped as part of the HapMap project [International HapMap Consortium, 2005], and have been found to be effective in capturing the genetic diversity in African American populations [Zakharia et al., 2009; Price et al., 2009].

There is substantial individual variation in the overall composition of ancestry (Figure 2.2), with a mean of approximately 20% European and 80% African ancestry. Genetic studies have estimated that the number of generations since admixture began varies mainly between $4-8$ generations[1], which is consistent with the historical record [Smith et al., 2004].

As a result of the action of a variety of demographic processes such as drift, popu-

[1]An individual's different lineages will have different histories, and likely to have different admixture times or possibly no admixture. An individual's *admixture time* is estimated using the particular pattern of their own genetic make-up and is therefore naturally averaged over all of their lineages.

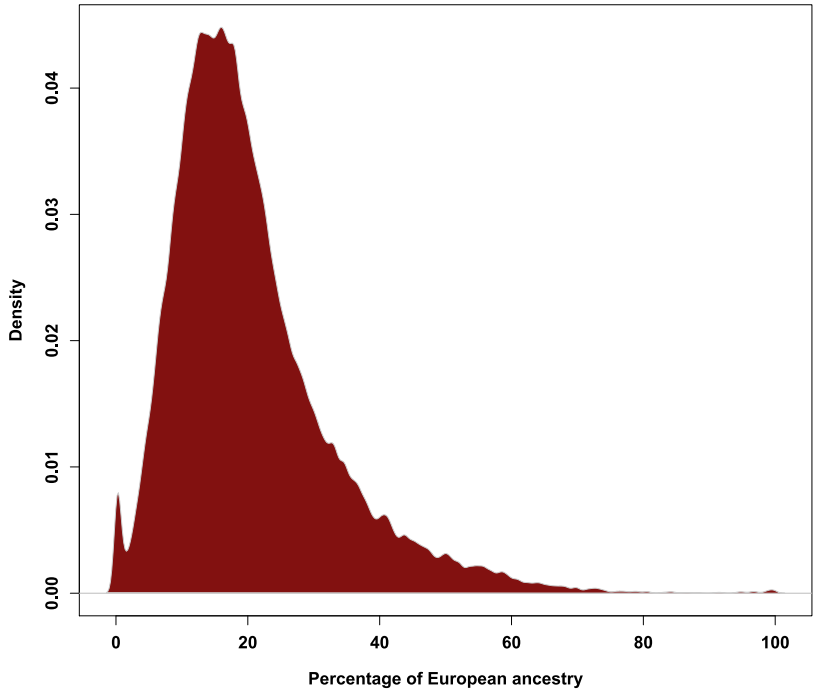


Figure 2.2: Distribution of the estimated fraction of European ancestry in about 30,000 African Americans from across the USA. HAPMIX [Price et al., 2009] was used to make the estimates for individuals described in Section 2.1. Some individuals show no evidence of admixture, while the mean is about 20% European ancestry.

lation bottlenecks and selection since the out-of-Africa migration ∼100,000 years ago, the patterns of diversity in Europeans and Africans are sufficiently different as to be distinguishable locally as we go along a chromosome [Price et al., 2009]. The resolution of chunk identification has improved with the availability of genotyping arrays with ever greater density of single nucleotide polymorphism (SNP) data (and more recently, sequence data) and the development of sophisticated statistical methods, as reviewed below.

2.3 Local Ancestry Inference

In the case of admixture between two populations, ancestry blocks were first inferred using SNPs which had very different allele frequencies in the two populations, and were in approximate linkage equilibrium with each other in the ancestral populations. The SNPs are called *ancestry informative markers* (AIMs) and a few thousand such SNPs would be used in a typical study, or approximately a few per megabase.

Hidden Markov Models (HMMs) have been used to model ancestry segments and their transitions [Falush et al., 2003], and are motivated by the process illustrated in Figure 2.1. Consider an admixture event, say, T generations ago, between individuals of fully African or European ancestry followed by multiple generations of panmictic mating within the admixed population. In the first generation, individuals inherit one African and one European chromosome. In the next generation, recombination will shuffle the chromosome, and chromosomes chunks will switch to another chunk at crossover break-points. This will occur at the rate of 1 per Morgan (by definition). Under the assumption of no interference, the placement of crossover events can be modelled as a Poisson process along the chromosome with a rate of 1 event per Morgan. Each subsequent generation will inherit blocks of DNA from chromosomes in the previous generation in a similar manner. It follows from the superposition of in-

dependent Poisson processes that switch points, between stretches of DNA inherited as a chunk, going along a chromosome in the present generation, will form a Poisson process of rate T per Morgan. Given a switch point, the chunk being switched to could be either African or European – it is independently drawn from the two populations with probabilities proportional to the relative fractions of the two ancestries in the population. Therefore, not all crossovers will result in change of ancestry, since chromosome chunks may re-join a chunk of the same ancestry after the crossover. The ancestry along a chromosome thereby forms a Markov chain. The genotype data of an admixed individual, whose ancestry segments are unknown, can be modelled using a Hidden Markov Model, with the ancestry of each locus as a hidden state. I will refer to this model later as the *admixture LD model*.

This model has been used in several approaches, for instance, structure [Falush et al., 2003], and ANCESTRYMAP [Patterson et al., 2004]. Variations of this approach have been designed to utilize genotyping chips designed for genome-wide association studies (GWAS) containing hundreds of thousands of markers. These approaches, for instance, LAMP [Sankararaman et al., 2008], apply previous HMM-based methods by thinning the SNPs in GWAS chips to sets of maximally informative unlinked markers. Another class of models [Bryc et al., 2010] adapts principal components analysis (PCA) methods used on genome-wide data in previous genetic studies for local inference using sections of chromosomes.

None of these models explicitly models linkage disequilibrium. Accounting for LD in a likelihood based framework offers the potential of better inference for two reasons. First, allowing linked markers enables information from a larger set of SNPs to be leveraged. Second, the joint probability distribution of SNPs in haplotypes evolves differently, since joint allele frequencies are influenced by different demographic processes, such as recombination, which achieve equilibrium slowly over a longer timescale (Section 1.3.3). Haplotypes, therefore, are potentially more differentiated between

populations than SNPs, and can lead to better ancestry inference. Algorithms accounting for LD, such as SABER [Tang et al., 2006], HAPAA [Sundquist et al., 2008], HAPMIX [Price et al., 2009], and LAMP-LD [Baran et al., 2012] (to name only a few), have led to large improvements in local ancestry inference. In this work, we have utilized HAPMIX, which has been shown in simulation to be best in class for local ancestry inference in African Americans [Churchhouse and Marchini, 2013] and has attractive statistical properties that are important for our application.

2.3.1 HAPMIX

HAPMIX is a likelihood-based approach that models haplotypes observed in a sample using haplotypes observed in ancestral 'reference' populations [Price et al., 2009]. I will only summarize the algorithm here, since a large number of parameters and equations are required to describe it completely.

As introduced in Section 1.3.3, calculating the full likelihood of observed genotypes in a sample is computationally intractable for all but the smallest datasets, and several approximations have been developed. A particularly popular approximation called the *product of approximate conditionals* (PAC) was introduced by Li and Stephens [2003] and is computationally quite fast. The joint likelihood of a set of sample chromosomes, say $s = \{s_1, s_2, \ldots s_n\}$ can equivalently be written as a product of conditional likelihoods $\Pr(s_1)\Pr(s_2|s_1)\ldots\Pr(s_n|s_1 s_2 \ldots s_{n-1})$. The PAC likelihood approximates the conditional probability $\Pr(s_k|s_1 s_2 \ldots s_{k-1})$ by modelling the k^{th} chromosome as an imperfect mosaic of the previous $k-1$ chromosomes. It essentially implements a 'copying' model, wherein s_k copies a segment or haplotype of one of $\{s_1, s_2, \ldots, s_{k-1}\}$, followed by a 'recombination event' at some rate, followed by copying another haplotype chosen uniformly at random from one of $\{s_1, s_2, \ldots, s_{k-1}\}$, followed by a recombination event and so on. The copying can be imperfect, reflecting mutation events. Since the 'correct construction' is obviously not observed, comput-

ing the probability of $\Pr(s_k|s_1 s_2 \dots s_{k-1})$, requires summing over the probabilities of all possible constructions. The Markov assumption used in constructing the chromosome enables the use of a hidden Markov model to compute the probability efficiently.

Intuitively, the haplotype copied by the sample chromosome at any point can be thought of as having recent shared ancestry with it. Though this model is motivated by ideas related to the unobserved underlying genealogy, it is flawed in that there is no actual genealogical interpretation for the construction. In reality, the constraints placed by the genealogy underlying a set of samples mean that the process along a chromosome is not Markov [McVean and Cardin, 2005]. Nevertheless, the method captures a lot of the information in the data, and is especially useful in that it provides a powerful and flexible way of representing and analyzing haplotypes.

The copying switch points or 'recombination' events, inferred by this scheme, have been used to estimate recombination rates [Li and Stephens, 2003]. Chromosomes copy haplotypes from other chromosomes, and the switch points approximate ancestral recombination events that have changed the structure of LD in the region. In our work, we wish to learn about recent recombination events, and therefore we do not base our inference on these switch points.

HAPMIX essentially merges the PAC copying model with the 'admixture LD model' outlined above. It models each admixed individual as a mosaic of ancestry blocks (as discussed in the previous section and illustrated in Figure 2.1), with the haplotypes *within* each block constructed by copying haplotypes from the respective ancestral populations. Intuitively, going along an African American chromosome, the chromosome will have African chunks and European chunks, which will switch at some rate T per Morgan, where T is related to the number of generations since admixture. *Within* an African chunk, the chromosome is more likely to copy from an African chromosome, and will switch to copy from another African chromosome, at some rate that depends on the extent of LD in Africans. Similarly, within a European chunk,

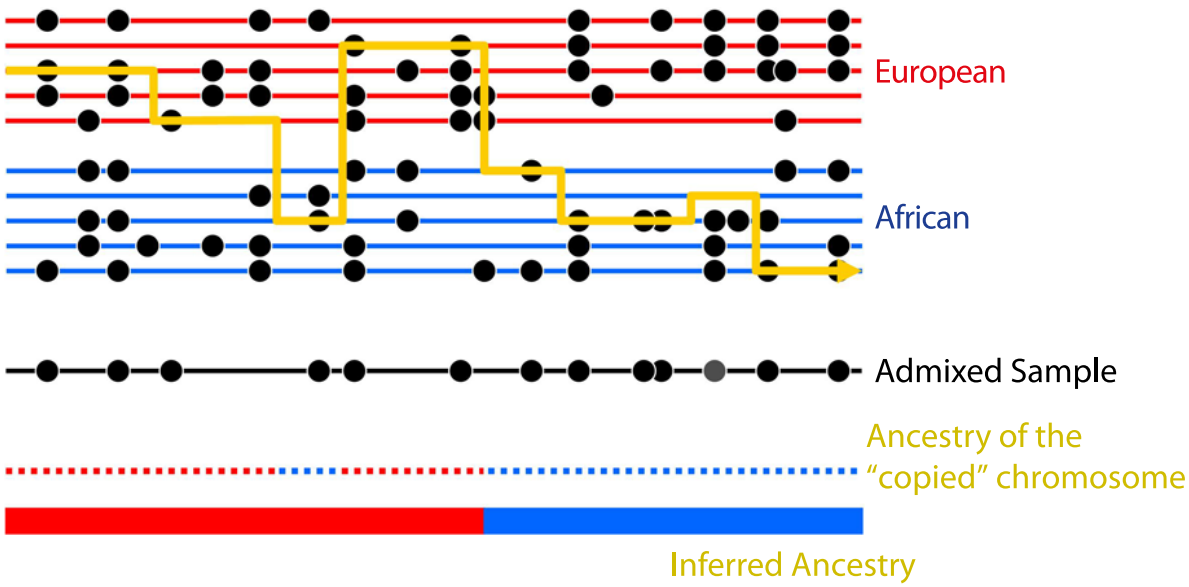


Figure 2.3: HAPMIX copying model for an admixed chromosome (Figure adapted from Price et al. [2009]). Densely typed phased chromosomes are obtained from the reference ancestral populations (red and blue). The admixed individual's chromosome (black) is constructed by copying segments from different chromosomes from these reference panels, and the yellow line shows the path traced to construct this mosaic.

Consider a segment in the admixed individual's chromosome that has (unobserved) African ancestry. Since Africans and Europeans are reasonably differentiated (at least at the scale of human variation), the segment will usually, but not always, copy from an African chromosome. Similarly, within a European chunk, the segment will usually copy from a European chromosome. Therefore, the underlying ancestry in the sample can be inferred by summing over the ancestries of the chromosomes copied at any location.

Switches between chromosomes within a population are frequent and depend on the extent of LD (among other factors), while switches between populations represent crossovers since admixture, which are recent and therefore relatively rare. The admixed sample may occasionally contain a modified version of the copied haplotype (grey circle), due to genotyping error or mutation.

HAPMIX allows a 'miscopying parameter' which allows a certain degree of copying without switching from a population. This provides robustness against spurious ancestry switches in the situation that the true ancestral haplotype is not represented in the reference population due to inadequate sampling, insufficiently well matched reference populations and/or deep coalescence time for a particular haplotype. Therefore, the dotted line below the sample, which shows the reference population being copied from at each point in the chromosome, may not exactly match the thick solid line below it, which shows the inferred ancestry.

the chromosome is more likely to copy from a European chromosome, and may switch to another European chromosome, at some rate that, in turn, depends on the extent of LD in Europeans. The ancestry of loci in the sample chromosome can therefore be learned by looking at the ancestry of the chromosomes likely to be copied at those loci. (Figure 2.3).

The scheme is illustrated and described in Figure 2.3. HAPMIX implements a nested HMM, which includes a 'broad-scale' HMM involving transitions between ancestry states, plus a 'fine-scale' HMM involving transitions between haplotypes within an ancestral population. Effectively, HAPMIX estimates, for each locus in the genome, the likelihood that the haplotype observed in an admixed individual is a better match to one ancestral population or the other. For a diploid African-American sample, the hidden state in the broad-scale HMM has 3 possibilities at every locus in the genome: 0, 1 or 2 alleles of European ancestry, and HAPMIX infers the probability of each of these states.

HAPMIX requires variation data for the reference populations to be provided as input. It also requires as input, an estimate of the number of generations since admixture, and a genetic map containing the genetic distance between every adjacent pair of SNPs included in the variation data. Since, in our setting, we will be using HAPMIX to build a genetic map, we run HAPMIX setting the input map to have a uniform rate.

A crucial property of HAPMIX is that its ancestry inferences are well-calibrated (Figure 2.4). This means that, in a set of simulated African American chromosomes, the probability of x copies of European chromosomes (where $x \in \{0, 1, 2\}$) predicted by HAPMIX closely matches the actual frequency of having x copies of European chromosomes. This implies that the inferred probabilities of x copies of European chromosomes can be used to calculate *the expected number of European alleles* at each locus in the genome. We use this information to build a genetic map as discussed

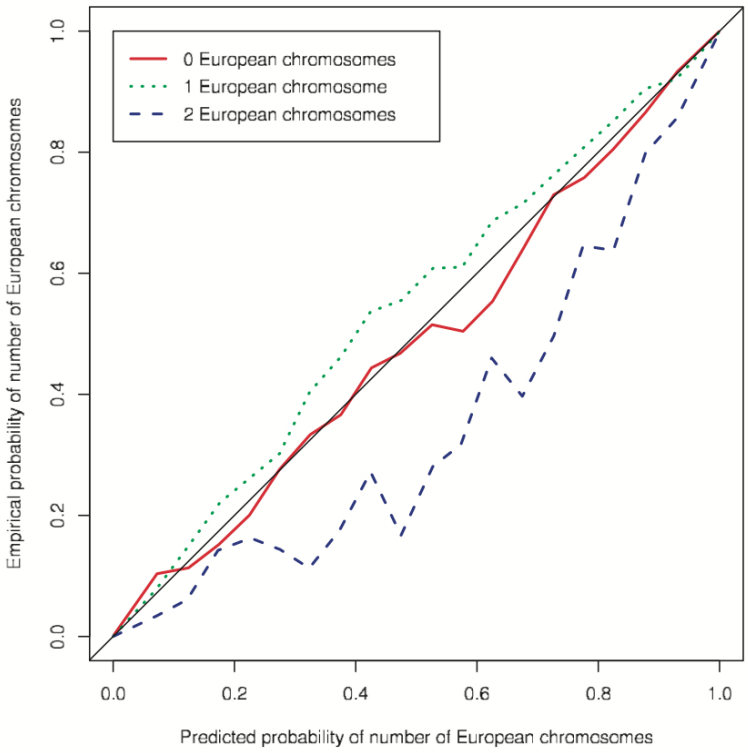


Figure 2.4: Calibration of ancestry inference produced by HAPMIX (Figure reproduced from Price et al. [2009]). For different numbers of European chromosomes, $x \in \{0, 1, 2\}$, Price et al. [2009] compared the probability of x copies of European chromosomes predicted by HAPMIX to the actual frequency of having x European copies in a set of simulated African-American chromosomes with 6 generations of admixture. The results would lie along the $y = x$ (diagonal black line) for a perfectly calibrated method.

below.

2.4 Building an African-American genetic map

In this section, I describe how we detect crossovers, filter them, and use them to build an initial genetic map. This map will be refined in the section using a model-based approach.

2.4.1 Detection of crossovers

I identify crossovers by inferring ancestry segments in the genome of every individual, and the transitions between them. As discussed earlier, transitions between ancestry segments are due to recombination in an ancestor of the present-day individual. To identify these transitions, we ran HAPMIX on each individual's genotypes. Inferring ancestry requires construction of reference panels from two populations, each of which approximates one of the two populations ancestral to the present-day samples. As discussed in Section 2.2, individuals from the Yoruba population (YRI) in Ibadan, Nigeria and the Centre d'Etude du Polymorphisme Humain collection (CEU) in Utah, USA, have been shown to form good reference panels for the African and European ancestries respectively in present-day African Americans [Price et al., 2009].

Phased chromosomes were obtained from the HapMap 3 dataset [International HapMap 3 Consortium, 2010] to construct ancestry reference panels[2]. The African reference panel comprised 226 and the European panel comprised 224 phased chromosomes.

Samples in HapMap 3 had been typed on Affymetrix 6.0 and Illumina 1M arrays. Most of the samples used in these study were typed on one of these two arrays, and some were typed on one of two older Illumina arrays (Table 2.1). Almost all of the SNPs in these arrays are subsets of the approximately 1.4 million SNPs that passed quality checks in HapMap 3. We intersected the SNP set of each consortium in our data with the HapMap 3 SNPs. Next, we filtered out SNPs that had observed allele frequencies that strongly differed from the expected $80 - 20\%$ (Figure 2.2) linear combination of YRI and CEU allele frequencies (t-statistic with an absolute value of greater than 3). These SNPs are likely to contain either genotyping errors in the reference panels or the samples, or have been incorrectly labelled (i.e., the SNPs may

[2]This information relates to map building for the autosomes and the non-autosomal portion of the X chromosome. The Pseudo Autosomal Region 1 (PAR1) had to be handled separately.

have been flipped).

We ran HAPMIX on each (unphased) sample in diploid mode, and with a prior hypothesis of 20% European and 80% African ancestry and 6 generations of admixture per individual, and the choices are justified on the basis of previous genetic studies (Section 2.2). HAPMIX requires users to input a recombination map as a prior distribution. To prevent our genetic map from being biased towards previous genetic maps, we provided HAPMIX with a piece-wise constant prior rate. Rates were assumed to be constant across each chromosome arm, with a total rate across each arm determined by the Rutgers genetic map [Matise et al., 2007]. Very low prior rates were used for the centromeric regions to reflect the known rarity of centromeric crossovers [Mahtani and Willard, 1998]. Table 2.2 details the Rutgers map rates used to derive prior rates for the autosomes and the non-autosomal portion of the X chromosome. X chromosome rates were scaled to reflect female-only recombination prior to running HAPMIX.

As described in Section 2.3.1, HAPMIX output can be used to calculate the expected number of alleles of each ancestry at each locus of the genome. The HAPMIX ancestry 'painting' of a chromosome of one of the individuals in our study is illustrated in Figure 2.5. For most of the chromosome, close to an integer number of European alleles is inferred (0, 1, or 2), reflecting near certainty in ancestry inference. Some loci, however, lie in regions where the ancestry of one of the two chromosomes in the individual transitions from European to African or vice-versa. We can infer that crossover has occurred in each such region in this individual's lineage since admixture.

I note that not all crossovers that have occurred in ancestors of the present day individual, since admixture, will be observed. Only crossovers that take place at a locus with heterozygous ancestry are visible (i.e., switches that actually result in a change of ancestry). Therefore, given the relative proportions of African and European ancestries inferred in our samples, we expect to find only about 32% of the crossovers.

Genetic map from unrelated African-American samples

Chromosome	p-arm end		q-arm start		q-arm end	
	Physical (Build 36)	Genetic (cM)	Physical (Build 36)	Genetic (cM)	Physical (Build 36)	Genetic (cM)
1	120,330,455	150.83	144,243,274	151.41	266,127,003	308.19
2	88,457,695	111.79	94,912,172	111.90	262,575,050	284.34
3	90,335,510	110.04	95,301,156	110.16	219,287,577	246.35
4	47,052,440	67.38	52,814,094	68.47	211,158,857	235.77
5	45,603,505	67.28	49,659,296	67.50	200,420,866	231.25
6	58,440,820	84.17	62,118,323	84.37	190,747,902	214.57
7	57,884,453	77.26	62,171,909	77.55	178,710,965	210.26
8	42,668,804	64.01	47,154,379	64.38	166,114,869	191.11
9	45,617,414	66.52	66,481,119	66.62	160,137,385	195.02
10	38,715,905	63.33	42,274,907	63.47	155,053,665	198.19
11	51,303,111	69.25	54,945,985	69.30	154,402,514	184.59
12	34,142,287	58.67	36,887,280	58.81	152,287,718	200.90
13	17,394,434	0.00	18,394,434	1.38	134,031,187	160.81
14	18,515,210	0.00	19,515,210	1.44	126,357,542	155.54
15	19,367,093	0.00	20,367,093	1.66	120,181,650	167.32
16	34,715,504	59.66	45,272,030	59.77	108,640,449	168.01
17	21,864,298	52.69	22,336,210	52.92	98,609,338 1	67.48
18	15,072,789	47.54	16,909,449	47.67	95,965,205 1	57.88
19	24,163,969	50.26	32,842,516	50.46	83,785,296 1	53.90
20	26,114,909	53.48	29,310,063	53.69	82,381,835 1	51.81
21	12,678,219	0.00	13,678,219	2.11	66,902,240 1	14.41
22	14,440,166	0.00	15,440,166	2.32	69,508,925 1	27.92
X	57,820,632	74.60	63,670,690	74.75	174,803,587	164.77

Table 2.2: Prior rates for map building were derived from this table, which is based on the Rutgers genetic map [Matise et al., 2007]. The recombination rate prior used for HAPMIX was piecewise uniform over three segments: (i) the p-arm start between position 0 of both the physical and genetic maps and the p-arm end, (ii) the centromere which is between the p-arm end and q-arm start, and (ii) the q-arm between the q-arm start and the q-arm end.

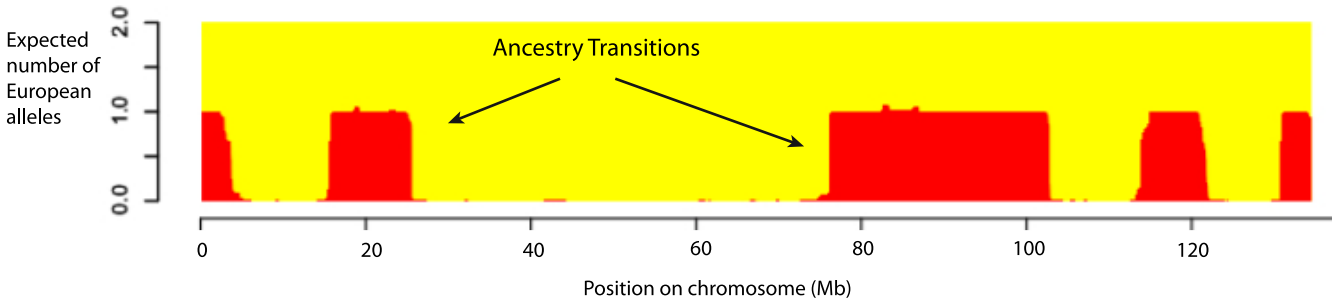


Figure 2.5: Sample chromosome with local ancestry inferred. European alleles are coloured red and African alleles yellow. Eight crossovers are observed in this example.

Further, this number is only a fraction of the potentially visible switches that have occurred in the ancestors, since most ancestral genetic material is not transmitted to the present-day individual (a present-day chromosome has 2^n chromosomal ancestors n generations ago). This number is further reduced by filtering to get only those ancestry switches which are likely to contain exactly one crossover, as described next.

2.4.2 Filtering of ancestry transitions to high confidence crossover locations

Although HAPMIX is expected to be $> 99\%$ accurate in making ancestry calls [Price et al., 2009], ancestry in some regions will be incorrectly inferred. Even though these regions are likely to be small, this issue is problematic for map building as every incorrect segment will lead to two false-positive crossovers being inferred. Therefore, filtering is important, and I have designed filters to reduce the impact of erroneous or uncertain inference of ancestry blocks.

There are three particular classes of false-positive transitions that the filters are designed to remove. They are illustrated in Figure 2.6, and discussed in turn below:

- *Confidence filter: restricting to transitions that HAPMIX is most confident about.* I observed that many transitions do not end in states where HAPMIX is confident about their ancestry (Figure 2.6). For example, an individual may

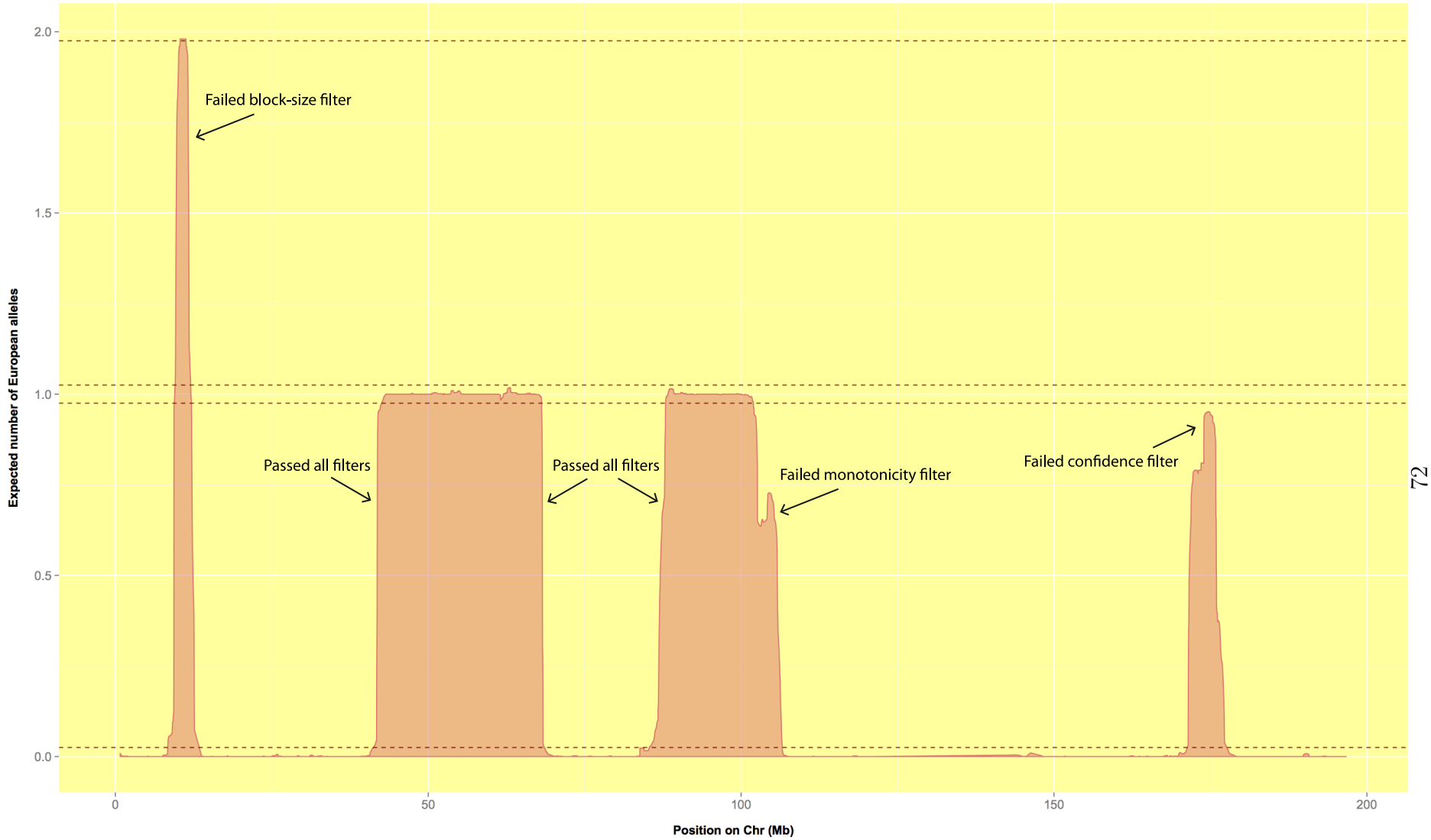


Figure 2.6: Chromosome painting constructed to illustrate filtering on putative ancestry-switch crossover events (does not show a real individual). The dashed vertical lines indicate the thresholds for calling an event, i.e. transition between 0.025 and 0.975, or between 1.025 and 1.975.

have, for a sequence of SNPs, a corresponding inferred ancestry sequence of $\{0, \ldots, 0, 0.1, 0.2, 0.9, 0.3, 0.1, 0, \ldots, 0\}$ European alleles, implying that there is a SNP with a 90% probability of European ancestry on one of the chromosomes in the middle of a block of African ancestry. This could happen, for example, when two crossovers have occurred in close proximity, thereby partially erasing the signal. This could also happen if HAPMIX has made a mistake, due to errors in the data or in the modelling assumptions. Either way, we cannot be confident that a unique event has occurred in that interval. I include only those transitions for which HAPMIX is at least 95% confident that they occurred. This is implemented by requiring that there has been a transition from less than 0.025 expected European alleles to more than 0.975 expected European alleles (or vice versa) or from 1.025 expected European alleles to 1.975 (or vice versa), as illustrated in Figure 2.6.

- *Monotonicity criterion: restricting to transitions with no evidence of multiple overlapping events.* Some transitions do not display a unidirectional gradient of ancestry change (Figure 2.6), in that a portion of the ancestry change suggests a transition in the opposite direction. This could happen, for example, when two crossovers have occurred in close proximity, and the signals have become merged. It is not possible, in this situation, to identify the correct end points of the transition(s). I therefore filter out events with non-monotonicity greater than 1%, where non-monotonicity is defined as the total change in ancestry in intervals whose direction is opposite to the rest of the transition.

- *Block size criterion: Restricting to transitions flanked by large blocks of contiguous African or European ancestry.* The distribution of sizes of ancestry can, under reasonable simplifying assumptions, be modelled as an exponential distribution. The reason is as follows: consider an admixture event between

individuals of fully African or European ancestry followed by multiple generations of random mating within the admixed population. In the first generation, individuals inherit one African and one European chromosome. In the next generation, recombination will shuffle the chromosome at the rate of 1 per Morgan. As discussed before, for the 'admixture LD model' (Section 2.3), placement of crossover events can be modelled as a Poisson process along the chromosome with a rate of 1 event per Morgan (though only some of the crossover events will be visible as a change in ancestry, which depends on the relative fractions of the two ancestries in generation 0). Each subsequent generation will undergo a similar shuffling. Since the superposition of independent Poisson processes is also Poisson, it follows that block switches will follow a Poisson process in the present day chromosome. The block sizes, therefore, will have an exponential distribution.

Across the African-Americans in our dataset, the exponential distribution is indeed an excellent fit for the most part, suggesting that the data are consistent with the assumption of a single admixture time. The exception is that too many very short ancestry blocks are inferred. On these grounds, I filtered out putative crossovers with flanking ancestry chunks shorter than 2 cM (using the HapMap population-averaged LD-based map [Myers et al., 2005] to assess the block size in Morgans). Figure 2.7 shows the distribution of European and African ancestry blocks flanking ancestry switches after filtering. Correctly identifying very short ancestry blocks is, in fact, a common problem for algorithms inferring local ancestry, and short ancestry blocks appear to be inferred due to a variety of reasons, for instance, insufficient diversity in the reference panel to capture the patterns in the sample or deep coalescence times of some haplotypes [Sundquist et al., 2008; Price et al., 2009].

Table 2.3 summarizes the number of crossovers inferred before and after filtering.

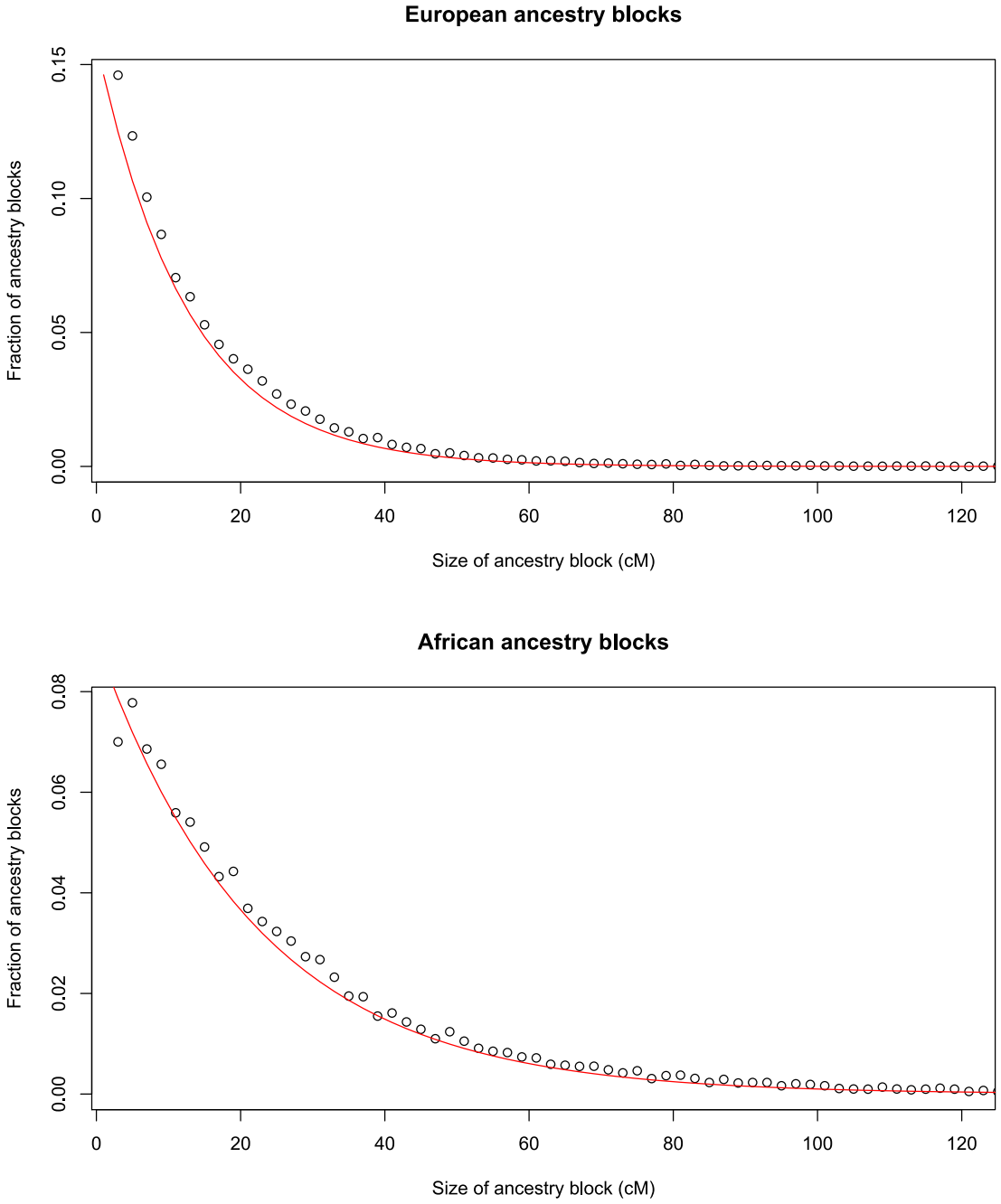


Figure 2.7: The distribution of African and European ancestry block sizes after filtering (black) together with their best-fit exponential distributions (red). African blocks are larger than European ancestry blocks since a greater fraction of crossovers within an African block lead to no change in ancestry.

Number of events called using the confidence criterion in 29,589 individuals	2,805,677
Events filtered by the monotonicity filter	7.9% or 221,208 events
Events filtered by the block-size filter	16.8% or 471,176 events
Number of events remaining	2,113,293

Table 2.3: Summary of the number of events before and after performing filtering steps. The final number of events is approximately 71 per person.

The genetic maps below were built using the highly confident 2.1 million events that remained after filtering. Figure 2.8 illustrates how we identify and validate crossovers inferred using this procedure in one nuclear family from the CARe consortium. I carried out a similar check for 15 additional nuclear families whose genotypes were available to us: there were no events that passed our filters, yet failed validation using families.

It is important, of course, that the filters we employ do not create biases in our rate estimates. The first two filters, the confidence criterion and the monotonicity filter, are naturally dependent on the ability to resolve events cleanly as poorly resolved crossovers that extend over a large chromosomal region are more likely to overlap with other crossovers. It is possible that such filters could introduce bias if they remove a larger fraction of true events in difficult to analyze regions. The final filter, the block-size filter, is not expected to create bias in rate estimates, in principle, since flanking block sizes, measured in Morgans or equivalent, have a distribution that is independent of rate at the break site (this is because block sizes are being measured in terms of *genetic distance*, not physical distance, and therefore should not depend on the rate itself). Nevertheless, there is the possibility of bias in regions

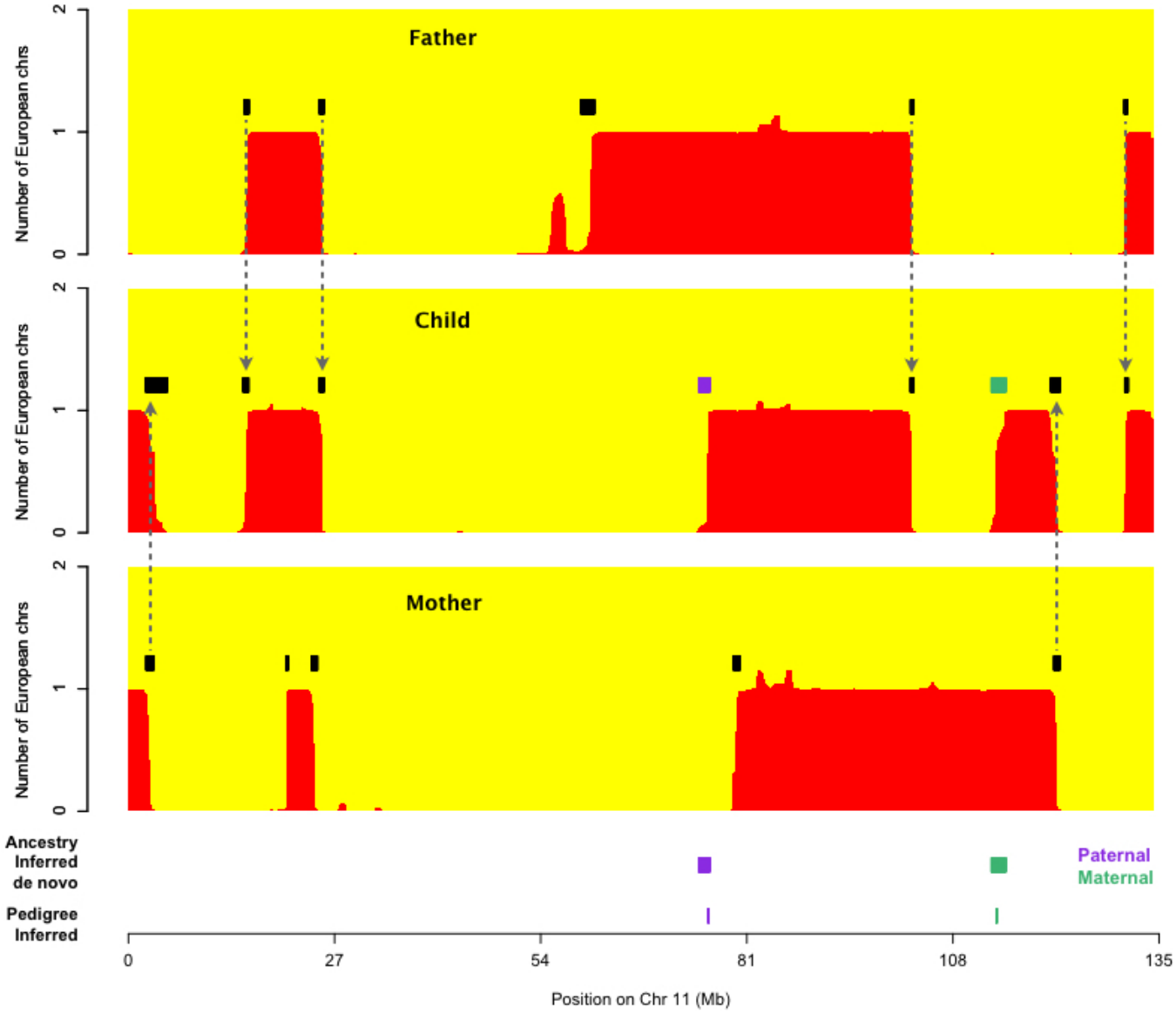


Figure 2.8: Detection of crossovers between blocks of inferred ancestry is illustrated in a father-mother-child trio. Black rectangles show inferred crossovers in each individual, with arrows showing the likely transmission of ancestral crossovers from parent to child (inferred by overlap). The purple and green segments show two *de novo* events. The purple event is inferred to be of paternal origin, since only the father is heterozygous in ancestry at this locus. Similarly, the green event is inferred to be maternal in origin. These events correspond precisely to crossovers identified directly using genotype data for an additional child in the same pedigree, as described in Chapter 3 (bottom of plot, marked Pedigree Inferred). Agreement with the independent pedigree-based approach confirms the validity of inferring a crossover in these segments.

where broad-scale LD-based recombination rates have large errors or deviate from present-day rates. We have no reason to believe that such occurrences are common based on strong genome-wide correlations between maps, particularly at broad scales (Chapter 4).

2.4.3 Probability distribution of a crossover and building a 'basic' genetic map

In this section, I use the expected number of European alleles estimated by HAPMIX at each SNP to construct a probability mass function for the location of a crossover inside an ancestry transition, under the assumption that a single ancestry-changing crossover has occurred inside that transition.

Consider an ancestry transition from 0 to 1 expected European alleles, without loss of generality. Let the transition take place between SNPs labelled $\{1, 2, .., \ldots n\}$. We define $X_j \in \{0, 1\}$ as the number of European alleles at SNP j. Further, crossover location $Z = j$ if the crossover occurred in the interval between SNPs j and $j + 1$. Let $E_j = \mathrm{E}[X_j]$, so that $E_1 = 0$ and $E_n = 1$. Therefore, assuming that exactly one crossover has occurred:

$$\begin{aligned} E_j &= P(X_j = 1) \times 1 + P(X_j = 0) \times 0 \\ &= P(X_j = 1 | Z = j - 1) P(Z = j - 1) + P(X_j = 1, Z \neq j - 1) \\ &= P(Z = j - 1) + P(X_{j-1} = 1) \\ &= \sum_{i=1}^{j-1} P(Z = i) \end{aligned}$$

Therefore, if follows that

$$P(Z = j) = E_{j+1} - E_j$$

If we assume that the underlying rates are uniform, we obtain E_j for $j \in \{1, 2, .., \ldots n\}$

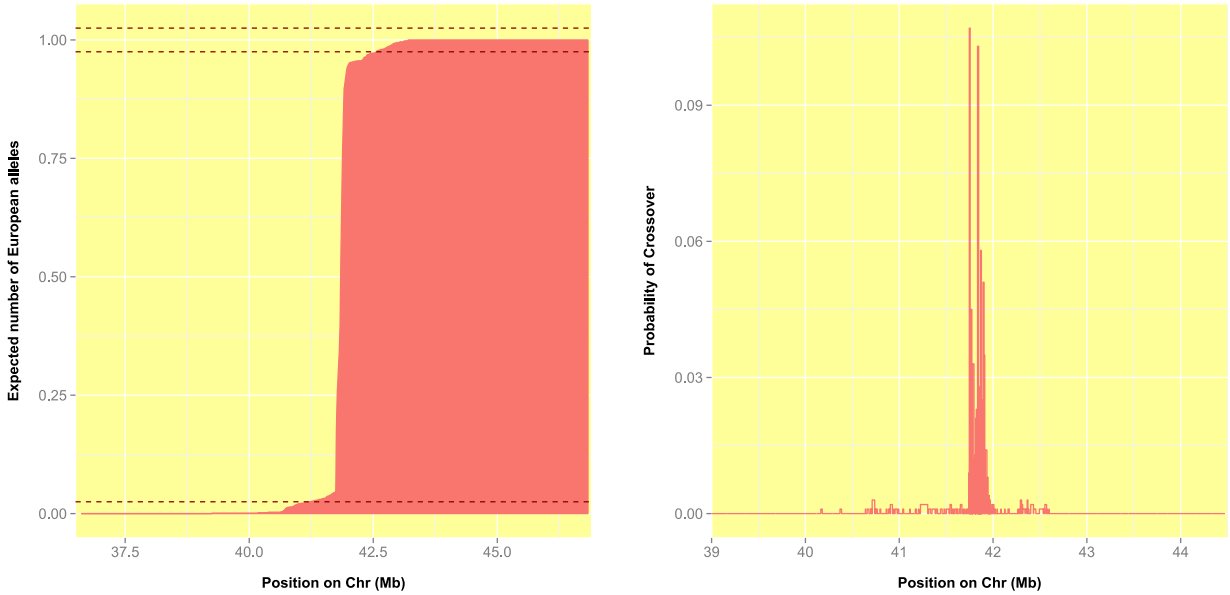


Figure 2.9: HAPMIX output can be used to calculate the distribution of the crossover's location in an ancestry switch, assuming that a single crossover has occurred in the region (details in Section 2.4.3).

from HAPMIX, as discussed earlier in this chapter. This gives us the distribution of crossover location in an ancestry transition, in principle. In practice, to deal with noise in the data, the following steps are taken:

1. Using the monotonicity filtering criterion, I filter out events with non-monotonicity greater than 1% (details in Section 2.4.2). For any events with non-monotonicity below that threshold: if $E_n > E_1$ then for any interval j where $E_{j+1} < E_j$, $P(Z = j) \equiv 0$, and vice versa.

2. To be consistent with the confidence criterion (Section 2.4.2), events are assumed to occur in restricted region between SNPs with $0.025 - 0.975$ and $1.025 - 1.975$ expected number of European chromosomes (as opposed to $0 - 1$ and $1 - 2$ respectively). The probabilities of crossover $P(Z = j)$, in intervals j in the restricted region, are re-normalized so that the probabilities over the region sum to one.

The calculations above produce a probability distribution for a crossover inside

an ancestry switch assuming uniform recombination rates. In Section 2.5, I describe how we calculate the distribution under variable rates.

To build a 'basic' genetic map I simply add the probability of a crossover from all 2.1 million events for each interval. Since the individuals in our dataset are genotyped on several different arrays (Table 2.1), I create a union of all SNP sets. This set contains a total of 1,271,992 SNPs. I transform the probability distribution of each event from its own SNP set to the 'union' SNP set by the natural linear interpolation, and re-calculate the PMF. I then added up the PMFs of all events to obtain a map *proportional to* the probability of crossover per meiosis at each interval in the genome. This gives us, on average, 1 crossover per 1.4 kb of the genome.

The total number of generations of recombination represented in the ancestry of each individual since European and African admixture began is not known. I also expect to miss many genuine crossover events as a result of the filtering procedure. As a result, this is a *relative* genetic map rather than an absolute map. I therefore normalized the total length of the map to make it equal to that of the HapMap population-averaged LD-based map [Myers et al., 2005][3]. Finally, the crossovers detected in an individual today may have occurred in maternal or paternal ancestors. Therefore, this map is informative about crossover positions in males and females jointly, but not separately.

2.4.4 Crossover Resolution

In this section, I assess how well-resolved ancestry-switches are. Transitions often appear to be quite long, for instance in Figure 2.9, where it is ∼1.2 Mb, and the median distance between the end-points of transitions defined by the confidence criterion is 1 Mb. However, the probability mass function (PMF) itself is sharply peaked and highly "informative" for the event in Figure 2.9, with most of the probability of crossover

[3]HapMap release 24 (NCBI Build 36) map was used for normalization. It was, in turn, normalized using the map built by Kong et al. [2002].

concentrated in a few inter-SNP intervals. To examine if this is generally the case, I looked at the probability of crossover across all inter-SNP intervals across all events. Each interval in each event was then associated with a crossover probability density – the probability of crossover per base of the interval. The greater the density, the more informative is the corresponding interval. This allows us to sort all intervals on the basis of how informative they are. By counting the most informative intervals first, and adding up how much of the total crossover probability they account for, I find that 50% of crossover probability fell within just over 70 kb per crossover (Figure 2.10). Informally, this means that, across the genome, the 'best' half of information about recombination is contained within 70 kb per event. While not directly comparable, I note that this is similar to the resolution of events in contemporary pedigree-based studies, for example Coop et al. [2008] and African Americans (Chapter 3).

2.5 Improved map resolution using Bayesian modelling

In this section, I build a map using information about crossovers in all individuals jointly, to achieve better resolution. There are three important facts which are not taken into account in the construction of the basic genetic map and which can be incorporated into a Bayesian model-based framework:

- The concentration of recombination in hotspots identified in pedigree-based and LD-based maps shows that recombination rates are shared, at least partially, with other individuals in a population (Section 1.4.4). Therefore, it may improve resolution if we model the observed crossovers as arising from a shared genetic map, or a set of genetic maps. Later chapters, particularly Chapter 5, shed light on the validity of this assumption.

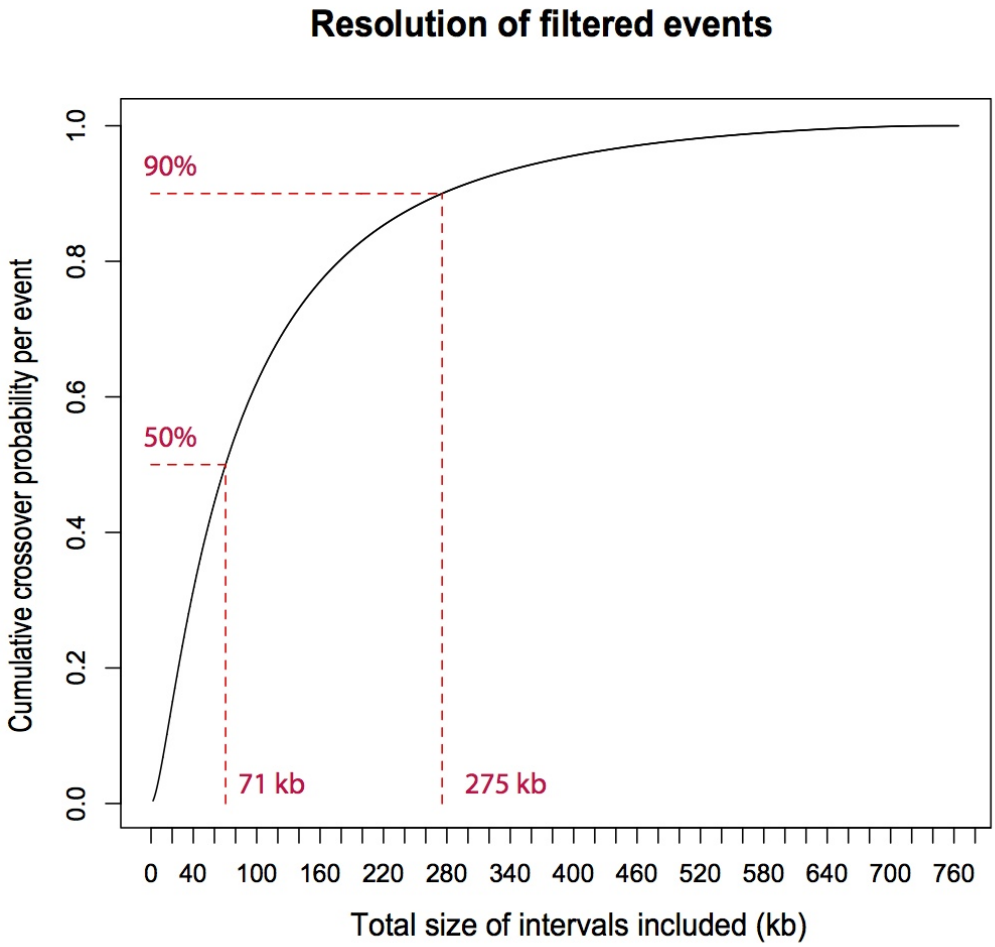


Figure 2.10: The regions most informative about crossover (details in text) contain half the probability of recombination genome-wide across 6,209 CARe individuals in just over 70 kb per crossover.

- High-resolution mapping of crossovers by sperm-typing has shown that crossovers are highly concentrated in hotspots, which are 1-2 kb wide, interspersed with regions which are much less likely to experience recombination (Section 1.4.4). Priors that reflect this distribution can, in principle, improve inference [Auton, 2007].

- Each ancestry switch is assumed to be highly likely to harbour a single crossover (the filters were designed to achieve this aim), which occurred in a unique inter-SNP interval. This information is helpful in dealing with the statistical noise associated with inferring ancestry switches.

Ancestry switches are observed quantities, while rates in the genetic map and the actual location of the crossover underlying each ancestry switch, are unobserved. A simple modelling choice for this problem is to model each inter-SNP interval (henceforth referred to simply as interval) as having a rate independent of all other intervals. Crossover events in an interval are assumed to occur independently of all other crossovers, with probability proportional to the rate in that interval. This model is illustrated in Figure 2.11. A crossover may lead to an ancestry switch (not all crossovers will be observed as ancestry switches; for instance, those resulting in no change of ancestry on either side of the crossover).

The assumption of independent rates implicitly assumes that hotspots are small enough to be contained within an interval. This may not be unreasonable, given that the average interval size is close to 3 kb, greater than the size of a typical hotspot (1-2 kb).

Human recombination rates, however, are known to vary systematically at broad scales (Section 6.1.1), and this introduces long-range correlations between rates. Figure 2.12 shows the variance of rates (in Morgans) in the HapMap2 LD-based genetic map and confirms that rates are correlated over hundreds of kilobases. There are several ways I could deal with this, for instance, by using a model that allows rates to be

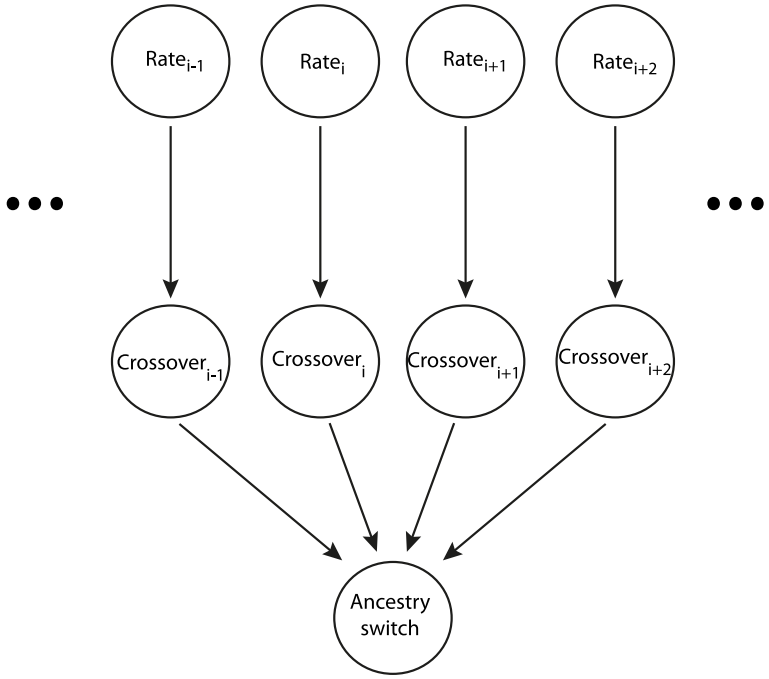


Figure 2.11: Our model assumes that rates in different intervals are independent of each other and I model crossover locations within an individual as independent conditional on these rates. A crossover may lead to an ancestry switch being observed.

estimated in intervals learned from the data as opposed to fixed inter-SNP intervals (e.g., reversible jump Markov Chain Monte Carlo [McVean et al., 2004]) or explicit modelling of the hierarchical nature of recombination-rate control (e.g., nested Hidden Markov Models). Both these options would result in very complex models with a large number of parameters, and it is not clear that there is sufficient resolution information in the data to learn these complex models. I also thought it unlikely that such models would be computationally tractable in our setting, and careful model choice would be needed to prevent over-smoothing of rates, which would result in loss of resolution.

Instead, I choose a prior on rates such that it reflects these long-range autocorrelations, at least partially. Analysis of ancestry switch resolution in Section 2.4.4, suggests that the basic genetic map, while it does not have resolution at the hotspot scale, should nevertheless be correct at the scale of hundreds of kilobases (where there

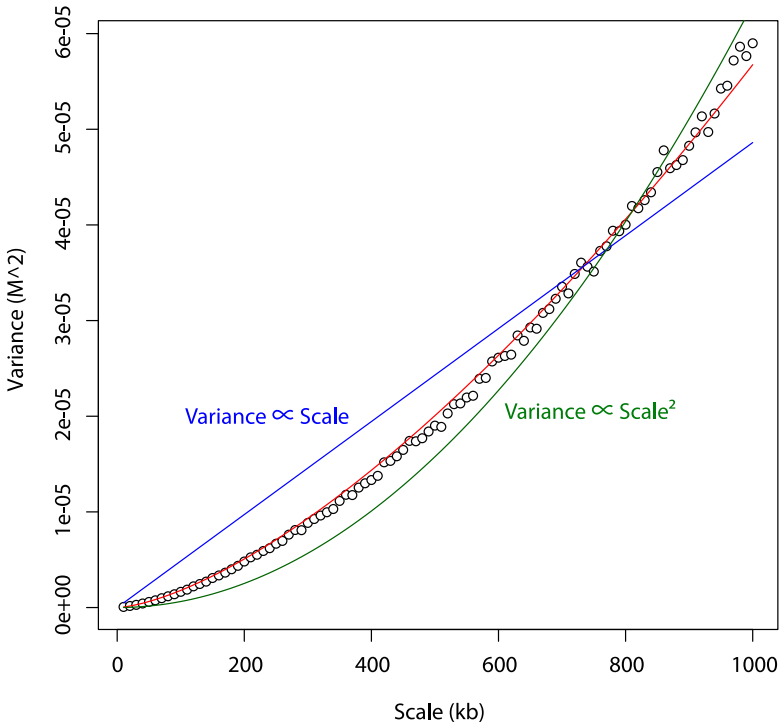


Figure 2.12: Variance of genetic distances at different scales in chromosome 1 of the HapMap2 LD-based genetic map [Myers et al., 2005].

If rates in different intervals were truly independent, as assumed in our model, the overall variance of the estimated genetic distance in a region would be the sum of the variances of genetic distances in sub-intervals of the region. Therefore, we would expect the variance to increase linearly with the size of that region (blue line).

If rates were constant throughout the chromosome, the genetic distance of a given region would simply be proportional to the size of the region. Since this is equivalent to multiplying the underlying random variable (the constant rate) by a number proportional to the size of the region, we would expect the relationship to be quadratic (green line).

The best fit curve with zero intercept and a single polynomial term (red) is between the two curves and varies as $\approx \text{scale}^{1.5}$.

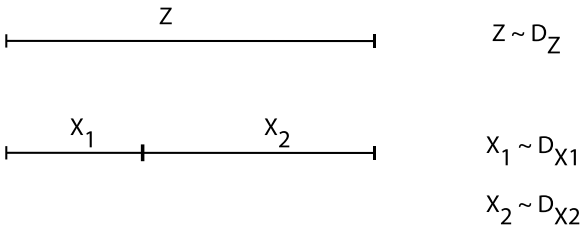


Figure 2.13: Distribution of rates should not depend on SNP set, i.e, D_Z should be equal to $D_{X_1+X_2}$

is considerable non-linearity in the variance curve in Figure 2.12). The basic map exhibits, as expected therefore, long-range correlation similar to the LD-based map: the basic map has a one-lag autocorrelation of 0.6 at the 100kb scale, comparable to the one-lag autocorrelation in the Yoruba LD-based map (0.5) and the CEU LD-based map (0.4). I reflect this in my choice of prior on rates by picking a distribution such that for each interval i, the prior mean rate $m_i = r_{basic_i}$, where r_{basic_i} is the rate estimated for interval i in the basic map.

An important consideration when specifying a prior in this setting is that it should be invariant to the choice of intervals used. The choice of SNPs, and by extension, intervals, is essentially arbitrary (different SNP chips are used, which undergo separate quality checks, and so on). The way in which a region is subdivided should not alter the rate in it. Consider a region between two SNPs (see Figure 2.13 for an illustration). I define the rate of recombination in this region to be Z, which has distribution D_Z, so $Z \sim D_Z$. Now if there was a third SNP between these two SNPs, it would split the region into two and I would model the rate in each subinterval (as illustrated) $X_1 \sim D_{X_1}$ and $X_2 \sim D_{X_2}$. However, given that the presence of the SNP is arbitrary, the total rate should equal the sum of the rates in the two intervals, i.e, $Z = X_1 + X_2$, and $D_Z = D_{X_1+X_2}$. Additionally, I require, for simplicity, that the form of the distribution is the same for all intervals where the expected rate is the same.

A simple distribution that satisfies these conditions is the gamma distribution if the inverse-scale parameter is fixed. If $X_i \sim \Gamma(\alpha_i, \beta)$ are gamma distributed random variables for $i = 1, 2, \ldots, n$ with the same inverse scale parameter β, then

$$\sum_{i=1}^{n} X_i \sim \Gamma\left(\sum_{i=1}^{N} \alpha_i, \beta\right)$$

Therefore, I chose the gamma distribution for specifying the prior rate. In order to specify the two-parameter gamma distribution completely, the variance of the distribution must also be estimated. Under the independence model (Figure 2.11), the number of crossovers in an interval can be modelled with a Poisson distribution, Poisson(λ) where λ is the rate in the interval, and is small. If n is the number of meioses observed, then the variance of the MLE estimator $\hat{\lambda}$ is λ/n. The total number of crossovers included in the basic map is about 2.1 million. This is the number of events I would expect to find, on average, in a pedigree-based map containing 60,000 meioses, given that the sex-averaged genome-wide rate in humans is ∼35 crossovers/meiosis. However, this number must be adjusted because the locations of crossovers are not known precisely. Since the mean resolution of crossovers is approximately 25 times larger than the mean inter-SNP interval in our data (rates cannot be determined to a finer resolution than the inter-SNP interval), I chose $n = 60,000/25 = 2,400$. Analysis of the resolution of the basic map in Section 2.6 suggests, using a very different approach, that this was a reasonable choice for n.

I calculated the parameters of the gamma distribution for each interval using the "methods of moments". If the mean of the distribution is m and the variance is v, the shape and inverse scale parameters are estimated to be $\hat{\alpha} = m^2/v$ and $\hat{\beta} = m/v$ respectively. Therefore, each each interval i, I specify $\alpha_i = n \cdot r_{basic_i}$ and $\beta_i = \beta = n$. I note that this specification preserves the additivity property of the prior distribution by choosing β that is invariant under choices of SNP intervals.

2.5.1 Markov Chain Monte Carlo (MCMC) scheme for estimation of rates

I perform Gibbs sampling [Geman and Geman, 1984] over the recombination rate in each interval and the location of the crossover within each ancestry switch transition. The MCMC proceeds as follows:

- **Initialisation of the MCMC:**

  The Gibbs sampler is initialised with recombination rates sampled from the gamma prior distribution for each interval, as defined above.

- **Running of the MCMC:**

  I set up the Gibbs sampler to sample the joint posterior distribution of the recombination rate in each interval, and where each individual's crossovers occur, within ancestry switch transitions, in the genome.

  To run the sampler from the random initial state described above, I must repeatedly sample:

  (1) Crossover locations given sampled rates, and

  (2) Rates conditional on sampled crossover locations.

  The two steps are described in detail below.

  1. *Sampling the crossover locations within each of the 2.1 million ancestry switch transitions, conditional on sampled rates.* For each filtered ancestry switch transition, I sample the interval in which the crossover occurred conditional on the most recent set of sampled rates. I sample this location under the assumption that exactly one crossover has occurred in an ancestry switch transition.

In principle, it would be ideal to re-run HAPMIX with sampled rates as prior, re-infer ancestry switch transitions and repeat until convergence. This is computationally prohibitive. However, I can alter the PMF produced using HAPMIX (Section 2.4.3) using the piece-wise uniform rate map as input and mathematically re-scale these probabilities to approximate the correct PMF as described below.

I define C as the set of observed ancestry switch transitions. I assume that ancestry switch transitions are independent between and within individuals. Due to shared ancestry, however, some ancestry switches may be inherited from the same ancestor and reflect the same crossover. I attempt to minimize this by filtering out close relatives (Section 2.1).

In this step of the Gibbs sampler, I sample the locations of all crossovers conditional on the vector of recombination rates in all intervals $\bar{r}$. I first compute the likelihood function $P(C|\bar{r})$.

Let the observed ancestry switch transition representing a crossover, say $c \in C$, have SNPs labelled from 1 to n as its end-points. Let corresponding inter-SNP intervals between them be labelled $i \in \{1, \ldots n-1\}$. I assume, without loss of generality, that the ancestry switch transition representing c happens from African to European ancestry. Specifically, if a_j is the ancestry and s_j is the allele type at SNP j, then c can be defined as the event that the ancestry at a_1 is African, the ancestry at a_n is European and that the allelic types s_j are observed at all SNPs on the recombining chromosome. Let r_i represent the crossover rate in interval $i \in \{1, \ldots n-1\}$. Let $\rho = i$ represent the event that a crossover occurs in interval i. I assume, as stated earlier, that exactly one ancestry switch happens in any

observed ancestry switch region. Therefore,

$$P(c|\bar{r}) = \sum_{i=1}^{n-1} P(c, \rho = i|\bar{r}) \tag{2.1}$$

For each interval i, I define $h_{i,l}$ as the haplotype to the left of i, and similarly $h_{i,r}$ as the haplotype to the right of i. Looking at recombination in interval i, therefore,

$$P(c, \rho = i|\bar{r}) = P(h_{i,l}, h_{i,r}, a_1 = \text{African}, a_n = \text{European}, \rho = i|\bar{r})$$

Conditional on crossover in interval i, the ancestries of the left and right haplotypes are fixed as African and European respectively. These two haplotypes are unlinked in the sense that they have separate genealogies prior to the crossover, and the haplotypic frequencies are independent conditional on their respective ancestries. Therefore,

$$\begin{aligned} P(c, \rho = i|\bar{r}) &= P(h_{i,l}, a_1 = \text{African}|\rho = i, \bar{r}) \times \\ &\quad P(h_{i,r}, a_n = \text{European}|\rho = i, \bar{r}) \times \\ &\quad P(\rho = i|\bar{r}) \end{aligned} \tag{2.2}$$

Let's say that the unconditional probability of African ancestry at any position in the genome, and therefore any SNP is θ and of European ancestry is $1 - \theta$ (based on the genome-wide ancestry distribution across a population of African Americans). The probability of which ancestry is found on the template chromosome after a double strand break on the initiating chromosome is obviously independent of recombination rates.

Therefore,

$$\begin{aligned} P(c, \rho = i|\bar{r}) &= P(h_{i,l}|a_1 = \text{African}, \rho = i, \bar{r})P(a_1 = \text{African}) \times \\ & \quad P(h_{i,r}|a_n = \text{European}, \rho = i, \bar{r})P(a_n = \text{European}) \times \\ & \quad P(\rho = i|\bar{r}) \\ &= P(h_{i,l}|a_1 = \text{African}, \rho = i, \bar{r}) \times \theta \times \\ & \quad P(h_{i,r}|a_n = \text{European}, \rho = i, \bar{r}) \times (1-\theta) \times r_i \end{aligned}$$

The probability of observing the haplotype $h_{i,l}$ among Africans and $h_{i,r}$ among Europeans is dependent on historical demographic events and is independent of the occurrence of crossover in the present day individual. Therefore

$$P(c, \rho = i|\bar{r}) = P(h_{i,l}|a_1 = \text{African}, \bar{r})P(h_{i,r}|a_n = \text{European}, \bar{r})\theta(1-\theta)r_i \tag{2.3}$$

Rearranging this equation, I obtain another equation, which will be useful:

$$P(h_{i,l}|a_1 = \text{African}, \bar{r})P(h_{i,r}|a_n = \text{European}, \bar{r})\theta(1-\theta) = P(c, \rho = i|\bar{r})/r_i \tag{2.4}$$

I make the assumption that the probability of observing the haplotype $h_{i,l}$ among Africans and $h_{i,r}$ among Europeans is not influenced by recombination rates $\bar{r}$ in the region. This is not strictly true as haplotypes (or, in other words, the pattern of linkage disequilibrium (LD)) are related to historical recombination rates and maps built in contemporary populations suggest that historical recombination rates and present-day recombination rates are correlated (Sections 1.3.3 and Chapter 4). The reason to make this assumption is that I wish to make a map indepen-

dently of LD patterns, and thus I ignore the effect of rates on LD patterns within a population. We have previously run HAPMIX with piece-wise uniform rates $\bar{r}_{flat}$, and now make the approximation that the probability of the haplotypes, in the absence of ancestry switches within them, are the same whether the rates are $\bar{r}$ or $\bar{r}_{flat}$. Under this assumption, I can rewrite equation 2.4 as

$$P(h_{i,l}|a_1 = \text{African}, \bar{r})P(h_{i,r}|a_n = \text{European}, \bar{r})\theta(1-\theta) \approx P(c, \rho = i|\bar{r}_{flat})/r_{flat_i} \tag{2.5}$$

Substituting equation 2.5 back into equation 2.4, I get

$$\begin{aligned} P(c, \rho = i|\bar{r}) &\approx \frac{r_i}{r_{flat_i}} P(c, \rho = i|\bar{r}_{flat}) \\ &= \frac{r_i}{r_{flat_i}} P(\rho = i|c, \bar{r}_{flat}) P(c|\bar{r}_{flat}) \end{aligned} \tag{2.6}$$

Importantly, all but the final term $P(c|\bar{r}_{flat})$ are known, and the final term is independent of i. I obtain $P(\rho = i|c, \bar{r}_{flat})$, i.e., the probability of crossover in interval i in an ancestry switch region c given uniform rates as discussed before in Section 2.4.3.

I add the possibility of crossover in all intervals by substituting equation 2.6 back into equation 2.1 to get

$$\begin{aligned} P(c|\bar{r}) &= \sum_{i=1}^{n-1} \frac{r_i}{r_{flat_i}} P(\rho = i|c, \bar{r}_{flat}) P(c|\bar{r}_{flat}) \\ &\propto \sum_{i=1}^{n-1} \frac{r_i}{r_{flat_i}} P(\rho = i|c, \bar{r}_{flat}) \end{aligned} \tag{2.7}$$

Therefore, assuming independence of events, I multiply over all $c_k \in C$:

$$P(C|\bar{r}) \propto \prod_k \sum_i \frac{r_i}{r_{flat_i}} P(\rho = i|c_k, \bar{r}_{flat})$$

The probability distribution of a crossover in an interval i conditional on rates in ancestry switch region c_k is therefore:

$$P(\rho = i|c_k,\bar{r}, C) = \frac{P(c_k, \rho = i|\bar{r})}{P(c_k|\bar{r})} = \frac{\frac{r_i}{r_{flat_i}} P(\rho = i|c_k, \bar{r}_{flat})}{\sum_j \frac{r_j}{r_{flat_j}} P(\rho = j|c_k, \bar{r}_{flat})}$$

I use this distribution to sample which interval the crossover occurred in, for each ancestry switch transition. Once I do this for all transitions, I get the set of sampled crossover locations, denoted χ. I count how many events were placed into each interval. This number will be referred to as n_i for interval i.

2. *Sampling the recombination rate in each interval.* In this step of the Gibbs sampler, I sample the recombination rate in each of the 1.3 million SNP intervals conditional on crossovers locations sampled in Step 1. Using Bayes rule,

$$P(\bar{r}|\chi, C) \propto P(\chi|\bar{r}, C)P(\bar{r})$$

Since I know the location of each sampled crossover, any further information about the crossover, such as ancestry, is irrelevant. At the end of Step 1, I calculated the total number of crossovers n_i in each interval i. Therefore, if the total number of intervals in the genome is G and $\{n_1, n_2, \ldots, n_G\}$ is the sequence of the crossover numbers in each interval of the genome,

$$P(\bar{r}|\chi, C) \propto P(\{n_1, n_2, \ldots, n_G\}|\bar{r})P(\bar{r})$$

Using the independence model (Figure 2.11), I have assumed that, given the rate r_i in interval i, the number of events n_i in that interval is independent of the rate in any other interval.

$$P(\bar{r}|C) \propto P(\bar{r}) \prod_{i=1}^{G} P(n_i|r_i) \tag{2.8}$$

I model the number of crossovers in each interval in any individual chromosome as having a Poisson distribution with rate r_i. Adding up over both chromosomes of each individual, I get that if the total number of individuals is I,

$$n_i \sim \texttt{Poisson}(2Ir_i)$$

I use the gamma prior for the recombination rate in each interval, defined previously in this section. If the shape and inverse scale parameters for interval i were α_i and β respectively, then

$$r_i \sim \Gamma(\alpha_i, \beta)$$

Using a gamma prior permits us to utilize the conjugacy of Poisson and gamma distributions to obtain a gamma posterior.

$$r_i|\chi, C \sim \Gamma(\alpha_i + n_i, \beta + 2I)$$

I can now sample a rate for each interval from its respective posterior distribution to complete one iteration of the Gibbs Sampler.

- **Completion of the MCMC:** At least 5 chains were initialized for each chromosome, each with independently sampled rates. At least 1 million iterations were performed for each chain, of which the first 300,000 were discarded as

burn-in.

I computed the mean recombination rate in each interval using every 500th sample for each chain. I assessed the convergence of the MCMC by comparing the mean recombination rate estimated from each chain. The correlation was 0.999 or greater for every pair of chains for all chromosomes. I then pooled all the chains together to produce a mean recombination rate. I normalized the rates so that the total genetic map length is equal to that of the HapMap2 LD-based map [Myers et al., 2005]. We call this map the *AA map* henceforth.

Figure 2.14 illustrates the action of the MCMC over three ancestry switch transitions in the same region.

2.6 Properties of the AA map

One important advantage of using the MCMC procedure is that I can use it to generate an uncertainty estimate for the recombination rate in each interval. This is because the MCMC re-samples the rates in each iteration and thus builds up a probability distribution.

I can, therefore, use samples collected during the procedure to estimate the resolution of the AA and basic maps at various size scales. For any chosen size scale, say S, I linearly interpolate the map (AA or basic) at fixed positions $p_1, p_2, \ldots, p_N$, which are S bases apart. I can pick a sample from the MCMC, and perform the same interpolation. Let the interpolated rates from the sample be $m_1, m_2, \ldots, m_N$ and the map be $r_1, r_2, \ldots, r_N$. I can now calculate the correlation of two sets of interpolated rates at scale S. I calculate the correlation coefficient of the map with every 10,000th sample from the MCMC (after discarding burn-in), and then calculate the mean correlation coefficient across samples to get the estimated correlation of the map to the averaged map inferred by MCMC. This is a measure of variation at a given scale.

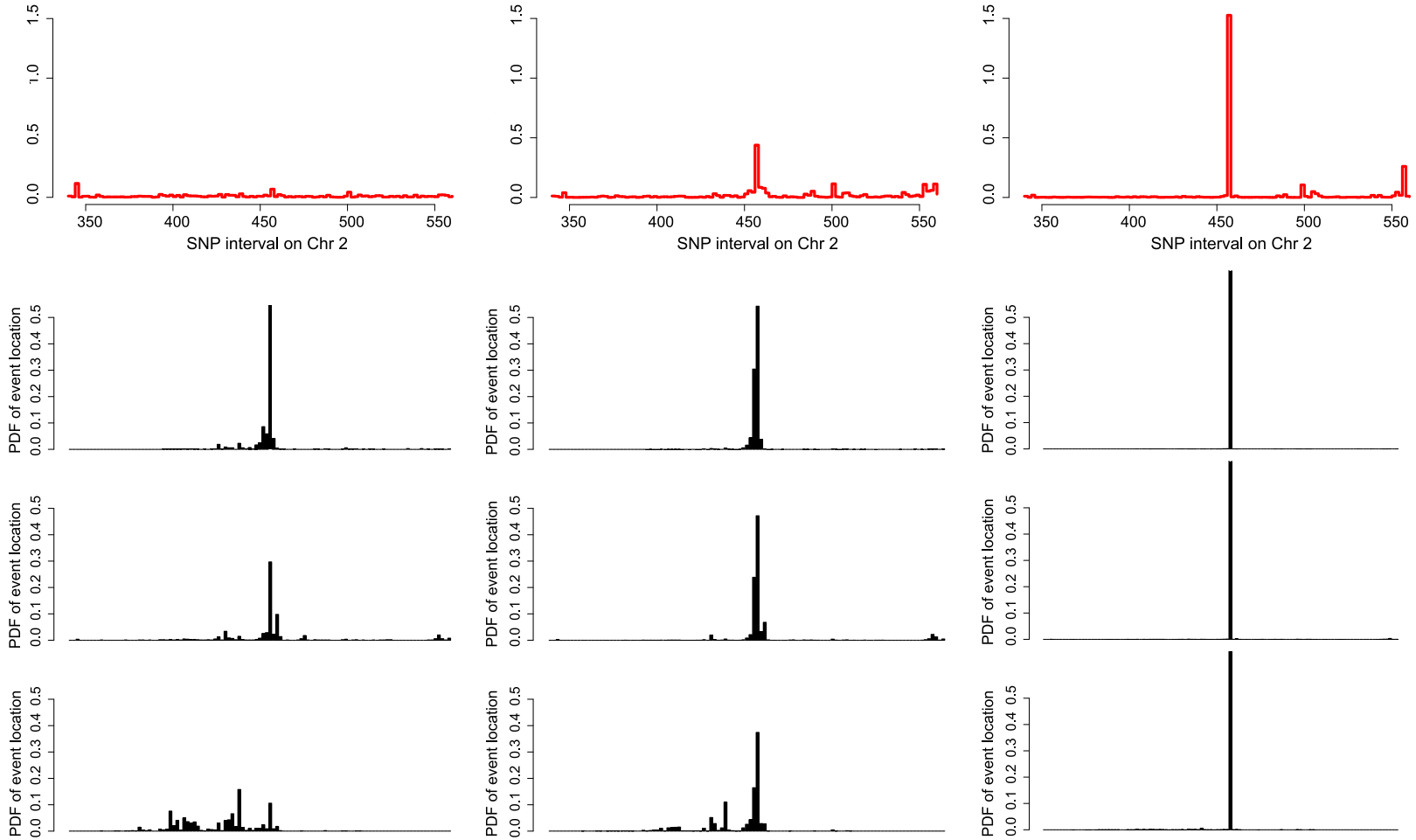


Figure 2.14: Localisation of hotspots by the Gibbs Sampler. The top panels show the recombination rate (cM) estimated by the MCMC at different stages in the chain for a small region in Chromosome 2. The bottom panels show the probability mass function (PMF) calculated for three events occurring in different individuals in the same region. The chain was started in the left panel using a uniform recombination rate per base of 1.1cM/Mb. The middle panel shows the state of the chain after 100 iterations and the final panel shows the state of the chain after 10,000 iterations.

Thinking of a sample as a possible "true" map, this can be thought of as an estimate of the correlation between the true recombination rate and our AA map, at any given scale. Estimated correlations for the AA map across a range of size scales are reported in Table 2.4.

Another way to think of map resolution may be as follows. The number of crossovers in an interval, assuming crossovers occur independently, can be modelled as having a Poisson distribution with mean λ and standard deviation $\sqrt{\lambda}$. Consider the coefficient of variation (CV), i.e., the ratio of the standard deviation of sampled rates with their expectation, i.e., $1/\sqrt{\lambda}$. CV of an interval decreases as the interval size increases, and at some scale becomes 1. If events were perfectly resolved, this would be the scale at which 1 event occurs, on average. Therefore, CV can be thought of as the 'density' of inferred events after accounting for location uncertainty. Therefore, we use it as a measure of resolution.

I estimate the coefficient of variation for a map at size scale S as follows: I calculate the mean rate of the map $\hat{\mu} = \frac{1}{N}\sum_i^N r_i$. For every 10,000th sample map, I calculate the mean standard error $e = \sqrt{\frac{1}{N}\sum_i^N (m_i - r_i)^2}$. If the posterior samples the correct distribution of potential true maps, the average of these values estimates the mean standard error of the map, which I call $\hat{\sigma}$. The coefficient of variation is $\hat{\sigma}/\hat{\mu}$. Coefficients of variation calculated at a range of size scales are reported in Table 2.4.

Using this definition, I find that the resolution of the basic map is 35 kb and the AA Map is greatly improved 6 kb. Given that the size of the human genome is approximately 3 Gb, a resolution of 6 kb corresponds to roughly 500,000 perfectly resolved crossovers[4]. Further, I demonstrate that our uncertainty estimates are well-calibrated in the next chapter.

Despite the high resolution of the AA map, there are some limitations. Ancestry

[4]Similarly, the equivalent number of 'perfectly resolved' events in the basic map is approximately 86,000. This is the number of events we would expect to find, on average, in a combined sample with equal numbers of male and female meioses of size ∼2500, given that the sex-averaged genome-wide rate in humans is ∼35 crossovers/meiosis.

Scale (interval size)	Estimated correlation of AA Map to the 'true' map	Estimated coefficient of variation of AA map
3 kb	0.93	1.41
10 kb	0.96	0.73
30 kb	0.98	0.36
100 kb	0.99	0.17
300 kb	1.00	0.08
1 Mb	1.00	0.04
3 Mb	1.00	0.02

Table 2.4: Assessment of map resolution at different size scales. The correlation of the AA map to the 'true' map is estimated as the mean correlation of the AA map with maps sampled during the MCMC procedure. The coefficient of variation is the ratio of the mean standard error of the maps sampled during the MCMC procedure relative to the AA map with the mean recombination rate (details in text). The coefficient of variation is 1 for the size scale 6 kb.

switches are inherited from both maternal and paternal ancestors. They are not distinguishable and therefore I cannot infer separate male and female maps. Second, I may miss crossovers in regions where crossovers are difficult to detect, for example, due to low SNP density or lower differentiation between African and European haplotypes. The resolution may be lower in such regions.

A final caveat is that the map building procedure assumes that all individuals are unrelated and that every ancestry switch is unique. In reality, there is likely to be some shared ancestry despite filtering of close relatives, resulting in multiple counting of some crossovers. Wegmann et al. [2011] have independently developed a procedure similar to the approach presented in this thesis to identify crossovers via ancestry switches. They use simulations to estimate the fraction of crossovers that are counted more than once due to segments inherited identical-by-descent among present-day individuals. They estimate that in a sample containing just under 3,000 individuals, with an effective population size of 20,000, the fraction of unique 'singleton' ancestry switches is about 94%, while in a more realistic population size of 200,000, over

99.8% of the switches are unique (the results of their analysis are reproduced in Appendix B). A detailed analysis of this question in our setting would require a similarly sophisticated simulation, or an estimate of identity-by-descent (IBD) among all pairs of samples, and is beyond the scope of the present work. However, a rough estimate can be obtained by using informal arguments based on the coalescent model without recombination. Consider a constant-sized panmictic population of size N, and a sample of size n. Both n and N are large, however, n/N remains constant, say p, in the limit $N \to \infty$. Without loss of generality, I assume that a mutation occurred i generations ago on lineage #1, and I estimate the probability that none of the remaining $n-1$ lineages coalesce with this lineage subsequent to appearance of the mutation. Here i is small, independent of n and N, and less than the number of generations since admixture, which is approximately 6 for African-Americans. This probability is lower bound by the probability that, for each generation back in time from 1 to i, each of the $n-1$ lineages finds a different common ancestor than lineage #1, i.e., it is $((\frac{N-1}{N})^{n-1})^i \approx ((1-\frac{1}{N})^N)^{pi}$. As $N \to \infty$, $((1-\frac{1}{N})^N)^{pi} \to e^{-pi}$. In other words, the fraction of singletons remains approximately unchanged if the size of the sample and the effective population size increase in the same proportion. Assuming that the effective population size is about 200,000 [Wegmann et al., 2011], the sample size in this work is about 15% of the total effective population size. As shown in Appendix B, this corresponds to the simulation in Wegmann et al. [2011] of population size 20,000, and leads to an estimate of approximately 6% of crossovers counted more than once in the sample. This is not expected to lead to differential bias in the map, as all regions in the genome are equally likely to come from a common ancestor. While this may lead to overestimation of map precision, it is very unlikely to lead to false positive inference of a hotspot, as the large sample size in this work implies that a typical hotspot will be supported by several dozen crossovers.

Chapter 3

Recombination in African-American families

In the process of filtering out closely related individuals in order to build the AA map (Chapter 2), I noticed many offspring-parent and sibling-sibling relationships represented in the data[1]. These relationships frequently formed nuclear families, and thus provided another resource for investigating recombination. As introduced in Section 1.3.2, children inherit one chromosome from each of their parents, and this inherited chromosome is a mosaic of each of the two homologous parental chromosomes. Therefore, differences in the chromosomes inherited by full siblings signal crossovers in their parents. With this in mind, I focussed on families which had two or more children genotyped. However, in the majority of these cases, only one of their parents was also genotyped.

Since these families are also African American, this resource gave us the opportunity to find direct biological insights into recombination, as will be shown in Chapter 5. Secondly, the approach in Chapter 2 of using ancestry detection to identify crossovers and building a genetic map is novel. Therefore, it was important to validate the

[1]I deduced these relationships from pairwise identity by descent (IBD) estimates.

approach using an independent method.

Identifying crossovers in these pedigrees, however, was not straightforward due to the missing parents. It meant, for instance, that the heuristic approach used in Coop et al. [2008] was inapplicable. Kong et al. [2010] use extended Icelandic pedigrees to inform their phasing of each individual, and that approach is also inapplicable here as we very rarely have grandparents or other close relatives genotyped. Existing methods for inferring haplotypes in families [Lander and Green, 1987; Kruglyak et al., 1996; Abecasis et al., 2002] handle missing data, however, they assume that markers are unlinked and do not explicitly account for genotyping errors. As I discuss below, both of these issues lead to spurious crossover calls.

To find crossovers in this setting, I have developed an HMM-based approach that infers crossovers at high SNP density and with missing parents. This approach extends and adapts the widely used Lander-Green algorithm [Lander and Green, 1987], which is introduced below in Section 3.1. I describe the data used in Section 3.2. I discuss our approach in Section 3.3, and the detailed algorithm in Section 3.4. An overview of the crossovers that I identified by this approach is presented in Section 3.5.

3.1 Inference in pedigrees

In humans, the chromosomes that mothers and fathers transmit to their children will typically have several dozen crossovers genome-wide. These crossovers can be detected by tracking how the transmission of alleles of heterozygous markers differs among multiple offspring of the same parents. Some of the parental genetic material will, in turn, be identical to that of other relatives descended from the same ancestors. This imposes a complex correlation structure among the genotypes of individuals in a family, but one that can be explicitly modelled if the genealogy connecting those individuals is known. As introduced in Section 1.3.2, calculating the likelihood of a

pedigree involves calculating the likelihood of observing the genotypes of the founder members in a pedigree and the transmission of alleles from them down into the rest of the pedigree. In the most general case, if there are m possible alleles per marker and s markers typed in a pedigree containing n individuals, this would involve calculating the likelihood of $O(m^{ns})$ states, which would quickly become intractable [Lander and Green, 1987]. However, this calculation ignores biological knowledge about the transmission of genetic information from parent to child. Certain algorithms, such as the Elston-Stewart algorithm [Elston and Stewart, 1971] leverage independence in special pedigree structures. However, they continue to require computation time that is exponential in the number of markers, making them unsuitable for use with densely typed SNP chips for finely resolved detection of crossovers.

Another approach to likelihood computation in pedigrees is the *Lander-Green algorithm* [Lander and Green, 1987]. Consider a trio of father, mother and child, and let us call the paternal chromosomes $H_{\text{paternal},0}$ and $H_{\text{paternal},1}$ and the maternal chromosomes $H_{\text{maternal},0}$ and $H_{\text{maternal},1}$. The child inherits one chromosome from the mother and one from the father. The maternally inherited chromosome may be a mosaic of $H_{\text{maternal},0}$ and $H_{\text{maternal},1}$ due to crossovers, and will be completely specified by the knowledge of which of those two chromosomes was copied at each locus. So, we define the maternal inheritance vector I_{maternal} at each locus j as

$$I_{\text{maternal}}(j) = \begin{cases} 0 \text{ if } H_{\text{maternal},0} \text{ is inherited at } j \\ 1 \text{ if } H_{\text{maternal},1} \text{ is inherited at } j \end{cases}$$

And similarly, for the paternal inheritance vector,

$$I_{\text{paternal}}(j) = \begin{cases} 0 \text{ if } H_{\text{paternal},0} \text{ is inherited at } j \\ 1 \text{ if } H_{\text{paternal},1} \text{ is inherited at } j \end{cases}$$

Going along a chromosome, I_{maternal} flips at locus j when there is a crossover in the maternally inherited chromosome between loci $j-1$ and j. Under the assumption of no interference, the probability of crossover between loci $j-1$ and j is independent of crossover in any other locus, and therefore I_{maternal} and I_{paternal} are first-order Markov processes. The trio can therefore be modelled as a Hidden Markov Model with the inheritance vectors as the hidden states. The forward-backward algorithm can be used for estimating the inheritance vectors of the child in time linear in the number of loci. Each meiosis is independent, and therefore, additional children in the family will have independent inheritance vectors. In general, only the genotypes will be known, and $H_{\text{maternal},0}$, $H_{\text{maternal},1}$, $H_{\text{paternal},0}$, and $H_{\text{paternal},1}$ will also be hidden states. The model is illustrated for a nuclear family with two children in Figure 3.1.

The Lander-Green algorithm extends this idea to arbitrarily large and complex pedigrees, where the calculation proceeds jointly over all members of the pedigree one locus at a time. Calculating the transmission of alleles at a single locus is still exponential in pedigree size in general. However, this is not a major problem for small pedigrees and, in addition, several implementations exist which extend the Lander-Green idea with computational improvements [Kruglyak et al., 1996; Abecasis et al., 2002].

A major issue with these approaches, however, is that they do not handle genotyping error. This is not a serious issue when the goal is to infer genotypes, however, it is problematic for detecting crossovers as this will frequently result in false inference of pairs of crossovers. Another difficulty when using dense SNP chips is specifying and calculating the joint probabilities of genotypes in founders and non-founders respectively. This is because markers are treated as independent in these approaches while they may in fact be in tight linkage. Both these problems are especially serious when there is a lot of missing data, and are addressed in our algorithm, described next.

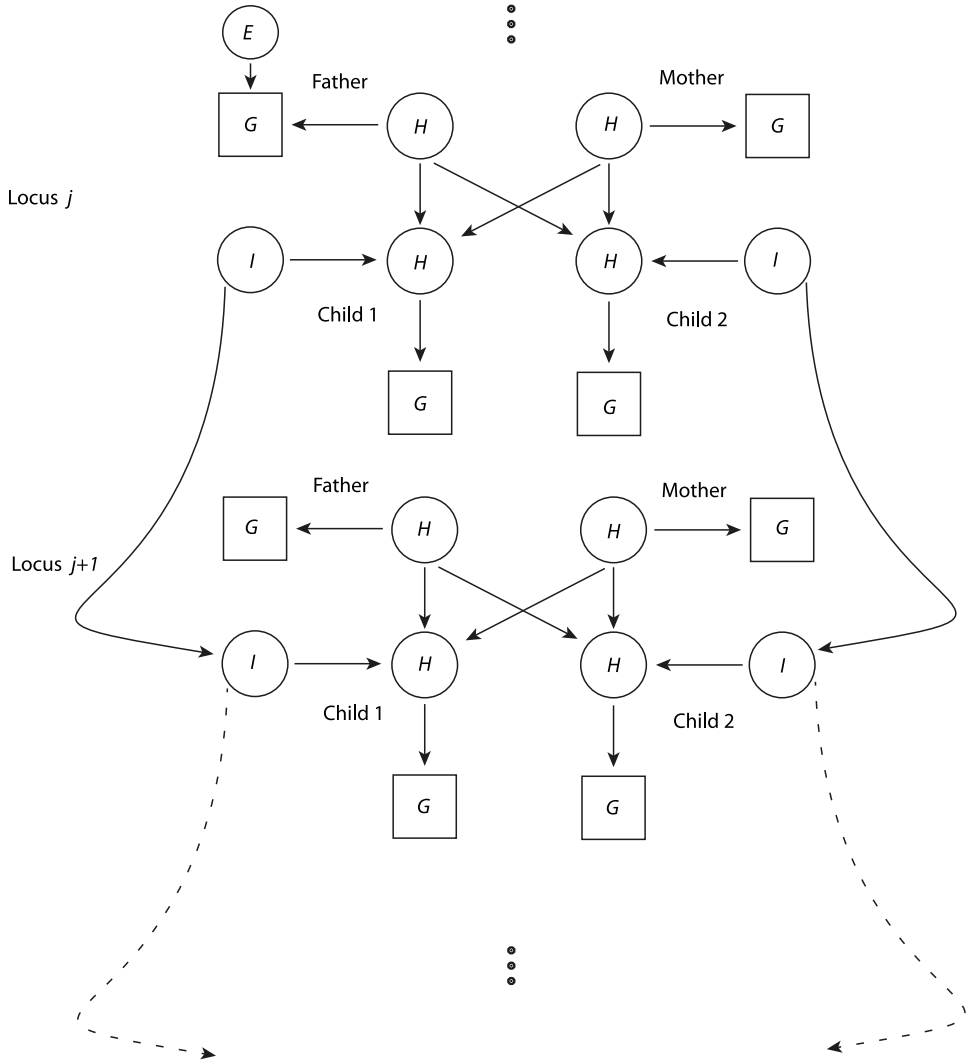


Figure 3.1: Hidden Markov Model for likelihood calculations in a nuclear family.

The observed variables are shown as squares and the hidden variables as circles. Genotypes G are dependent on the underlying phased haplotypes H. Observed genotypes may be subject to error E (errors are modelled in our algorithm but not in the Lander-Green algorithm). Only one instance of E (top left, for father) is shown for clarity, but is replicated for all G states. The parental haplotypes are assumed to be independent at each loci, which is valid only if the loci are unlinked. The probabilities of the parental haplotypes may be calculated from the population frequencies of the alleles in the haplotypes, under the assumption of Hardy Weinberg equilibrium. In some pedigrees, one parent may not be genotyped, in which case the genotypes of that parent are treated as missing data.

Children's haplotypes depend on the parental haplotypes, and their own maternal and paternal inheritance vectors I. The inheritance vectors follow a first-order Markov process as explained in the main text. Crossover is inferred when there is a change in the inheritance vectors of a child.

3.2 Data

In this study I analyzed recombination in 222 nuclear families, representing 1,056 meioses. The studies that contributed samples were CARe (135 families) and CHOP (87 families). In a few cases, the familial relationships were known *a priori*, however, in most situations, first-order relationships were deduced from the mean and variance of pairwise IBD estimates produced by running PLINK [Purcell et al., 2007]. Families were chosen for analysis on the basis that they had at least two full-siblings and one parent genotyped. Of the 222 families thus selected, 27 had genotype data for both parents while the rest had data for only one parent.

CARe samples were typed on Affymetrix 6.0 arrays and had 580,000 SNPs after data curation. CHOP samples were genotyped on either Illumina 610-Quad or Illumina HumanHap550 arrays and had 491,572 SNPs after data curation.

3.3 Calculating the likelihood of nuclear families with missing data and genotyping error

I implement a three-step algorithm for detecting crossovers in these families. In the first step, I implement the Hidden Markov Model illustrated in Figure 3.1, on a thinned set of approximately unlinked markers. The expected value of the inheritance vector of each child is inferred *a posteriori* at every locus using the forward-backward algorithm. If the genetic data of one parent is unavailable, the HMM treats it as missing data, and jointly infers the parental genotype together with the inheritance vector of each child. How well the parental genotype can be imputed in this procedure depends on the number of genotyped children. In a family with 2 children we expect to see both alleles from the missing parent half the time. That fraction is three-quarters in a family with 3 children, seven-eighths with 4 children and so on.

In the second step, I identify crossovers by post-processing the expected inheritance vectors for each child.

In the third step, I run a modified HMM *separately* in each small region containing a crossover, with the *full* set of SNPs. This HMM calculates the likelihood of the family in this region *conditional on no more than a single crossover* in the region. Section 3.4 provides a detailed description of the algorithm.

The reason for this multi-step procedure is as follows: the HMM as depicted in Figure 3.1 has a correctly specified transition matrix for the parental haplotypes only if the marker loci are unlinked. Reducing the set of genotyped SNPs to an unlinked set causes the resolution of crossover breakpoints to be greatly impaired (approximately 90% of SNPs are removed in an adequate thinning procedure). An alternative may be to run the HMM with the full set of SNPs, and hope that the mis-specification of the transition probability does not lead to biases in inference. This does, in fact, work well for families in which both parents are genotyped. There is sufficient information in the data and the prior probability of recombination in any inter-SNP interval is low, which prevents spurious crossovers from being inferred. Consider, however, the case where one of the parents is not genotyped. Let's say this parent is homozygous for a haplotype containing an extensive LD-block spanning multiple SNPs. All children will therefore carry this haplotype. The HMM will integrate over the following two possibilities: (i) the truth, i.e., the parent is homozygous for the haplotype and all children carry the haplotype regardless of whether they are identical-by-descent (IBD) therein, or (ii) the parent carries one copy of the haplotype, and all children are IBD at that haplotype. In case of the second possibility, if the children are not IBD in either region flanking the haplotype, the HMM will have to infer two crossovers, one at each end of the haplotype. Depending on how likely the haplotype is under the (false) assumption of unlinked markers and the number of children in the family, this may lead to pairs of false crossovers being inferred.

The multi-step approach ameliorates this problem. The use of unlinked markers in the first step prevents false crossovers from being identified in the first instance[2]. I then restrict the algorithm to a region in the vicinity of the *true* crossover and condition the algorithm on identifying no more than one crossover. The only way for the algorithm to call a spurious crossover in this region when run with the full set of markers then is for the false event to appear more likely than the true event. The choice of the region in which to run the final HMM in the third step makes this unlikely: the algorithm includes several hundred SNPs on either side of the crossover (details in the algorithm description below). This means that the true crossover will generally be supported by several megabases of IBD among a pair of children on one side of the crossover, and several megabases of non-IBD on the other. On the contrary, there are two possibilities for false inference of IBD for a haplotype: (i) the entire haplotype is included in the region, or (ii) only a portion of it is included. In the first case, two crossovers would have to be inferred to achieve IBD, and that is disallowed by the HMM. In the second case, the partial haplotype is biologically highly unlikely to have the several megabase support of the true crossover, except possibly if the missing parent has very high homozygosity. Parents where there is evidence of such recent inbreeding are identified and their inferred crossovers are discarded in a post-processing step.

3.4 Algorithm for detecting crossovers

Here I give a detailed description of the scheme outlined above.

1. *Crossover detection using unlinked markers.* In this step, I thinned SNPs to create a set of unlinked and informative markers. I selected SNPs with minor

[2]This is not strictly true, and would fail in the case where the missing parent has high homozygosity, for example, due to recent inbreeding. This case has a distinctive signature and is easy to detect and filter out (Figure 3.3).

allele frequency (MAF) greater than 0.4, as long as the r^2 between any pair of selected SNPs was less than 0.2. The number of SNPs remaining after thinning was 50,468 for CARe and 51,613 for CHOP.

The HMM implemented in this step is illustrated for a family with two children in Figure 3.1.

States. Let the family contain n children, labelled C_1, C_2, ..., C_n. The observed variables are the genotypes of all available members in the family, called G_{maternal}, G_{paternal}, G_{C_1}, ... G_{C_n}, collectively referred to as $\mathcal{G}$. The underlying phased haplotypes are hidden variables of all members in the family. They are labelled $H_{\text{maternal},0}$, $H_{\text{maternal},1}$, $H_{\text{paternal},0}$, and $H_{\text{paternal},1}$ for the parents. Since one chromosome is inherited from each parent, the children's haplotypes are labelled $H_{C_1,\text{maternal}}$, $H_{C_1,\text{paternal}}$ for child C_1 and so on. All SNPs are assumed to be biallelic and labelled 0 or 1. Therefore, there are four possibilities each for the maternal and paternal haplotypes at each site, and the total number of sites is N.

The maternal and paternal inheritance vectors for child C_1 are called $I_{C_1,\text{maternal}}$ and $I_{C_1,\text{paternal}}$ respectively. As defined above, for SNP j

$$I_{C_1,\text{maternal}}(j) = \begin{cases} 0 \text{ if } H_{\text{maternal},0} \text{ is inherited at } j \text{ in child } C_1 \\ 1 \text{ if } H_{\text{maternal},1} \text{ is inherited at } j \text{ in child } C_1 \end{cases}$$

and similarly for the paternal inheritance vector and other children. The haplotypes for the children are completely specified from the parental haplotypes and the inheritance vectors.

$$H_{C_1,\text{maternal}}(j) = \begin{cases} H_{\text{maternal},0}(j) \text{ if } I_{C_1,\text{maternal}}(j) = 0 \\ H_{\text{maternal},1}(j) \text{ if } I_{C_1,\text{maternal}}(j) = 1 \end{cases}$$

Genotypes may be observed with error. The error model is assumed to be such that each haplotype is independently subject to an error probability p_e, which I fix at 0.1%. For instance, the genotype is observed to be 2 if both haplotypes are 0 with probability p_e^2. For maternal genotypes,

$$P_{\text{emit}}(G_{\text{maternal}}(j)) = \begin{cases} p_e^2 & \text{if } \text{abs}(G_{\text{maternal}}(j) - (H_{\text{maternal},0}(j) + H_{\text{maternal},1}(j)) = 2 \\ 2 \cdot p_e \cdot (1 - p_e) & \text{if } \text{abs}(G_{\text{maternal}}(j) - (H_{\text{maternal},0}(j) + H_{\text{maternal},1}(j)) = 1 \\ (1 - p_e)^2 & \text{if } G_{\text{maternal}}(j) = H_{\text{maternal},0}(j) + H_{\text{maternal},1}(j) \\ \propto 1 & \text{if } G_{\text{maternal}}(j) \text{ is unknown, wlog} \end{cases}$$

and similarly for paternal and child genotypes.

All hidden states, i.e., haplotypes, inheritance vectors and errors, are collectively denoted as $\mathcal{H}$.

Transition Probabilities. Since SNPs are assumed to be unlinked, transition probabilities between the parental genotypes at locus s and $s+1$ are simply the prior probabilities of the genotypes at $s+1$. These are set to be the population allele frequencies of the respective alleles, estimated from panels of unrelated African Americans in whom these SNPs are typed (datasets of unrelated individuals were described in Chapter 2).

Transition in the inheritance vector of a child represents a crossover. If there are n children in the family, there may be up to $2n$ crossovers (one per meiosis) in a small interval. To keep the computation tractable, however, I assume that no more than one crossover occurs across all $2n$ inheritance vectors in any inter-SNP interval. Recombination rates in humans are low and the SNP density in this analysis is high enough that this is a reasonable assumption. If r is the recombination rate in the interval between SNPs j and $j+1$, then the (Poisson) probability that there is no crossover at all in this interval is e^{-2nr}. The

probability of crossover for a single inheritance vector is therefore $(1-e^{-2nr})/2n$. This assumption introduces dependency in the transition probabilities of the different inheritance vectors, and the transition probability between loci j and $j+1$ is defined jointly for all inheritance vectors:

$$P_{\text{trans}}(I_{.,.}(j) \rightarrow I_{.,.}(j{+}1)) = \begin{cases} e^{-2nr} & \text{if } I_{.,.}(j) = I_{.,.}(j+1) \\ \frac{1-e^{-2nr}}{2n} & \text{if } \exists! l : I_{C_l,\text{maternal}}(j) \neq I_{C_l,\text{maternal}}(j+1) \oplus \\ & \quad \mathrm{I}_{C_l,\text{paternal}}(j) \neq I_{C_l,\text{paternal}}(j+1) \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The first case represents zero crossovers, the second case represents a single crossover ($\oplus$ denotes the XOR binary operation), the final case represents more than one crossover, which is not permitted.

Forward backward algorithm. I run the forward-backward algorithm and use the forward and backward probabilities [Rabiner and Juang, 1986] to calculate the posterior probability of parental haplotypes and inheritance vectors at each locus. The probability of crossover between loci j and $j+1$ is simply the the posterior joint probability of all states $\mathcal{H}_j$ and $\mathcal{H}_{j+1}$ that involve inheritance vectors representing a crossover, as defined in equation 3.1. In other words, if $\alpha_j(\mathcal{H}_j)$ and $\beta_j(\mathcal{H}_j)$ are the forward and backward probabilities respectively for the hidden states, then the probability of crossover between loci j and $j+1$ is

$$P_{\text{crossover}}(j, j+1|\mathcal{G}) = \frac{\sum_{\mathcal{H}_j} \sum_{\mathcal{H}_{j+1} \in \mathcal{R}} P(\mathcal{G}, \mathcal{H}_j, \mathcal{H}_{j+1})}{P(\mathcal{G})}$$

where $\mathcal{R}$ is the set of states $\mathcal{H}_{j+1}$ whose inheritance vectors are such that they differ from the inheritance vectors in $\mathcal{H}_j$ by exactly one crossover. Maternal and paternal crossover probabilities can be calculated by restricting $\mathcal{R}$ to contain

vectors that differ in exactly one maternal or paternal crossover respectively. Using the forward and backward probabilities, I get

$$P_{\text{crossover}}(j, j+1|\mathcal{G}) = \frac{\sum_{\mathcal{H}_j} \sum_{\mathcal{H}_{j+1} \in \mathcal{R}} \alpha_j(\mathcal{H}_j) P_{\text{emit}}(\mathcal{G}_{j+1}) P_{\text{trans}}(\mathcal{H}_j, \mathcal{H}_{j+1}) \beta_{j+1}(\mathcal{H}_{j+1})}{P(\mathcal{G})} \quad (3.2)$$

where $P(\mathcal{G}) = \sum_{\mathcal{H}_j} \alpha_j(\mathcal{H}_j)\beta_j(\mathcal{H}_j)$, which is the same for all j.

2. *Identifying individual crossovers.* The cumulative recombination probability going along a chromosome is illustrated for a typical family in Figure 3.2. In this step I use a dynamic programming algorithm to identify the end points of crossovers. The endpoints are chosen such that the PDF is maximally steep and contains a crossover with probability of at least 95%. A region including the crossover and 700 flanking SNPs from either side of the crossover showing no evidence of an additional crossover are used to refine the breakpoints in Step 3. If 700 SNPs are not available, then as many as SNPs as possible are used. In the event that there are overlapping maternal and paternal crossovers, they are filtered out at this stage and not taken to Step 3.

   The assumption of approximately unlinked markers in Step 1 breaks down when a parent has a high degree of homozygosity, for instance, due to recent inbreeding. This results in inference of recombination throughout large portions of a chromosome (Figure 3.3), as the paucity of heterozygous markers means there is insufficient information to localize underlying crossovers. Crossovers from such parents are discarded and not considered for Step 3.

3. *Refining crossover breakpoints using a second HMM.* In this step I fine-map crossover breakpoints using an HMM similar to one in Step 1, but conditional on having exactly one crossover. In general, it is not possible to condition on a fixed number of crossovers *a priori* in this framework (i.e., without introducing additional states), as the first-order Markov property would no longer hold.

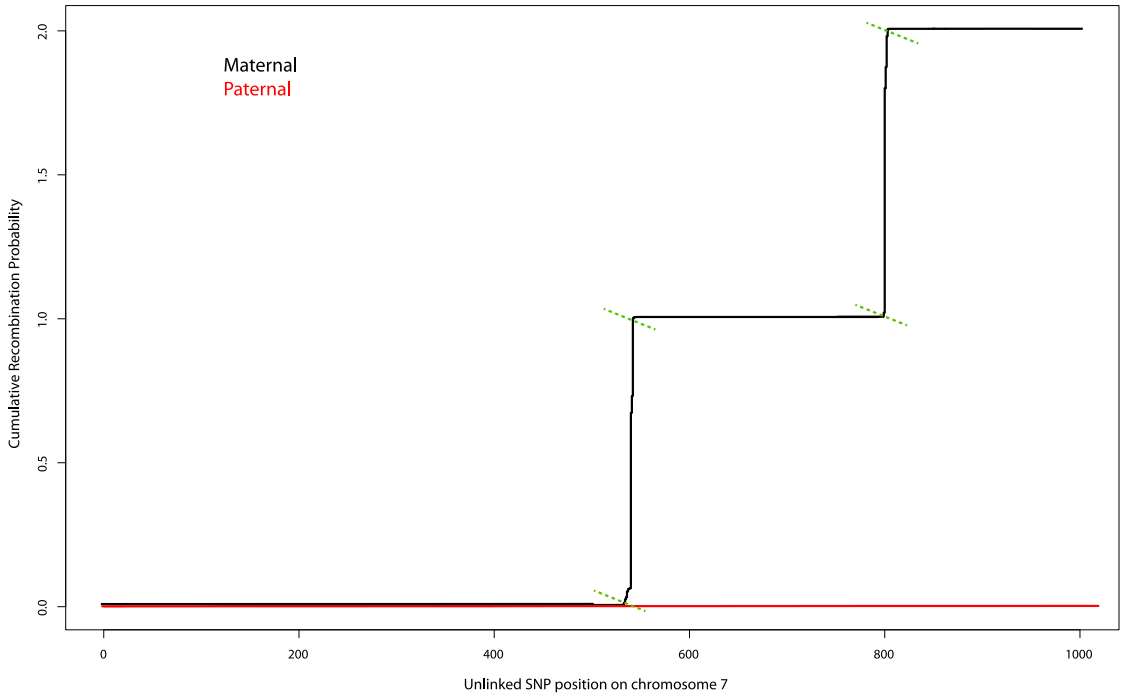


Figure 3.2: Cumulative recombination probability along a chromosome in a typical family. Each unit change in probability corresponds to one crossover. The green dashed lines show the boundaries of the crossover determined by Step 2 of the algorithm.

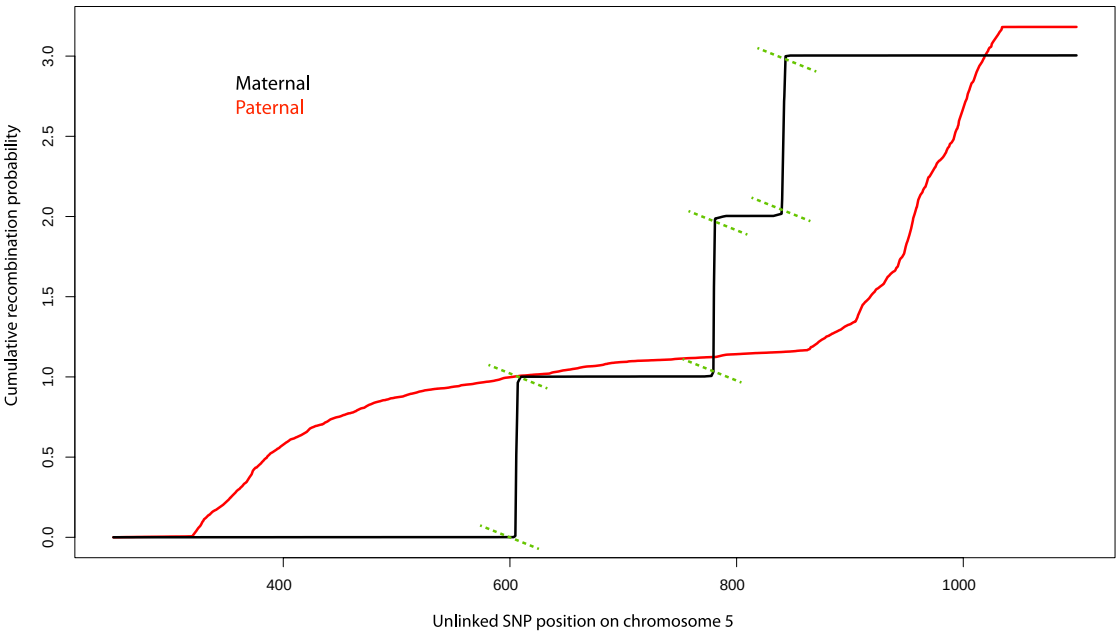


Figure 3.3: Similar to Figure 3.2 except in this case the father was highly homozygous. Such cases are easily detected and removed from the analysis.

In the particular case of one crossover, however, it is possible to adapt the framework without loss of computational performance by implicitly running $s-1$ first-order Markov chains (where s is the number of loci) such that in any chain one and only one inter-SNP interval has a non-zero probability of crossover. This implies that only a single crossover can happen in the whole region since only a single crossover is permitted in any inter-SNP interval. This can be achieved by changing the calculation of the forward probability and backward probabilities of the HMM in Step 1. I keep track of forward and backward probabilities $\tilde{\alpha}$ and $\tilde{\beta}$ respectively, based on a transition vector permitting *no crossovers*, i.e.,

$$\tilde{P}_{\text{trans}}(I_{.,.}(j) \rightarrow I'_{.,.}(j+1)) = \begin{cases} 1 & \text{if } \forall l : I_{C_l,\text{maternal}}(j) = I'_{C_l,\text{maternal}}(j+1) \;\&\; \\ & \quad I_{C_l,\text{paternal}}(j) = I'_{C_l,\text{paternal}}(j+1) \\ 0 & \text{otherwise} \end{cases}$$

The probability of crossover in the j^{th} Markov chain, permitting crossover between loci j and $j+1$ only can then be calculated by modifying equation 3.2 as follows:

$$P_{\text{crossover}}(j, j+1|\mathcal{G}) = \frac{\sum_{\mathcal{H}_j} \sum_{\mathcal{H}_{j+1} \in \mathcal{R}} \tilde{\alpha}_j(\mathcal{H}_j) P_{\text{emit}}(\mathcal{G}_{j+1}) P_{\text{trans}}(\mathcal{H}_j, \mathcal{H}_{j+1}) \tilde{\beta}_{j+1}(\mathcal{H}_{j+1})}{P(\mathcal{G})}$$

As before, $\mathcal{R}$ is the set of states $\mathcal{H}_{j+1}$ whose inheritance vectors are such that they differ from the inheritance vectors in $\mathcal{H}_j$ by exactly one crossover. Calculation of $P(\mathcal{G})$ requires summing the likelihood over each of the n chains, plus the probability of no crossover at all.

$$P(\mathcal{G}) = \sum_{\mathcal{H}_N} \tilde{\alpha}_N(\mathcal{H}_N) + \sum_j \sum_{\mathcal{H}_j} \sum_{\mathcal{H}_{j+1} \in \mathcal{R}} \tilde{\alpha}_j(\mathcal{H}_j) P_{\text{emit}}(\mathcal{G}_{j+1}) P_{\text{trans}}(\mathcal{H}_j, \mathcal{H}_{j+1}) \tilde{\beta}_{j+1}(\mathcal{H}_{j+1})$$

This calculation can be done without significant increase in computing time since it requires calculating only one additional forward probability per site.

3.5 Discussion

I validated this three-step procedure in a subset of the CARe dataset containing 118 families, of which 16 families had both parents genotyped. To determine its effectiveness in the face of missing parental data, I ran it twice for all 16 families where genotype data was available for both parents: once with the genetic data of both parents and once by masking the genotypes of the father. All crossovers inferred with both parents were also inferred with one parent and vice versa, although the resolution was impaired when only one parent was available. This is expected since the genotype of the missing parent is being inferred jointly with crossovers, and it may require more data to disambiguate IBD of children from homozygosity in a parent. The median event resolution in families with both parents genotyped was 27 kb, whereas it was 40kb in families with a missing parent.

I inferred 32,180 crossover events in total in the full dataset containing 222 nuclear families, of which 26,664 were resolved within 100 kb and 12,953 within 30 kb. Table 3.1 compares the total autosomal genetic map length estimated in our study with other published maps, all of which were estimated in Europeans. While estimates from the different maps are similar, the rates in this study are estimated to be somewhat higher than the other maps. This may be, in part, due to a greater number of SNPs genotyped in this study, particularly in the sub-telomeric regions. Another factor may be the larger number of crossovers in a subset of individuals with African ancestry, as discussed further in Section 5.3.2.

The median event resolution was 38 kb. I added the posterior probabilities of events that were resolved within 100kb to produce male, female and sex-averaged

	Female	Male
African Americans	42.6	29.9
deCODE [Kong et al., 2010]	40.7	22.9
Hutterites [Coop et al., 2008]	39.6	26.2
Rutgers [Matise et al., 2007]	44.0	28.3
deCODE [Kong et al., 2002]	42.8	25.9

Table 3.1: Autosomal genome-wide rate (in Morgans) in different pedigree maps

genetic maps.

I was able to validate and test the properties of the AA map using these crossovers in Chapter 4.

In addition to directly identifying crossovers, this procedure enabled me to impute genotypes in parents with missing data, because they are part of the hidden state vector (Figure 3.1). As discussed in Chapter 5, this increased the available data size for association mapping of recombination-related phenotypes in pedigree parents by 78%, thereby providing significantly greater power in testing.

Chapter 4

Assessing the AA map using families and existing genetic maps

I built the AA map in Chapter 2 using crossovers identified from the switching of ancestry blocks in African American genomes, followed by refinement using an MCMC scheme. This is a fundamentally new approach for building genetic maps, and therefore it needed testing and validation. Secondly, it was unclear at this stage if humans do or do not have systematic differences in their genetic maps. For this reason, an important initial aim was to test the accuracy of the AA map at both fine and broad scales, before using it to make biological inferences.

In this chapter I validate the AA map and demonstrate its accuracy and precision by comparing it with crossovers directly identified in African American pedigrees (Chapter 3) and with other existing resources for studying recombination in humans. These resources are

- Hotspots identified by sperm typing (Section 1.3.1) in two well-studied regions of the human genome. These allow us to gain insights at fine-scales (~ 1 kb).
- A pedigree-based map built in the Icelandic population, who are a European population (Section 1.3.2). Kong et al. [2010] from deCODE genetics published

maps for males and females separately. The map used in the analysis below combines the sex-specific maps to make a sex-averaged map, and is referred to as the deCODE map. This allows us to compare and contrast rates at scales greater than approximately 10 kb.

- LD-based maps estimated from HapMap2 variation data (Details in Section 1.3.3). Three LD-based genetic maps are used in this chapter:

    1. Map estimated using the CEPH samples from Utah, who have European ancestry. This is referred to as the CEU map.

    2. Map estimated using the Yoruba population, who are a West African people. This is referred to as the YRI map. This map allows us to compare our map to one built in a population whose ancestry is more similar to our samples, relative to the European resources listed above (although the comparison will be with crossovers in the distant past).

    3. Map estimated by pooling four populations: CEU, YRI and the Asian populations of JPT (Japanese) and CHB (Han Chinese). This is referred to as the Combined map.

4.1 The AA map finds sperm-typing hotspots

As discussed in Section 1.3.1, sperm typing of targeted genomic regions is the most fine-scaled and unbiased way of identifying crossover hotspots in humans. However, sperm typing does not scale genome-wide as it is labour-intensive and involves individual targeting of each candidate hotspot. I compare the AA map rates in two regions targeted in sperm typing studies. The first region I examine is the Major Histocompatibility Complex (MHC) region on chromosome 6 [Jeffreys et al., 2001]. Figure 4.1 shows the rate estimates for four maps (AA Map, deCODE Map, CEU

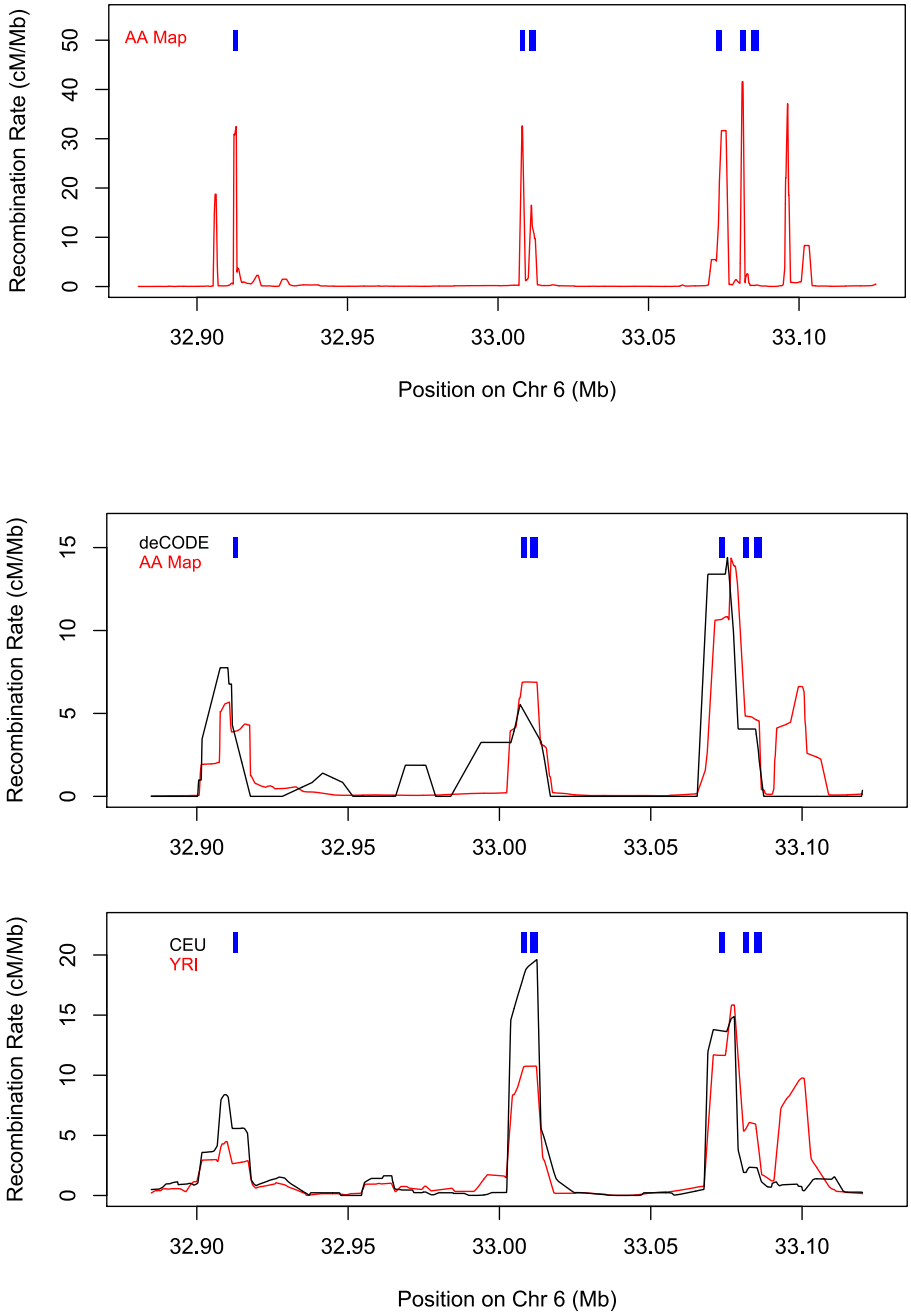


Figure 4.1: Recombination Rates in the HLA region. (Top) The AA Map is able to accurately localize five hotspots to a 200kb region of the human MHC region whose positions (blue marks) were mapped by sperm typing [Jeffreys et al., 2001]. (Bottom Two) Comparison of the same region for several maps (I smooth all maps to 10 kb to allow comparison with the deCODE map, which has lower resolution). One, and possibly two, African-specific hotspots are apparent near position 33.1 Mb; here, both the AA Map and the YRI LD Map show elevated rates, whereas the deCODE Map and CEU LD Map do not.

LD Map, and YRI LD Map) over a 200 kb region within the MHC. In the top panel, the AA Map is plotted at 1 kb resolution and detects five of six known hotspots, and localizes them to within 1 kb (the sixth hotspot is weak, with a peak male recombination rate significantly below the genome average). The bottom two panels show all four maps smoothed to a 10kb scale (in order to match the resolution of the deCODE map). The deCODE Map detects hotspots in the same regions, although one hotspot appears shifted by around 5kb.

The most striking finding is the presence of an intense hotspot between 33.09 Mb and 33.11 Mb that is detected in the AA map but not in the deCODE map. This suggests that it may be a hotspot that is active specifically in Africans, but not in Europeans. This is confirmed by comparing the YRI (African) and CEU (European) maps, wherein a hotspot was identified in the YRI map, but not in the CEU map. I discuss the genome-wide prevalence of such *African-specific hotspots* in Chapter 5.

Another region which has been well studied in sperm typing studies is the highly variable MS32 minisatellite on chromosome 1 [Jeffreys et al., 1998, 2005]. Figure 4.2 shows rates in all four maps in this region. The higher resolution of the AA map, relative to the deCODE map is evident.

4.2 The AA map predicts hotspots at sites of crossover in African-American families

I examine the precision and accuracy of the AA map by assessing the crossover rate it estimates in regions where crossovers have been precisely mapped in African American families in Chapter 3. I identified the start and end points of 3,068 crossover events in pedigrees that were resolved to within 10 kb. I then calculated the mean recombination rate in the AA map 100 kb upstream and downstream of the midpoints of these events. Figure 4.3 confirms the high resolution of the AA map. While the

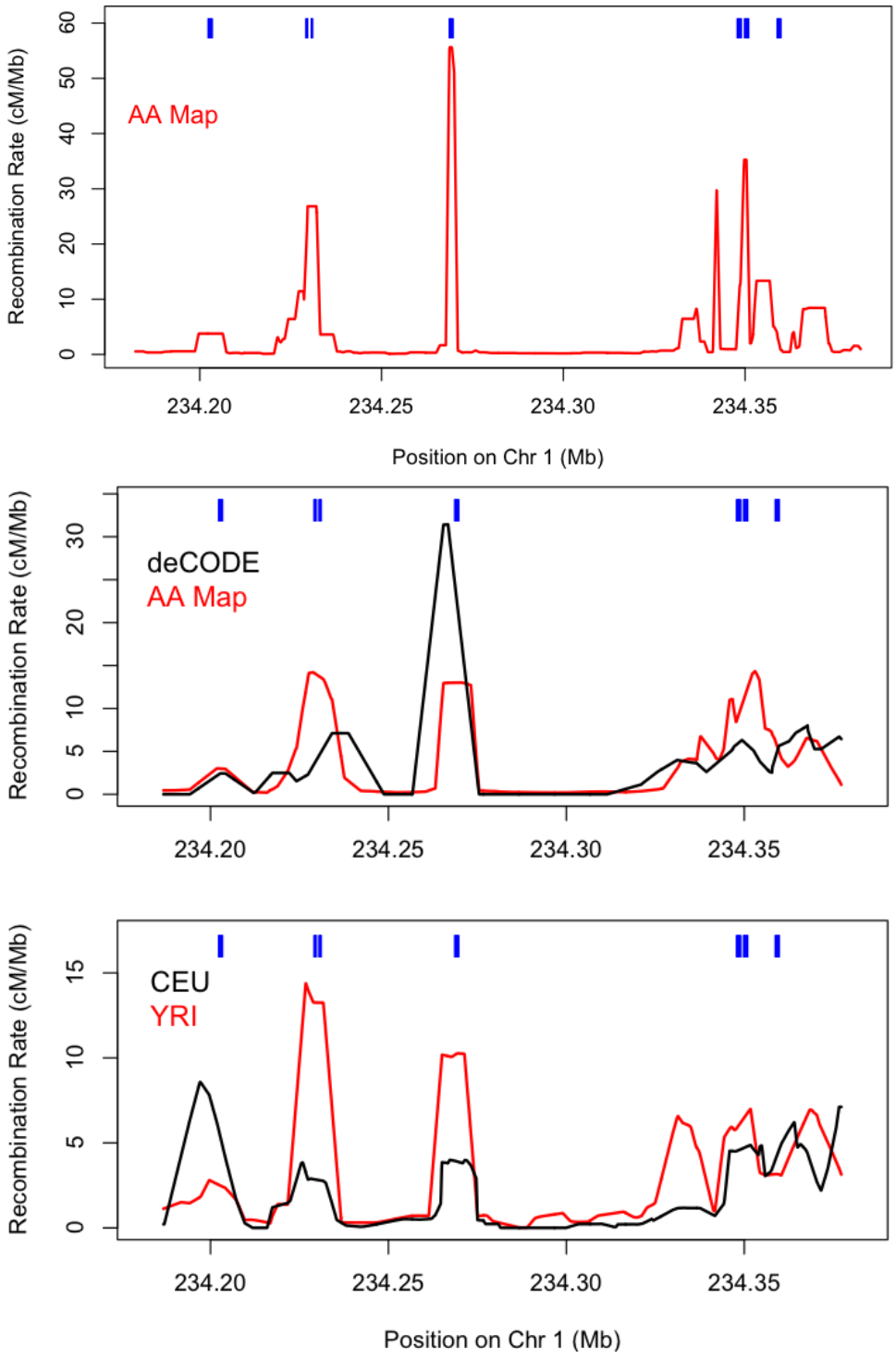


Figure 4.2: Recombination Rates in the MS32 minisatellite region. (Top) The AA Map localizes hotspots in the region whose positions (blue marks) were mapped by sperm typing [Jeffreys et al., 2005]. (Bottom Two) Comparison of the same region for several maps (all maps smoothed to 10kb resolution to facilitate comparison with the deCODE map).

LD-based map has slightly more tightly resolved hotspots, possibly due to higher SNP density, the AA map predicts hotspots with a higher mean rate at the crossover sites observed in African American families.

4.3 The AA map is well-calibrated

Here I assess whether the probability of recombination reported in the AA map accurately reflects the true probability of recombination. Specifically, I wish to examine if the AA map systematically overestimates or underestimates the rate depending on the true recombination rate. Underestimating the rate might represent recombination 'leaking' into nearby intervals and poor resolution. Overestimating the rate might represent overfitting hotspots and under-representation of recombination outside of hotspots. Although we do not know the true rates genome-wide, I check calibration in individual crossover events by asking if the AA map produces appropriately sampled event positions. Since these positions are used to build the map, it is reasonable to infer that the map may be accurate if it results in correct sampling of where events occur.

To do this I use 16 African-American nuclear families, whose genotypes were available for both parents and two or more children. I identify crossovers using pedigree-analysis as described in Chapter 3. Further, I run HAPMIX on each member of the nuclear family and identify *de novo* ancestry switches in the children, as described in Figure 2.8. These events represent a subset of the crossovers in the parents, and match up with events in the pedigree analysis.

Using this procedure, I have two views of 370 crossovers, a pedigree view and an ancestry-switch view. I re-normalize the ancestry-switch PMF with the AA map, as described in Section 2.5, to estimate the PMF of each crossover under an AA map prior. For each crossover, say c_j where $j \in \{1, 2, \ldots, 370\}$, I have two PMFs each, a

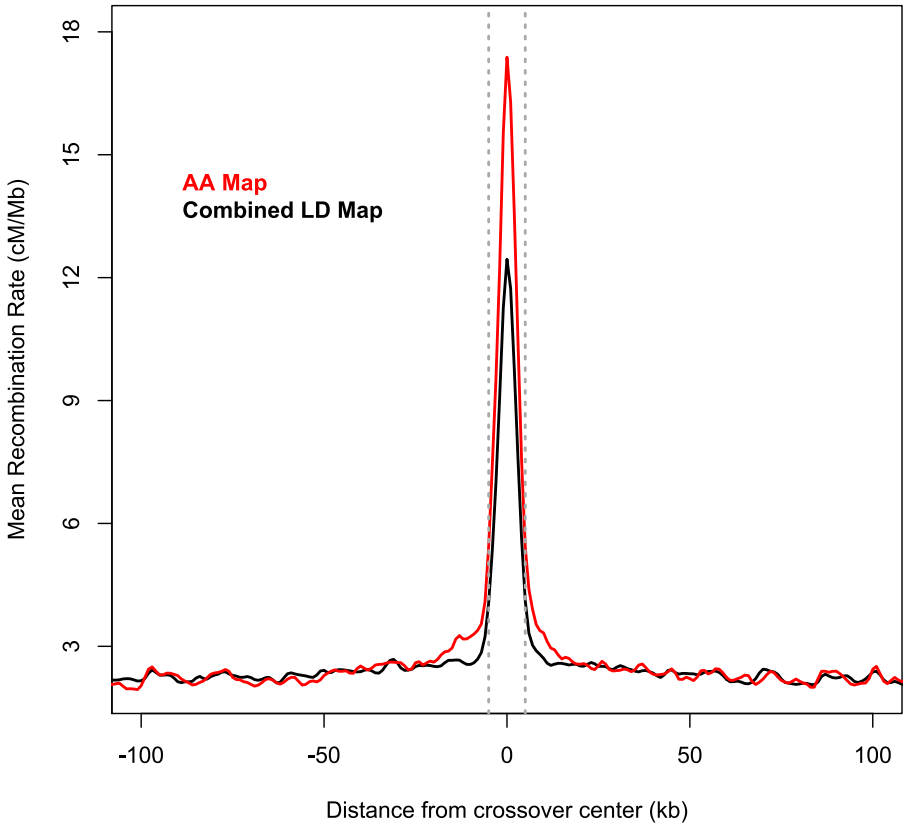


Figure 4.3: Mean recombination rate in the AA Map (red) and the population-averaged HapMap2 LD-based map (black) around 3,068 narrowly defined crossover events directly identified in African American families. The grey dashed vertical lines mark the maximum extent of pedigree events included in this analysis. The AA Map has a resolution similar to the LD-based map, and also has a rate peak at least as high as the LD-based map.

PMF from pedigree analysis, $f_{ped,j}$ and one from the AA map, $f_{AA,j}$.

Figure 4.4 compares f_{ped} and f_{AA}, pooled together for all events, with the hottest f_{AA} intervals on the left and the coldest on the right on the x-axis. The intervals are pooled into 10 bins such that each bin has equal total probability of recombination according to f_{AA}. Encouragingly, f_{ped} in those intervals is essentially constant. In other words, the hottest regions in the AA map, while accounting for much less than 10% of sequence, still contain the estimated 10% of recombination (as do the coldest regions). This shows that f_{AA} is not biased in favour of or against hotspots, up to the resolution of pedigree events of approximately 30-40 kb, so it is well-calibrated in that sense.

It is illuminating to compare this with the basic map. I repeat the procedure above to get f_{basic} for the same crossovers. Figure 4.4 shows that the curve of the basic map slopes downwards. For example, the hottest intervals containing 10% of recombination according to the basic map, should in fact contain approximately 16% of recombination according to pedigrees. This shows that the basic map underestimates the rate in its hottest regions. This implies that it has lower ability to finely resolve events than the AA map, showing that the MCMC procedure in Chapter 2 resulted in clear map improvements.

4.4 Correlation analysis with existing genetic maps

Having observed good results at the finest scales, I now extend our analysis up to broader scales. I compared the correlation of the AA map with the CEU and YRI LD-based maps. Table 4.5 summarizes the Pearson and Spearman (rank-based) correlations at a range of size scales, in comparison to the best available map of present-day recombination to date, from deCODE [Kong et al., 2010].

At fine scales ($\approx$ 3-30 kb) there are significant differences in the patterns of cor-

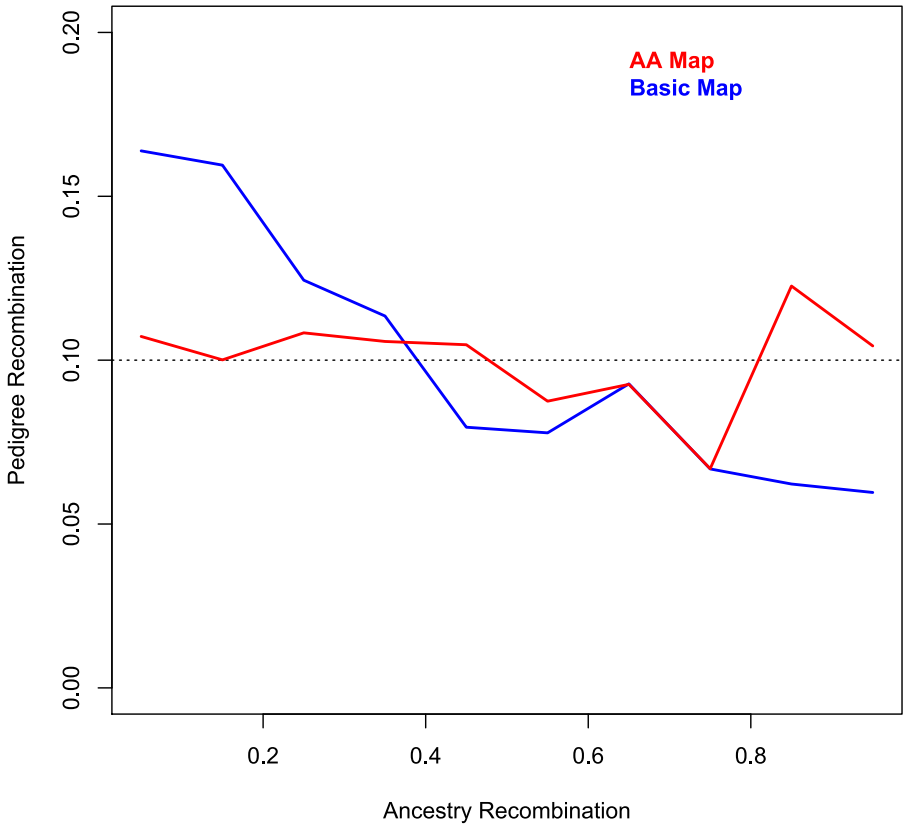


Figure 4.4: Calibrating recombination probabilities estimated by the AA Map using 370 crossover events observed in pedigrees as well as from ancestry painting. The x-axis has 10% bins of recombination probability, with the left-most containing SNP intervals with the highest probability density per base of sequence and the right-most bin having the lowest, according to the AA map. The y-axis shows the corresponding probability observed in the pedigree crossovers. The slope of the curve of the Basic Map shows that it underestimates the recombination rates of hotspots due to over-smoothing, while the flat line for the AA Map suggests accurate estimation.

Scale (interval size)	Pearson correlation of the AA map (deCODE) to the specified LD map		Spearman correlation of the AA map (deCODE) to the specified LD map	
	CEU	YRI	CEU	YRI
3 kb	0.66 (0.58)	0.71 (0.53)	0.64 (0.31)	0.69 (0.32)
10 kb	0.73 (0.70)	0.78 (0.65)	0.72 (0.47)	0.78 (0.48)
30 kb	0.78 (0.78)	0.83 (0.74)	0.83 (0.70)	0.87 (0.69)
100 kb	0.84 (0.85)	0.87 (0.81)	0.90 (0.88)	0.92 (0.86)
300 kb	0.89 (0.90)	0.92 (0.88)	0.92 (0.92)	0.94 (0.91)
1 Mb	0.94 (0.94)	0.95 (0.95)	0.95 (0.95)	0.96 (0.94)
3 Mb	0.97 (0.97)	0.98 (0.97)	0.97 (0.97)	0.97 (0.97)

Figure 4.5: Pearson and Spearman (rank) correlations of the AA and deCODE maps with specified LD-based maps at different scales.

relation. The deCODE Map correlates more strongly with the European ancestry CEU map ($\rho_{\text{deCODE,CEU}} = 0.58$ at 3 kb scale) than it does with the African ancestry YRI map ($\rho_{\text{deCODE,CEU}} = 0.53$), while the AA Map correlates more strongly with the YRI map ($\rho_{\text{AA,YRI}} = 0.71$) than the CEU map ($\rho_{\text{AA,CEU}} = 0.66$). This suggests population-level differences in the patterns of recombination. I find that this is due to significant biological differences in recombination initiation, as shown in Chapter 5. At coarse scales ($\geq$ 3 Mb), both the AA Map and deCODE Map are about equally correlated to CEU and YRI LD-based maps ($\rho \geq 0.97$), indicating that the African and European recombination landscapes are nearly identical at this resolution. This appears to indicate strong broad-scale conservation of rates, and is explored further in Chapter 6.

The AA map correlates more strongly than the deCODE map with the YRI map at all scales. This might be expected, considering the ancestry of the groups used to build these maps. More surprisingly, at very fine scales ($\leq$ 10kb), I find that the AA map map still has a higher correlation than the deCODE map with the CEU

map. This is true despite the fact that both CEU and deCODE maps are built in Europeans, and at broader scales the correlations are similar. This could be due, in part, to the CEU map containing historical recombination events that happened prior to the out-of-Africa split, and which may reflect patterns that continue to exist in contemporary African populations. However, it may also be due to the AA map having fundamentally lower error rate than the deCODE map. Three factors likely contribute. First, there is a much greater number of crossover events in the AA Map (2.1 million vs. 0.5 million), allowing us to identify more subtle hotspots. Second, the AA map leverages a larger number of SNPs (1.2 million vs. 0.3 million), permitting better localization of events. Finally, our algorithm for inferring recombination rates is different. The method used for the deCODE Map used an Expectation Maximization (EM) algorithm for maximum likelihood estimation of recombination rates without the use of prior distributions. Prior distributions naturally smooth rate estimates, and not using a prior distribution might overfit, for example by overestimating the rates at the most active hotspots and therefore underestimating in the more frequent colder regions, or even by misplacing the less active hotspots. These factors predicts an even greater improvement in the correlation of the CEU map with the AA Map relative to the deCODE Map when assessed using Spearman rank correlations. This is because rank correlation is less dominated by the hottest hotspots. This prediction is indeed borne out: $\rho^{Spearman}_{\text{AA,CEU}} = 0.64$ vs. $\rho^{Spearman}_{\text{deCODE,CEU}} = 0.31$. Table 4.5 summarizes the Spearman correlations at a range of size scales. Comparing the Pearson correlations when the rates are measured in log-scale shows a similar pattern, as that also reduces the impact of the hottest hotspots.

4.5 Concentration of recombination in hotspots in different human populations

The extent to which recombination is concentrated into short segments of the genome can be visualized by so-called "80-20" plots. In these plots, the fraction of recombination explained genome-wide is shown as a function of the fraction of genomic sequence included in the analysis going from the hottest sequences to the coldest. Figure 4.6 plots this for the AA, deCODE, YRI and CEU maps. All maps are smoothed to 30kb, which is chosen because this scale is coarser than the resolution of all four maps. I do this so that I am not comparing the extent of noise among the maps in the ability to define hotspots, but rather the underlying differences in rates. Typically, in Europeans, a 30 kb region will generally have either none or one hotspot (or less often, two hotspots) [Jeffreys et al., 2004, 2005]. Thus, this scale is expected to be sufficiently fine to reveal general differences in hotspot frequency and/or intensity between the populations.

Examining Figure 4.6, I note that AA and YRI maps reveal comparable concentration of recombination in hotspots, while both CEU and deCODE maps are more concentrated. This suggests African ancestry populations have more of their genome active for crossovers, consistent with them having more hotspots than Europeans. This shows that the shorter extent of LD in Africans, which is the flip-side of the number of LD-based hotspots, is not due to differences in demographic history alone but instead also reflects differences which influence contemporary populations.

4.6 Discussion

In this Chapter, I sought to test the accuracy of the novel map building procedure utilized to build the AA map, and to assess the resolution of the hotspots identified

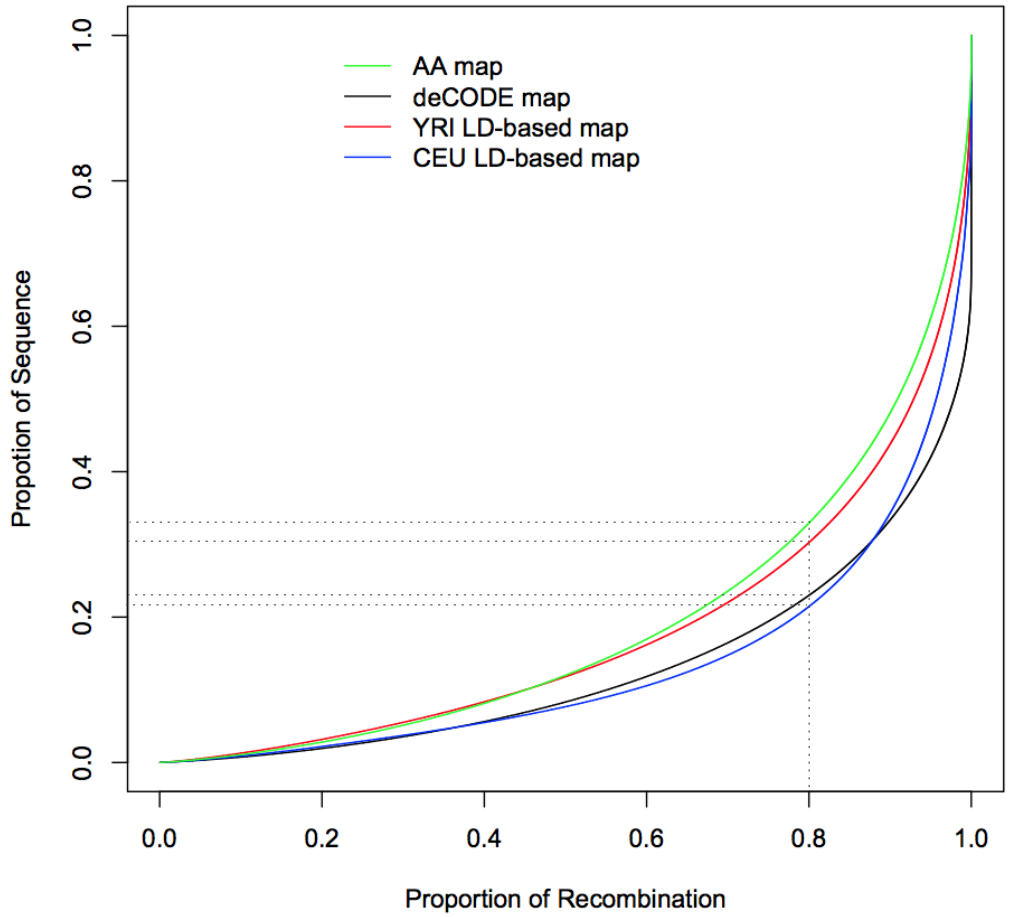


Figure 4.6: Proportion of genomic sequence needed to account for increasing fractions of recombination, with the hottest sequences included first. For example, this plot shows that 80% of recombination in the deCODE map happens in 23% of the genome, while in the AA map, it requires about 32% of the genome (dotted lines). Concentration of recombination in the AA and YRI maps suggests similar density of hotspots, while the deCODE and CEU maps show crossover is significantly more concentrated. This suggests a lower density of hotspots in the genome in the European ancestry maps relative to the African ancestry maps.

therein. I showed that the AA map recapitulates all hotspots which have rate estimated to be greater than the genome-wide average rate, in two regions of the genome well characterized by sperm typing. I also showed that the AA map predicts hotspots in regions where African Americans have crossovers identified from pedigree analyses, and that the map is well calibrated. Comparison of the two regions above, as well as correlation analysis, showed that the AA map has greater accuracy than the most advanced pedigree-based map available to date. These analyses enable me to use the AA map with confidence in examining the biological drivers of recombination in the next chapter.

This chapter also revealed an initial biological insight: comparison of hotspots between different maps in the HLA region revealed a hotspot that appears to be active only in populations with African ancestry. Correlation analysis suggested that the AA map is more similar to the African ancestry LD-based map than the European one, while the reverse was true for the deCODE map. Comparison of "80/20" plots also indicated that the density of hotspots in the two European ancestry maps was similar, which was less than the two African ancestry maps. I explore these differences further, and pursue their causes in Chapter 5.

Chapter 5

The biological basis of differences in human recombination landscapes

In Chapter 2, I built a map of crossovers that have occurred in the last few generations in African Americans, and the results of Chapter 4 show that this map is accurate and that it has high resolution. Multiple lines of evidence suggested, in Chapter 4, that African Americans often have crossovers in places different from most Europeans. In this chapter, I aim to confirm if this is true genome-wide, and, if so, to identify hotspots that are active in one population but not in the other. I then search for the biological factors behind these differences.

I start with a description of known details about positioning of hotspots in humans and other mammals, as well as factors influencing genome-wide rates, in Section 5.1. In Section 5.2, I report my findings that thousands of hotspots genome-wide are active in Africans, with much less or no activity in Europeans, while the converse is not true. To identify what factors play a role in creating these differences, I develop a genetic association testing approach that we can apply to both unrelated individuals

and those in pedigrees, in Section 5.3. I perform follow-up studies on these findings (some of the follow-up studies were also performed by my collaborators based on my results) to understand the association signals. I present, in Sections 5.4 and 5.5, how particular DNA features mark the newly found hotspots, and how variation in a single gene, *PRDM9*, explains nearly all of the individual differences in their use. (The work presented in this chapter was done by me unless I explicitly state otherwise at the start of a section).

5.1 Known factors influencing recombination in humans and other mammals

Major breakthroughs have recently occurred in our understanding of how hotspot locations are specified. An epigenetic modification commonly associated with open chromatin and transcriptionally active regions, H3K4me3 (the tri-methylation of histone H3 on lysine 4), was shown to be an important and pre-existing mark for the initiation of recombination in yeast [Borde et al., 2009] and mice [Buard et al., 2009; Smagulova et al., 2011]. Further, it was found in mice that differences in an approximately 5 Mb long region containing the gene *Prdm9* led to a difference in recombination patterns among strains [Grey et al., 2009; Parvanov et al., 2009]. *Prdm9* fit the bill for being the causal gene, being a gene transcribed specifically during meiosis [Hayashi et al., 2005], and containing a SET domain that trimethylates H3K4 (Figure 5.1). In addition, *PRDM9* contains a highly conserved KRAB domain of unknown function and an array of DNA-binding C2H2 zinc fingers.

Myers et al. [2005] investigated whether particular repeat features or simple DNA motif sequences are associated with LD-based hotspots in humans. They created hotspot and coldspot sets matched for size, GC content, SNP density, and other features and looked for enrichment of a variety of features in each set relative to the

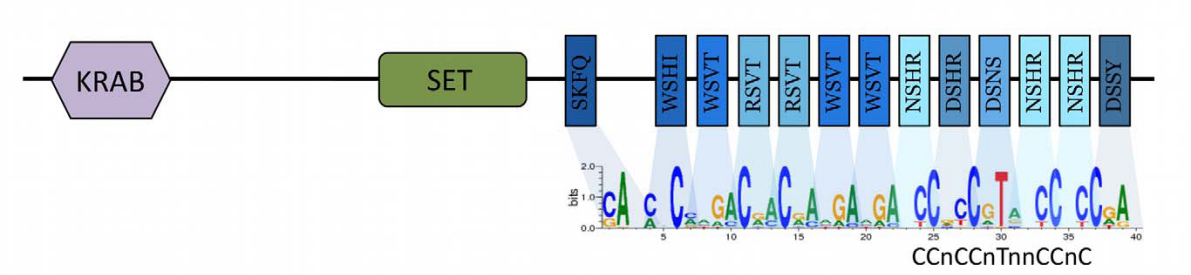


Figure 5.1: The three domains of PRDM9 and the predicted binding sequence of the most common human allele, the A allele. The KRAB domain is thought to be a transcriptional repressor. The SET domain is also highly conserved and trimethylates H3K4, a mark associated with recombination in mice and yeast [Huang, 1998; Dillon et al., 2005]. The zinc finger domain is exceptionally variable within and across species and the DNA binding residues of several zinc fingers show signs of multiple bouts of positive selection across diverse taxa. The predicted binding sequence of the A allele is shown below the zinc finger domain, aligned with the 13-mer Myers motif CCnCCnTnnCCnC found to be enriched in human hotspots (Figure from Ségurel et al. [2011]).

other. The long terminal repeats of two retrovirus-like retrotransposons, THE1A and THE1B were strongly overrepresented in hotspots. In addition to differential behavior of certain dinucleotide repeats in hotspots and coldspots, they identified a 7-nucleotide motif CCTCCCT as being strongly enriched in hotspots. Direct experimental evidence for the role of this motif came from Jeffreys and Neumann [2002], who found that a single nucleotide variant in this motif reduced crossover initiation at the DNA2 hotspot in humans by more than three-fold. Myers et al. [2008] refined their analysis using hotspots identified from HapMap Phase 2 data [International HapMap Consortium, 2007] (the hotspots were described in Section 1.4.4) and showed that the degenerate 13-mer DNA sequence motif CCnCCnTnnCCnC[1] recruited crossovers in at least 40% of human hotspots (this 13-mer will be referred to as the *Myers motif* henceforth). Further, they predicted that this motif might be bound by a zinc finger protein, based on an extended sequence context influencing hotspot probabilities. The Myers motif was observed to have different penetrance in different genomic

[1]In the motif CCnCCnTnnCCnC, n can be any base A/T/C/G, and is referred to as a *degenerate* base.

backgrounds, with decreasing likelihood of hotspots when found, in order, in THE1, L2 and Alu repeat elements. A further motif CCCCACCCC was also found to be enriched in hotspots.

The dots were connected in humans in 2010 with the demonstration that (a) a common *PRDM9* variant, called the A allele, is the only zinc finger gene in the human genome predicted *in silico* to bind the Myers motif in a manner consistent with the observed degeneracy [Myers et al., 2010], (b) the *PRDM9* A allele does indeed bind the motif *in vitro* [Baudat et al., 2010], (c) individuals with a particular rare variant of *PRDM9*, called the *I* allele, have statistically different recombination in LD-based Myers hotspots (Baudat et al. [2010]), and (d) the motif does not appear to operate in chimpanzees, and chimpanzees have different *PRDM9* allele(s), and finally (e) the Myers motif has been lost from the human genome at an accelerated rate relative to chimpanzees [Myers et al., 2010], a situation which could be explained by ongoing biased gene conversion (Section 1.4.5).

Similarly, in mice, the 5Mb region previously implicated in recombination control was refined to a 180 kb region containing *Prdm9* [Parvanov et al., 2010] and a consensus motif enriched in hotspots was later found to match the binding prediction of *Prdm9* in two distinct mouse strains [Smagulova et al., 2011]. Direct and highly specific binding of Prdm9 with DNA at the centre of several mouse hotspots was shown in vitro by Grey et al. [2011], who further demonstrated a change in hotspot activity, H3K4me3 levels and chromosome-wide distribution of crossovers by the sole modification of Prdm9 zinc fingers in transgenic mice. *Prdm9* variation has also recently been associated with genome-wide patterns of hotspot activity in cattle [Sandor et al., 2012].

In humans, variants of *PRDM9* were found to control the activity of a dozen sperm-typing hotspots from across the genome [Berg et al., 2010] and polymorphism near the *PRDM9* gene was associated with the degree of recombination occurring

within LD-based hotspots in Icelandic pedigrees [Kong et al., 2010]. At this stage, several mysteries remained (some continue to be so):

- *The fraction of hotspots controlled by PRDM9.* Myers et al. [2008] estimated that the Myers motif played a causal role in at least 40% of human hotspots. Another study [Baudat et al., 2010] suggests that the number may be higher. This is based on the observation that individuals with two copies of the *A* allele of *PRDM9* have 60% of their crossovers in LD-based hotspots, while those carrying a single copy of the rare *I* allele of *PRDM9* (i.e, AI heterozygotes) have only 18% of their crossovers in those hotspots. If the *A* allele motif was activating only 40% of the hotspots, we would expect a reduction not greater than 40% in the fraction of crossovers in hotspots, even if the I allele is dominant and activates no LD-based hotspots. The observed reduction of 70% suggests that PRDM9 may be specifying a higher fraction of hotspots. Sperm-typing data [Berg et al., 2010] suggests PRDM9 may control nearly all hotspots, though the sample size in that study is modest.

- *Relationship between PRDM9 allelic types and the DNA sequences they bind.* *PRDM9* is highly diverse in humans with about two dozen alleles sequenced, which contain differences in both the number and type of the zinc fingers (Berg et al. [2010, 2011]; Figure 5.2). They can be categorized into classes based on their respective binding motifs predicted computationally using algorithms that predicts the DNA binding specificity for C2H2 zinc fingers [Persikov et al., 2009]. The biggest classes in humans are the *A*-like class (matching all 8 non-degenerate bases of the Myers motif) and the *C*-like class (matching 5 non-degenerate bases of the Myers motif). These classes have different allele frequencies in different populations, with most Europeans carrying A-like alleles and greater variability among Africans (Figure 5.3). Despite having similar predicted consensus motif sequences, alleles *within* a class have sometimes been observed to lead to

differential activities of some hotspots [Berg et al., 2010; Baudat et al., 2010]. An important question is to understand the impact of this variation on the recombination landscape of an individual. Equally, it is interesting to understand whether *PRDM9* variants from different classes share hotspots.

Sperm typing studies have shown that single mutations in the 13-mer within hotspots can abolish the activity of a hotspot, which is consistent with the activity of a highly specific and sensitive motif [Jeffreys et al., 2001; Jeffreys and Neumann, 2005]. A puzzling observation, however, from a more recent sperm-typing experiment [Berg et al., 2010] is that the activity of 5 intense recombination hotspots were all influenced by the *PRDM9* genotype, even when the hotspots did not contain an obvious match to the motif, and with even subtle changes in the zinc finger array abolishing crossover at these hotspots. Despite the absence of a clear motif, several of these hotspots were shown to bind *PRDM9 in vitro* via highly degenerate matches to the motif [Nudrat Noor, personal communication]. In general, changes in a zinc finger array can influence affinity as well as cooperation between zinc fingers, and interactions between zinc fingers are not yet well understood [Ramirez et al., 2008].

- *Penetrance of the Myers motif and the role of its genomic context.* As noted above, the Myers motif is not necessary to initiate hotspot activity in humans. It is not sufficient either. It occurs approximately 300,000 times in the genome, yet the number of hotspots is approximately a tenth of that number. Epigenetic context may play a part, including the presence of nucleosomes [Smagulova et al., 2011]. Other possibilities include additional zinc fingers influencing the stability of binding and possible unknown co-factor proteins. The genomic context also plays a role, for reasons that are not understood. Specifically, the motif has far greater penetrance in certain repeat regions – THE1A/B and L2, than it does on average in non-repeat regions [Myers et al., 2008].

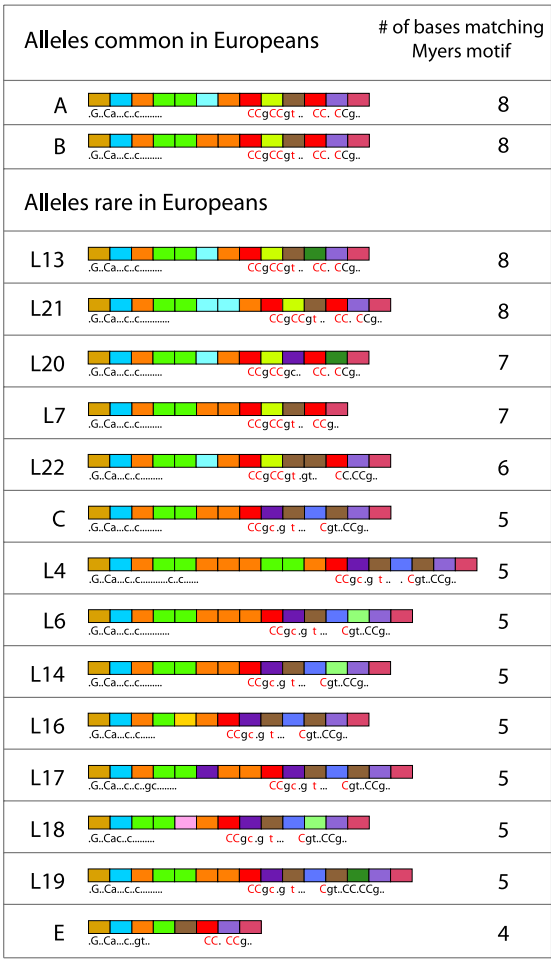


Figure 5.2: A subset of *PRDM9* variants found in human populations. Each rectangle represents a zinc finger (ZF) in the array, with different colours representing distinct ZFs. The predicted binding motifs are shown underneath the array. The predicted motifs, aligned to the Myers motif, are based on the final six ZFs. For instance, the A and B alleles each have 13 ZFs, and have the same predicted motif as they differ only in one upstream zinc finger. (Figure adapted from Berg et al. [2010]).

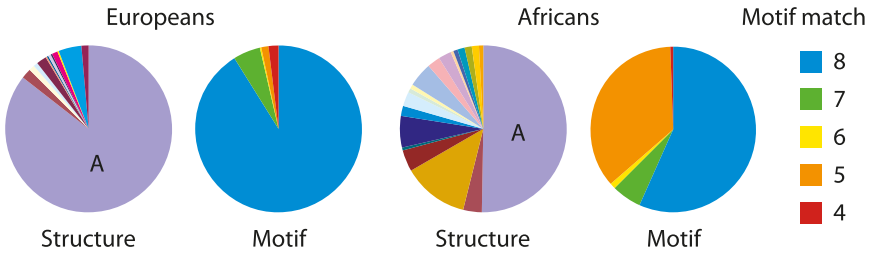


Figure 5.3: The diversity of *PRDM9* alleles in African populations is significantly greater than that in Europeans. 88% of European alleles have a predicted 8/8 match to the motif, while in Africans only 55% do so. 33% of the African alleles are predicted to match only 5 bases of the Myers motif. (Figure from Berg et al. [2010]).

- *Interaction of PRDM9 variants with each other.* Comparison between recombination patterns of individuals containing *AI* PRDM9 variants as opposed to the *AA* homozygotes suggests that the *I* allele may be dominant over the *A* allele [Baudat et al., 2010]. It is not clear how this dominance is mediated biologically, though recent data in mice suggests that it may be due to intrinsically different binding affinities of Prdm9 variants for DNA [Brick et al., 2012]. Ségurel et al. [2011] reanalyze hotspot activity in sperm typing studies and infer partial dominance of the C-type allele over the A allele, but this is far from clear. The dominance is at best much weaker, with clear co-dominance signals.

  An intriguing aspect of this problem is that *Prdm9* is the only known speciation gene in mammals, with certain *Prdm9* heterozygous mice failing to form sex bodies[2] or repair DSBs, resulting in meiotic arrest and hybrid male sterility [Mihola et al., 2009]. In sterile male hybrids, PRDM9 interacts, among other genes, with a locus that contains itself (though it has not been determined if it is in fact interacting with itself, or a nearby functional element). Interaction between particular PRDM9 variants, thus, potentially plays a role in the hybrid sterility phenotype. Hybrid male sterility is discussed further in Chapter 6 in the context of recombination in the PAR.

- *Conservation and evolution of the PRDM9 gene. PRDM9* acts early in the recombination process and clearly influences the location of programmed DSBs [Smagulova et al., 2011; Brick et al., 2012]. However, how it influences Spo11 behaviour mechanistically is unknown. The KRAB and SET domains of PRDM9 are highly conserved across taxa from snails to humans [Ponting, 2011], suggesting a critical role for the gene. Mice with PRDM9 loss-of-function alleles are infertile – they show meiotic arrest in both males and females, impairment of

[2]The sex body is the cytologically visible change in the unsynapsed portions of the X and Y chromosomes during Prophase I in male meioses. These regions undergo transcriptional silencing and formation of heterochromatin.

DSB repair, chromosome synapsis and disrupted sex body formation [Hayashi et al., 2005]

In contrast, residues of PRDM9 in contact with DNA during binding, and thus controlling binding sites, show an exceptionally high rate of change in mice, primates, and other taxa [Oliver et al., 2009], with strong evidence for positive selection at those bases. This suggests that selection has occurred for novel binding targets multiple times independently across diverse taxa. Several theories for this have been proposed, among them: creating new hotspots to replace those lost due to biased gene conversion (Section 1.4.5) in order to ensure sufficient crossovers for proper segregation [Myers et al., 2010], and the evolution of centromeric repeats [Oliver et al., 2009].

Nevertheless, the importance of the PRDM9 gene remains unclear. Several taxa lack PRDM9 altogether, such as non-metazoan eukaryotes, fruit fly, birds, lizards and reptiles [Oliver et al., 2009]. Even a mammal, dog, appears to have lost the use of PRDM9 [Axelsson et al., 2012]. This leads to the apparently paradoxical conclusion of both a critical and non-essential role for *PRDM9*.

The impact of genetic variation on other aspects of recombination, such as the number of crossovers per meiosis, has also been studied. Baudat et al. [2010] report an association between *PRDM9* allele B and genome-wide rate. They find that a combined sample of males and females, who are heterozygous with one A and one B *PRDM9* allele, have a significantly longer genetic map length than those homozygous for the A allele.

Variants in *RNF212*, the mammalian ortholog of *Zip3* (Section 1.2.4), have been found to be associated with recombination rate in humans in both males [Kong et al., 2008; Fledel-Alon et al., 2011] and females [Kong et al., 2008] and also in male cattle [Sandor et al., 2012]. Curiously, the haplotype associated with high male crossover rate in humans is found to associate with a decrease in rate in females [Kong et al.,

2008], though this signal has not been replicated. *RNF212/Zip3* is required for correct synapsis in *S. cerevisiae* and appears to specifically inhibit SC initiation between centromeres in both *S. cerevisiae* and mice [Qiao et al., 2012].

Another confirmed human association for genome-wide crossover rate is with inversion 17q21.31 [Stefansson et al., 2005; Kong et al., 2008; Fledel-Alon et al., 2011] in females. This locus appears to be under selection in Europeans, possibly due to the effect of the inversion on the sequence of one gene *KANSL1* [Boettger et al., 2012]. Kong et al. [2010] also find evidence for a very modest effect of this inversion on the fraction of crossovers in LD-based hotspots.

Finally, genetic variants in *Rec8,* a component of the meiosis-specific cohesin complex (Section 1.2.4), are associated with crossover rate in male cattle [Sandor et al., 2012].

To illuminate some of the biological processes underlying human crossover location and number, I analyzed the differences in human populations I identified in earlier chapters. In the rest of this chapter, I discuss our findings on differences in hotspot locations, and our work in understanding the causes of those differences.

5.2 Population differences in hotspot locations

In Chapter 4, I noted using correlation analysis that the AA map is more similar to the YRI LD-based map than it is to the CEU LD-based map. Conversely, the deCODE map, built in the Icelandic population, is more correlated with the CEU map than with the YRI map. This suggests that recombination patterns may differ consistently between the European and African populations.

I also noted in Figure 4.6 that the landscape of recombination was dominated by hotspots to a comparable degree in the AA and YRI maps. The same fraction of recombination was concentrated in a smaller portion of the genome in the European

maps of CEU and deCODE.

In a detailed analysis of the recombination hotspots in the Major Histocompatibility Complex (MHC) region illustrated in Figure 4.1, I found that there is a strong hotspot that is shared by the maps built in individuals with African ancestry (AA map and YRI map), while being absent in both European derived maps (deCODE and CEU maps). Another example of an intense African-specific hotspot is shown in Figure 5.4. To assess whether the existence of such *African-specific* hotspots is a general phenomenon, I identified 2,375 loci with recombination rate peaks in the YRI map ($\geq$ 5 cM/Mb) but not the CEU map ($<$ 1 cM/Mb). If these are genuine African-specific hotspots, we would expect to see a rate increase in the AA map, but not in the deCODE map. We find that to be the case, as shown in Figure 5.5. In the reciprocal experiment, I search for candidate rate peaks in the CEU map ($\geq$ 5 cM/Mb) that are absent in the YRI map ($<$ 1 cM/Mb). We find only 1,263 such locations, and they show weak rate peaks in both deCODE and AA maps (Figure 5.6). This analysis provides no support for European specific hotspots. Instead, it suggests that the rate peaks are weak hotspots in both African and European populations, and that the differing rate estimates in the LD-based CEU and YRI maps are likely due to statistical noise in the estimation of weak hotspots. Thus, hotspots active in Europeans are consistently shared with YRI and African Americans, which we refer to as *shared hotspots,* whereas populations with African ancestry have additional, non-shared hotspots that we call *African-specific hotspots.* In the next section, I search for genetic variants that may be causing differences in the activity of African-specific versus shared hotspots.

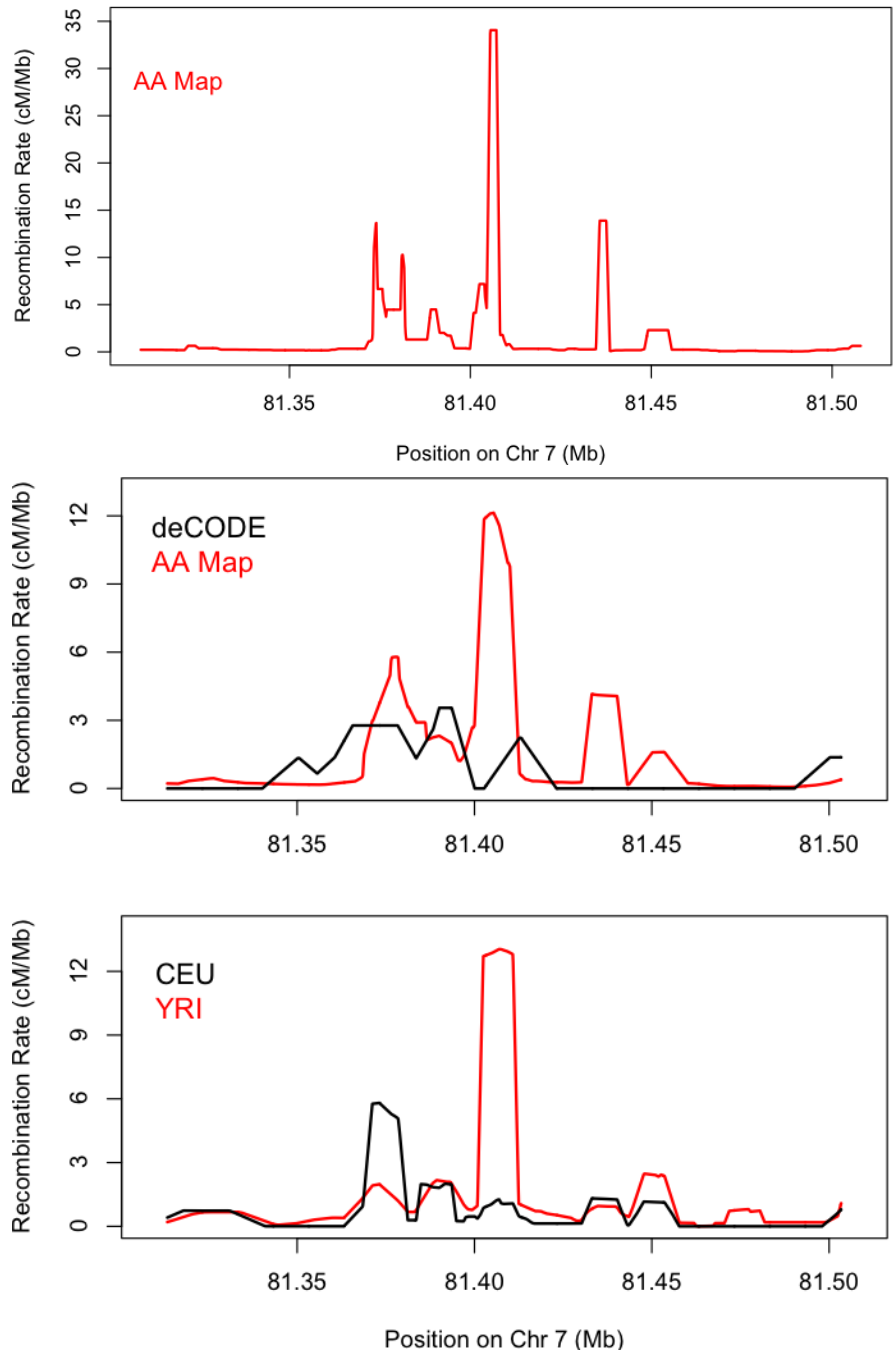


Figure 5.4: Recombination Rates in a 200 kb region on chromosome 7. This region presents another example of an African-specific hotspot that is inferred by both the AA and YRI Maps. This site is cold with a low recombination rate in both the deCODE and CEU maps.

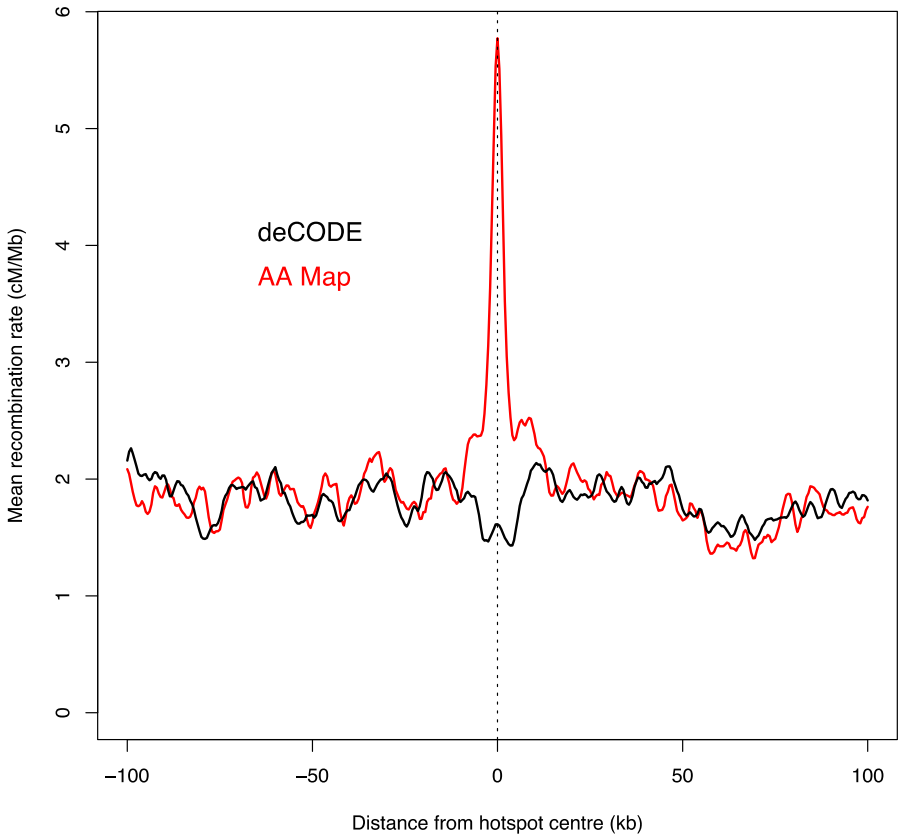


Figure 5.5: Evidence for hotspots active in African populations but inactive in Europeans. This plot shows mean estimated rates in the AA Map (red) and the deCODE Map (black) surrounding 2,375 regions with high rates in the YRI Map ($\geq$ 5 cM/Mb over 2kb) but not the CEU Map (<1 cM/Mb). The maps are plotted in 2 kb bins with a 500bp sliding window. The peak in the AA Map only (with no signal in the deCODE Map) demonstrates the presence of genuine African-enriched hotspots in individuals with African ancestry.

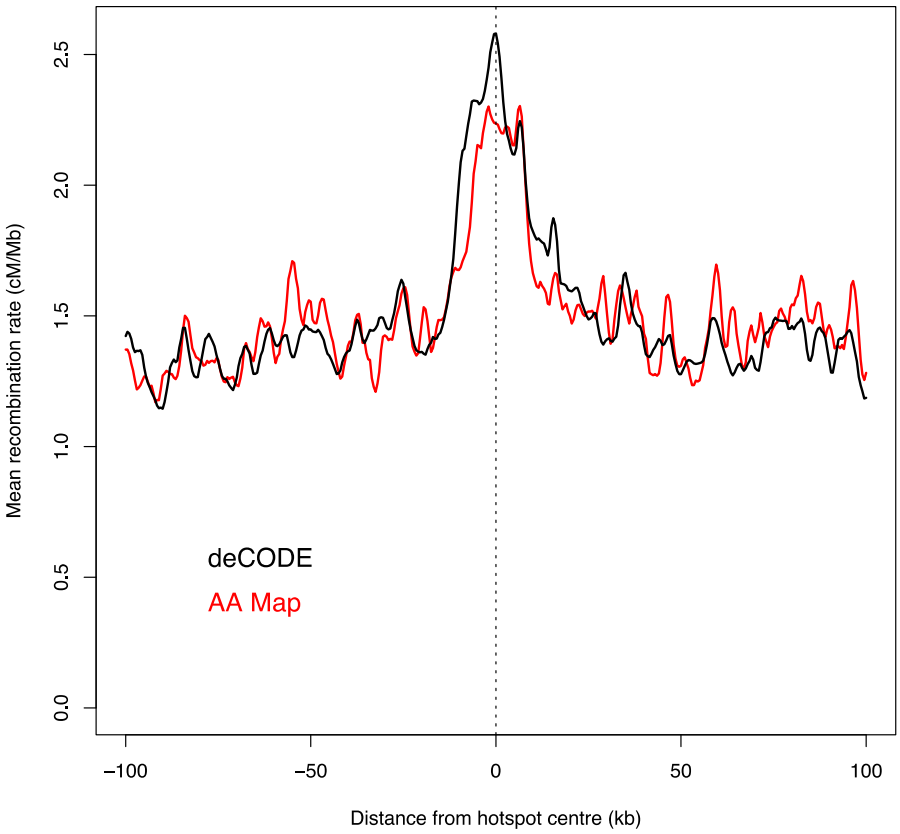


Figure 5.6: No evidence for hotspots specific to European populations. This plot shows mean estimated rates in the AA Map (red) and the deCODE Map (black) surrounding 1,263 regions with high rates in the CEU Map ($\geq$ 5 cM/Mb over 2kb) but not the YRI Map (<1 cM/Mb). The maps are plotted in 2 kb bins with a 500bp sliding window. Both AA map and deCODE map contain weak hotspots in these regions, showing the lack of ancestry specificity of hotspots active in Europeans.

5.3 Mapping variants underlying the use of African-specific hotspots

We inferred above that African Americans harbour two types of recombination hotspots: *shared hotspots* that are in common with Europeans, and *African-specific hotspots* that appear to have only been active in African history. To better understand these differences, I estimate the degree to which crossover events detected in an African-American individual differ in their distribution from what is expected in people of European descent. Motivated by the pattern of two classes of hotspots, I modelled the crossover events of each African American as arising from a mixture of a *Shared genetic map* and an *African-specific genetic map*. I estimate both maps in this section, and utilize them to infer a phenotype for each individual that corresponds to his or her mixing proportion of the two maps. We refer to that proportion of an individual's crossovers that originate from the African-specific map as their *African-enrichment (AE)* phenotype. I start with describing how I estimate Shared and African-specific maps.

5.3.1 Constructing Shared and African-specific genetic maps

I extend the Gibbs Sampler from Chapter 2 to estimate Shared and African-specific genetic maps, and the African-enrichment of each individual. Since there is no evidence of hotspots that are specific to the European population, we expect that the Shared map is very similar to the European map. The basic idea then is to model each individual as a mixture of two maps, one of which is the European map and the other is, in some sense, maximally different from it. I initialise both maps to reasonable starting points, such as rates sampled from the basic map. Next, I populate the Shared map with events from the deCODE map. I then sample each African American crossover as arising from either the Shared map or the African-specific

map. Intuitively, since we fix the total genetic map length of both maps to be the same, crossovers which are not likely in the Shared map will be sampled more often to the African-specific map. This will, therefore, nudge the Gibbs Sampler towards producing two distinct maps.

In Chapter 2, the Gibbs sampler iterated over sampling rates and crossover locations. In this Gibbs sampler, I additionally sample each crossover to one of the two maps, and also sample the African-enrichment of each individual based on how African-specific their crossovers appear to be. The algorithm is described in detail below.

The notation used is as follows: let the sequence of inter-SNP intervals be numbered $(1, 2, \ldots, M)$, where M is the total number of intervals. The Shared map is denoted $\mathcal{S}$ with crossover rates per interval $(r_1^{\mathcal{S}}, r_2^{\mathcal{S}}, \ldots, r_M^{\mathcal{S}})$ and the African-specific map $\mathcal{A}$ with rates $(r_1^{\mathcal{A}}, r_2^{\mathcal{A}}, \ldots, r_M^{\mathcal{A}})$. Let n_k refer to the number of ancestry switches detected in individual Υ_k, and N to the total number of crossovers from all individuals. Therefore, $N = \sum_k n_k$. I also count the aggregate number of crossovers across all individuals per inter-SNP interval, denoted N_i for interval i. Therefore, it is also true that $N = \sum_i N_i$. When referring to the number of crossovers sampled in specific maps, I denote them $\vec{N}^{\mathcal{S}} = (N_1^{\mathcal{S}}, N_2^{\mathcal{S}}, \ldots, N_M^{\mathcal{S}})$ and $\vec{N}^{\mathcal{A}} = (N_1^{\mathcal{A}}, N_2^{\mathcal{A}}, \ldots, N_M^{\mathcal{A}})$, for the Shared and African-specific maps respectively. Finally, I denote the African-enrichment (AE) of each individual Υ_k as ϕ_k.

- **Initialisation of the MCMC.** The Gibbs sampler is initialised with two maps, Shared map $\mathcal{S}$ and African-specific map $\mathcal{A}$, both of which have recombination rates sampled from a Dirichlet prior distribution. We use a Dirichlet prior here, as opposed to the gamma in Chapter 2, to force the total genome-wide rate in the two maps to be the same. This makes both sets of rates to sum to 1, so the prior represents the fraction of all events in each interval. Thus, the maps produced by this procedure are relative and not absolute maps. The prior

distribution is defined such that mean rate in each interval is proportional to the rate in the corresponding interval of the basic genetic map. Note that this prior distribution also has the property of being invariant to the particular choice of SNP set (details in Section 2.5). I denote prior mean rates as $(\lambda_1, \lambda_2, \ldots, \lambda_M)$.

I initialise the African-enrichment ϕ_k for individual Υ_k by sampling from a uniform $U(0, 1)$ distribution. Another possibility for a prior for ϕ may have be one that is informed by the proportion of African ancestry genetically, as inferred by HAPMIX (denoted θ in Chapter 2). However, since the link between 'Africanness' of recombination and the 'Africanness' of the genome is not known at this stage, a non-informative prior is chosen.

Ten chains, each with independently sampled starting maps and African-enrichment values, were initialised.

- **Running of the MCMC:** Five steps are performed in each iteration of the MCMC as detailed below. At least 70,000 iterations were performed for each chain.

    1. *Initializing the iteration.* At the start of an iteration, we have sampled rates for the $\mathcal{S}$ and $\mathcal{A}$ maps available from the previous iteration. I initialize the number of events sampled to each map to be 0 for all intervals, i.e., $\forall_i N_i^{\mathcal{S}} = 0$ and $\forall_i N_i^{\mathcal{A}} = 0$.

    2. *Sampling European crossovers learned from the deCODE sex-averaged genetic map.* We do not have knowledge of the start and end points of crossovers identified by deCODE. I therefore use the sex-averaged map and multiply it by the number of meioses the map represents, to get the approximate numbers of crossovers per inter-SNP interval. The intention is to simply add these events to $\vec{N}^{\mathcal{S}}$, the set of events sampled to the Shared map.

A subtlety is that the deCODE map is estimated across 300,000 intervals, while in this work we use 1.2 million intervals. Instead of simply linearly interpolating the deCODE map to our intervals, we *sample* events within a deCODE interval to the set of the intervals corresponding to it in our study. This allows us to leverage the new data to improve the resolution of the deCODE map. Let the number of crossovers thus obtained from the deCODE map in every interval be $\vec{N}^{\mathcal{D}}$. I add them to the set of crossovers for the Shared map, i.e., $\forall_i N_i^{\mathcal{S}} = N_i^{\mathcal{D}}$.

3. *Sampling African-American ancestry-switch crossovers.* For each individual Υ_k, ϕ_k is the probability of a crossover arising from the African-specific map. Therefore, for a crossover c detected in Υ_k, I set the prior probability of coming from $\mathcal{A}$ as simply his or her ϕ_k sampled in the previous iteration. The prior probability of coming from $\mathcal{S}$ is correspondingly $1 - \phi_k$.

   In Chapter 2, I derived the likelihood of a crossover conditional on a given map, using HAPMIX output. I showed in equation 2.7 that the likelihood of a crossover c conditional on a map is

$$P(c|\mathcal{S}) \propto \sum_i \frac{r_i^{\mathcal{S}}}{r_{flat_i}} H(c,i) \tag{5.1}$$

$$P(c|\mathcal{A}) \propto \sum_i \frac{r_i^{\mathcal{A}}}{r_{flat_i}} H(c,i) \tag{5.2}$$

   where $H(c,i) \equiv P(\rho = i|c, \bar{r}_{flat})$ is obtained from HAPMIX output as described in Section 2.4.3. Note that the constant of proportionality is $P(c|r_{flat})$, which is the same for both maps and can be ignored.

   We therefore obtain the posterior probability of the crossover, up to a

constant of proportionality

$$\begin{aligned} P(c|\mathcal{S}, \Upsilon_k) &\propto (1-\phi_k) \cdot \sum_i \frac{r_i^{\mathcal{S}}}{r_{flat_i}} H(c,i) \\ P(c|\mathcal{A}, \Upsilon_k) &\propto \phi_k \cdot \sum_i \frac{r_i^{\mathcal{A}}}{r_{flat_i}} H(c,i) \end{aligned} \tag{5.3}$$

I sample whether the crossover c arose from $\mathcal{S}$ or $\mathcal{A}$ in proportion to the posterior probabilities above.

I next sample the inter-SNP interval in the map that the crossover arose in, in the same way as in Chapter 2. If the crossover was sampled to the $\mathcal{S}$ map, the sampling is done relative to the rates in that map, and the $\mathcal{A}$ map otherwise. I then increment the corresponding event counts. For example, if the crossover was sampled to come from the Shared map, and was then sampled to interval j in that map, I increment $N_j^{\mathcal{S}}$.

I repeat this procedure for every ancestry switch crossover. This completes sampling for $\vec{N}^{\mathcal{S}}$ and $\vec{N}^{\mathcal{A}}$ for one iteration.

4. *Sampling recombination rates from crossover locations.* In this step, I sample $\mathcal{S}$ and $\mathcal{A}$ conditional on $\vec{N}^{\mathcal{S}}$ and $\vec{N}^{\mathcal{A}}$. By Bayes rule, we have that $P(\mathcal{S}|\vec{N}^{\mathcal{S}}) \propto P(\vec{N}^{\mathcal{S}}|S)P(\mathcal{S})$ and $P(\mathcal{A}|\vec{N}^{\mathcal{A}}) \propto P(\vec{N}^{\mathcal{A}}|A)P(\mathcal{A})$. We model the numbers of crossovers $\vec{N}^{\mathcal{S}}$ and $\vec{N}^{\mathcal{A}}$ as having a multinomial distribution with rates in the $\mathcal{S}$ and $\mathcal{A}$ maps respectively.

$$\begin{aligned} \vec{N}^{\mathcal{S}} = (N_1^{\mathcal{S}}, N_2^{\mathcal{S}}, \ldots, N_M^{\mathcal{S}}) &\sim \texttt{Multinomial}\left(\sum_i N_i^{\mathcal{S}}, (r_1^{\mathcal{S}}, r_2^{\mathcal{S}}, \ldots, r_M^{\mathcal{S}})\right) \\ \vec{N}^{\mathcal{A}} = (N_1^{\mathcal{A}}, N_2^{\mathcal{A}}, \ldots, N_M^{\mathcal{A}}) &\sim \texttt{Multinomial}\left(\sum_i N_i^{\mathcal{A}}, (r_1^{\mathcal{A}}, r_2^{\mathcal{A}}, \ldots, r_M^{\mathcal{A}})\right) \end{aligned}$$

In order to incorporate the long-range correlation of rates in humans (Section 2.5), I calculate a new prior distribution in each iteration for the rates

in both maps, in the spirit of empirical Bayes. I construct basic maps, as described in Chapter 2. The Shared basic map is constructed from crossovers sampled to $\mathcal{S}$ in Step 3. The rates are then normalized to sum to 45,000 (events), reflecting a prior weight of approximately 1,250 meioses. I denote the rates of the Dirichlet prior for $\mathcal{S}$ as $(\lambda_1^{\mathcal{S}}, \lambda_2^{\mathcal{S}}, \ldots, \lambda_M^{\mathcal{S}})$, and the corresponding rates for $\mathcal{A}$ are $(\lambda_1^{\mathcal{A}}, \lambda_2^{\mathcal{A}}, \ldots, \lambda_M^{\mathcal{A}})$. The conjugacy of the multinomial and Dirichlet distributions is used to construct a Dirichlet posterior distribution for the rates:

$$\begin{aligned}(r_1^{\mathcal{S}}, r_2^{\mathcal{S}}, \ldots, r_M^{\mathcal{S}})|\vec{N}^{\mathcal{S}} &\sim \texttt{Dirichlet}\left(N_1^{\mathcal{S}} + \lambda_1^{\mathcal{S}}, N_2^{\mathcal{S}} + \lambda_2^{\mathcal{S}}, \ldots, N_M^{\mathcal{S}} + \lambda_M^{\mathcal{S}}\right)\\ (r_1^{\mathcal{A}}, r_2^{\mathcal{A}}, \ldots, r_M^{\mathcal{A}})|\vec{N}^{\mathcal{A}} &\sim \texttt{Dirichlet}\left(N_1^{\mathcal{A}} + \lambda_1^{\mathcal{A}}, N_2^{\mathcal{A}} + \lambda_2^{\mathcal{A}}, \ldots, N_M^{\mathcal{A}} + \lambda_M^{\mathcal{A}}\right)\end{aligned}$$

We use these distributions to sample rates for each map.

5. *Sampling the African-enrichment for each individual.* Every crossover was sampled to either the $\mathcal{S}$ or the $\mathcal{A}$ map in Step 3. For individual Υ_k, the total number of crossovers is denoted n_k. Of these, let the number of her crossovers that were mapped to the $\mathcal{S}$ map be $n_k^{\mathcal{S}}$ and the $\mathcal{A}$ map be $n_k^{\mathcal{A}}$. It is natural to model these counts as a binomial distribution based on her mixing proportion of the maps, i.e., ϕ_k,

$$n_k^{\mathcal{A}} \sim \texttt{Binomial}(n_k, \phi_k)$$

We have assumed a uniform prior on ϕ_k, which is the same as the Beta(1,1) distribution. We use the conjugacy of the beta and binomial distribution to obtain the posterior distribution for ϕ_k

$$\phi_k|n_k^{\mathcal{S}}, n_k^{\mathcal{A}} \sim \texttt{Beta}\left(n_k^{\mathcal{A}} + 1, n_k^{\mathcal{S}} + 1\right)$$

We sample ϕ_k for each individual using their respective posterior distribution. This completes one iteration of the MCMC.

- **Completion of the MCMC**. Ten chains with at least 70,000 samples each were run, and 15,000 samples per chain were discarded as burn-in.

Unlike the Gibbs sampler of Chapter 2, which is run separately for each chromosome, this MCMC works with the whole genome at once to maximize the information available to estimate the African-enrichment of each individual. Due to the high memory requirements of this process, I performed this analysis with a subset of the data. Autosomal data from 18,000 unrelated individuals were included (in contrast with the AA map which is built using 30,000 individuals and includes the X chromosome). I computed the mean recombination rate for every SNP interval using every 25^{th} sample for both maps, to produce Shared and African-specific maps.

5.3.2 Association testing of recombination phenotypes

I perform association testing of crossover-related phenotypes in both pedigrees and unrelated individuals. As discussed in Chapter 3, crossovers were directly identified in pedigrees by comparing the genotypes of parents and children in a model-based framework. In families with only one parent genotyped, we can confidently infer the alleles of the missing parent from the children's alleles in a substantial fraction of the genome. In this way, we can directly test association between genotypes and phenotypes in the parents in nuclear families. The testing in unrelated individuals is indirect and we justify doing so below.

5.3.2.1 Association mapping of recombination phenotypes is meaningful in unrelated individuals

Crossovers that are detected by ancestry switches in an individual have occurred in a recent ancestor of the individual, not in the individual himself. Therefore, for any crossover, the genotypes of the tested individual are only sometimes descended from the ancestor in whom the crossover occurred. This adds noise to the association signal, which I have estimated in an idealized haploid setting. Let the number of generations of admixture in present-day individual Υ be n. Υ inherits his genetic material from 2 ancestors in his parents' generation, 4 ancestors in his grandparents' generation, 8 from his great-grandparents' generation, and so on. The total amount of genetic material inherited by Υ, however, is the same from each generation, i.e., one genome. Therefore, the number of new crossovers that are passed on is also the same, in expectation, per generation. In other words, the probability that a crossover detected in Υ occurred in an ancestor k generations ago (with Υ's parents being $k = 1$), is $\frac{1}{n}$. Now we calculate the probability that, given that the crossover occurred k generations ago, it occurred in the same ancestor whose allele is inherited by Υ at a particular (unlinked) locus. Since crossovers and alleles from an ancestor are equally likely to be inherited as any other ancestor in the same generation, and there are 2^k ancestors in the k^{th} generation, the probability of a match is simply $\frac{1}{2^k}$. Summing over the generations we get, therefore,

$$\sum_{k=1}^{n} \left(\frac{1}{2^k} \cdot \frac{1}{n} \right) = \frac{1}{n} \cdot \left(1 - \frac{1}{2^n} \right)$$

As discussed in Chapter 2, the number of generations of admixture in African Americans is approximately 6, on average. Therefore, the expected number of crossovers that occurred in ancestors with an allele that is ancestral to the one an individual carries today is about 1/6. The extension of this argument to a diploid setting is

straightforward. If exactly one of two alleles at the locus under consideration is causal, then 1 in 6 crossovers co-occur with the causal allele, as in the haploid case. However, twice as many, or 1 in 3, are detected on a genetic background containing at least one of the two alleles carried by the individual in whom they occurred. These considerations suggest that association testing in unrelated individuals is expected to be noisy. Nevertheless, given the large sample size of almost 30,000 individuals, I perform association testing on them also.

5.3.2.2 Recombination phenotypes

I calculated and mapped three phenotypes: African-enrichment, fraction of crossovers arising from LD-based hotspots, referred to as *hotspot usage*, and the genome-wide crossover rate.

- *African-enrichment (AE).* African-enrichment was estimated for 18,000 unrelated individuals as part of the algorithm to generate Shared and African-specific maps in Section 5.3.1. Due to computational constraints, however, I could not implement the algorithm while including all the individuals in the dataset. Here I present the procedure I used to estimate African-enrichment for all the unrelated individuals as well as the pedigree parents.

  We model crossovers in African Americans as arising from a mixture of the Shared and African-specific maps, denoted $\mathcal{S}$ and $\mathcal{A}$, as discussed previously. For unrelated individuals, as well as for parents in nuclear families, I estimate the mixture proportion of the African-specific map, referred to as their African-enrichment, and denoted ϕ. I use a Bayesian approach to estimate ϕ. Let C denote the set of crossovers in an individual Υ (C can contain either ancestry switches, with probability distributions as obtained in Section 2.4.3, or crossovers detected in families whose distributions are calculated in Section 3.4).

We have, by Bayes rule,

$$P(\phi|C, \mathcal{A}, \mathcal{S}) \propto P(C|\phi, \mathcal{A}, \mathcal{S})P(\phi|\mathcal{A}, \mathcal{S})$$

I impose a uniform (uninformative) prior on ϕ. I assume that there is no interference and each crossover is independent to compute the likelihood

$$\begin{aligned} P(C|\phi, \mathcal{A}, \mathcal{S}) &= \prod_{c_k \in C} P(c_k|\phi, \mathcal{A}, \mathcal{S}) \\ &= \prod_{c_k \in C} (\phi P(c_k|\mathcal{A}) + (1-\phi)P(c_k|\mathcal{S})) \end{aligned}$$

How to compute $P(c_k|\mathcal{A})$ and $P(c_k|\mathcal{S})$ required for this calculation is shown in Equation 2.7 for unrelated individuals. In pedigree individuals, the calculation is the same, with the exception that the PMF of the crossover is derived not from ancestry transitions but from family analysis (Section 3.4). I use the likelihood and the prior distribution to obtain the posterior distribution $P(\phi|C, \mathcal{A}, \mathcal{S})$ for each individual, up to a normalizing constant. I calculate $P(\phi|C, \mathcal{A}, \mathcal{S})$ over a dense grid of $\phi \in [0, 1]$ and numerically integrate to obtain the expected value of ϕ under the posterior.

Events overlapping the terminal 5 Mb of each chromosome were removed from the analysis since estimated rates are less reliable near the telomeres. This is because our study as well as deCODE [Kong et al., 2010] are more likely to miss sub-telomeric crossovers and thereby underestimate rates in these regions. For estimating ϕ in pedigree parents, I further restrict to crossover events that are resolved within 100 kb.

- *Hotspot Usage.* We wished to estimate the fraction of each individual's crossovers that arise from hotspots identified previously from genome-wide HapMap2 LD data (Myers et al. [2005]). The authors identified 32,991 hotspots that were

active in at least two of the three constituent populations (Europeans, Africans and Asians). LD-based hotspots are known to be enriched for the 13-bp Myers motif CCnCCnTnnCCnC [Myers et al., 2008], as discussed before, and hotspot usage is a phenotype that has previously been shown to associate with variation in the *PRDM9* gene in humans [Baudat et al., 2010; Kong et al., 2010]. Different studies have estimated hotspot usage using calculations that are similar, in spirit, but different in implementation. We model crossovers as arising from a mixture of (i) a map containing only LD-based hotspots, and (ii) a map that does not have any hotspots, but correct broad-scale rates. The intuition behind this is that broad-scale rates are highly similar across populations (Section 4.4) These maps, therefore, offer a choice between placing crossovers in the set of LD-based hotspots, that is known to be active and shared among populations, versus a map that does not distinguish between this set of hotspots or any others that may be more active in particular populations. I construct the 'hotspot map' by distributing the total probability of recombination in a chromosome equally between the hotspots in that chromosome, and assigning a zero probability of crossover everywhere else. I calculate the 'broad-scale map' by smoothing the LD-based map in moving windows 5 Mb long. I then estimate the mixing proportion of the maps in the same way as I did for African-enrichment above, and define it to be the hotspot usage.

- *Genome-wide crossover rate.* I count the total number of crossovers observed for each parent in the pedigrees. I do not test for genome-wide rate in unrelated individuals due to its strong dependence on the number of generations since admixture in every lineage in the genealogy of the individual, which is unknown.

5.3.2.3 Genome-wide association testing procedure

I performed association testing for 29,589 unrelated individuals and 444 parents from the pedigrees. Genotypes at up to 3,058,149 HapMap2 SNPs were imputed into the unrelated individuals using MaCH software [Li et al., 2010] with YRI and CEU haplotypes as reference panels. I tested for association both at genotyped and imputed SNPs. I performed association testing separately for each of the seven datasets (unrelated individuals from the CARe consortium, pedigrees from the CARe consortium, unrelated individuals from CHOP, pedigrees from CHOP, and unrelated individuals from the African American Prostate, Breast and Lung Cancer Consortia) as follows.

1. *AE and hotspot usage phenotypes.* I restricted the testing of unrelated individuals to those who had at least 35 filtered ancestry-switch crossovers (approximately half of the mean number of events per unrelated individual). This is done in order to reduce statistical noise; it ensured that for all individuals included in the analysis, we had a sufficient number of crossovers to reliably estimate the phenotype. In pedigrees I re-scaled AE and hotspot usage for male and female individuals to have the same mean and variance.

   I tested for association by performing linear regression of individuals' phenotypes against their genotypes. For pedigree parents, no additional regressors were included. Genome-wide ancestry, for example, was significantly correlated neither with AE (R^2=0.006, P-value=0.24) nor with hotspot usage (R^2=0.005, P-value=0.27).

   For unrelated individuals, however, genome-wide ancestry was included as regressor. It is crucial to include genome-wide ancestry as a regressor when testing in unrelated individuals. As discussed in Section 5.3.2.1, the alleles of the present-day individual at a causal locus may not be inherited from the same ancestor in whom a crossover happened. However, other segments of the genome

may be inherited from that ancestor, and the ancestry of those segments may be informative about the *alleles in the ancestor at the causal locus*. They may, therefore, correlate indirectly with the phenotype and lead to false positive associations.

Further, I noticed that the resolution of ancestry-switch crossovers was correlated with genome-wide ancestry, and the resolution of a crossover had an effect on corresponding measurements of AE and hotspot usage. I observed that this correlation was largely explained by the better resolution of ancestry switches that transitioned from 1 to 2 European alleles (and vice versa), relative to switches that transitioned from 0 to 1 European alleles (and vice versa). In other words, the ancestry of the *non-recombining chromosome* had an effect on resolution, with African ancestry resulting in more poorly resolved switches. I suspect that this is due to the difficulty in phasing African chromosomes due to greater haplotype diversity in that population. Including genome-wide ancestry as a regressor accounts for both effects.

2. *Genome-wide crossover rate phenotype.* I tested mothers and fathers from the pedigrees separately as well as together. For the joint test I re-scaled the phenotypes for mothers and fathers to have the same mean and variance. I performed linear regression of their crossover counts with respect to their genotypes, including their genome-wide ancestry as a regressor.

3. *Meta-analysis for all phenotypes.* Subsequent to testing separately in each of seven datasets, I performed a meta-analysis to increase statistical power. I computed Z-scores from the P-values from all seven datasets (including the pedigree data and the unrelated individuals), and then computed the combined P-value. The basic idea is as follows: consider a linear regression model with intercept and residuals $\epsilon \sim N(0, \sigma^2)$, i.e., $\vec{y} = \beta_0 + \beta\vec{x} + \epsilon$. Under the null

hypothesis of no association, the estimate of the parameter $\hat{\beta}$ has a normal distribution $N(0, \sigma^2_{\hat{\beta}})$, where $\sigma^2_{\hat{\beta}}$ depends on aspects of the data, i.e., σ^2, variance of $\vec{x}$ and the number of observations. $\sigma^2_{\hat{\beta}}$ is unknown, however, it is estimated from the data. Estimates from different datasets can be combined linearly, and the combined estimator with the lowest total variance is obtained by weighting $\hat{\beta}$ from each dataset in proportion to the inverse of its respective $\sigma^2_{\hat{\beta}}$. A combined P-value can thus be calculated.

4. *Genomic Control.* Finally, I correct P-values using genomic control [Devlin and Roeder, 1999]. The intention is to compare the distribution of the observed test statistics (e.g., $\hat{\beta}^2/\sigma^2_{\hat{\beta}}$) with the theoretically expected test distribution under the null hypothesis of no association (χ^2_1). Devlin and Roeder [1999] showed, in the context of case-control studies, that population substructure results in over-dispersion of test statistics of association, which may lead to spurious rejection of the null hypothesis. In the next section, I report the finding of a true signal of association at SNP rs6889665 on chromosome 5. I calculated a genomic control inflation factor after removing SNPs within 10Mb of rs6889665. I identify a very mild inflation of test statistics, with inflation factors of $\lambda = 1.046$ for AE and $\lambda = 1.038$ for hotspot usage. P-values reported in the next section are corrected for these factors.

I have used a threshold of 3.3×10^{-9} to determine genome-wide significance. This corresponds to a significance threshold of 0.01 after performing Bonferroni correction for multiple hypothesis testing.

5.3.2.4 Results of association testing

Genome-wide P-values of association for AE are shown in Figure 5.7. The SNP showing the strongest association with AE is rs6889665 ($P = 1.5 \times 10^{-246}$), which has a derived allele frequency of 29% in YRI and 2% in CEU, and is within 4 kb

of the C-terminal end of the *PRDM9* zinc finger array (Figure 5.8). This SNP is also the most strongly associated SNP for AE in pedigree individuals separately ($P = 3.3 \times 10^{-54}$). The highly significant P-value in unrelated individuals confirms that there is considerable power to detect associations despite a noisy phenotype (Section 5.3.2.1). rs6889665 is also the SNP most strongly associated with hotspot usage ($P = 1.8 \times 10^{-52}$). No locus outside *PRDM9* is significant for either phenotype ($P < 0.01$ after Bonferroni correction). Table 5.1 lists the top hits for AE and hotspot usage. No genome-wide significant associations were observed for genome-wide crossover rate.

Comparing the P-values of association of rs6889665 for AE and hotspot usage phenotypes (Table 5.1) demonstrates the greater power of using genome-wide maps, as opposed to only strongly signalled hotspots, in studying recombination. It also demonstrates that the African-specific map is, as intended, highly able to capture individuals carrying crossovers in regions that are not recombinogenic in Europeans, and that our indirect phenotyping approach still has power to identify genuine associations.

To further explore variants of *PRDM9* that influence the crossover landscape, I tested association of SNPs within 20 Mb of *PRDM9* conditional on rs6889665. To do this, I performed linear regression as before, but with the genotype at rs6889665 as an additional predictor variable. Several SNPs continued to have significant associations conditional on rs6889665 (Figure 5.9), of which the most significant is rs10043097 ($P = 8.3 \times 10^{-14}$), which is upstream of the *PRDM9* transcription start site. I dissect the underlying causes of association of rs6889665 and additional SNPs with AE in the next section.

For the genome-wide crossover rate phenotype, I find that the derived allele of rs6889665 is weakly associated with an increase in the total genetic map length ($P = 0.037$). I estimate that the effect size is approximately 1.7 Morgans per copy of the

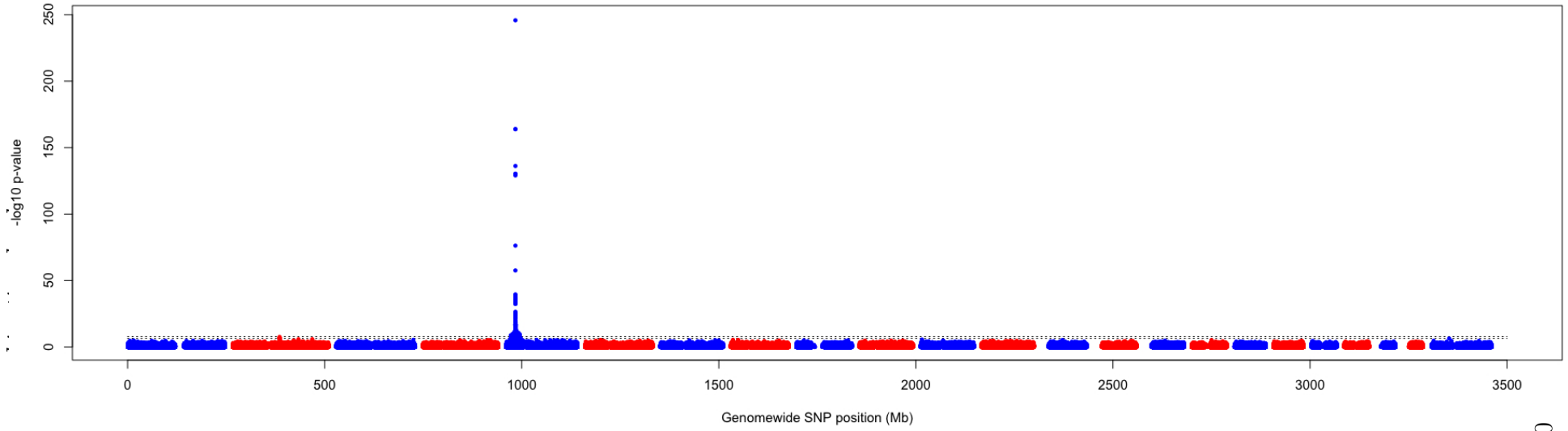


Figure 5.7: Manhattan plot of P-values for genome-wide association testing of more than 3 million SNPs for the African-enrichment (AE) phenotype in pedigrees and unrelated individuals. Changing colors indicate alternating chromosomes, with the non-autosomal portion of the X chromosome as the 23^{rd} chromosome (in blue). The y-axis corresponds to -log10(P-values) for association. Dashed lines indicate thresholds for genome-wide significance (P <0.01 after Bonferroni correction) and suggestiveness (P<1). The most associated SNP is rs6889665 on chromosome 5 which is 4 kb downstream of the *PRDM9* gene ($P = 1.5 \times 10^{-246}$). No SNP outside the 10 Mb window on either side of rs6889665 achieves genome-wide significance. rs11888485 on chromosome 2 has a P-value of 0.09 after Bonferroni correction.

SNP	P-value	YRI allele frequency*	CEU allele frequency*	# of SNPs in 10 Mb with $P < 3.3 \times 10^{-7}$	Chr: Hg18 position	Genes within 50 kb/ notes
African enrichment						
rs6889665	1.5×10^{-246}	0.29	0.02	506	5:23,568,400	*PRDM9*
rs11888485	2.9×10^{-8}	0.92	1.00	4	2 : 118,403,702	*CCDC93*
Hotspot usage						
rs6889665	1.8×10^{-52}	0.71	0.98	36	5:23,568,400	*PRDM9*
rs9987353	7.8×10^{-9}	1.00	0.70	4	8 : 9,153,759	*cis*-effect†
rs10015037	1.5×10^{-7}	0.83	0.85	1	4: 98,101,343	

Table 5.1: Three recombination-related phenotypes were tested: African-enrichment, hotspot usage, and genome-wide crossover rate (genome-wide rate was tested pedigrees only). This table reports SNPs with a suggestive $P < 3.3 \times 10^{-7}$ (less than one hit expected in a scan in the absence of a true signal). Due to the clustering of significant SNPs, I pick the SNP with the smallest P-value to represent each region (defined as a 10 Mb window in either direction of the most associated SNP).

*The frequency of the allele that confers a higher value for the phenotype is quoted.

†rs9987353 occurs within the *β-Defensin* cluster at 8p23.1, a region which is known to harbour complex rearrangements in humans [Giglio et al., 2001]. We suspected that the signal may be an artefact or due to an interaction between recombination and these rearrangements. Therefore, I retested SNPs in chromosome 8 for association to a new AE phenotype, which was defined using crossovers on all chromosomes other than chromosome 8 itself. The signal disappeared, thereby confirming that this is a local effect. Unusual behaviour in this region was also noted by Wegmann et al. [2011] who also used ancestry inference to detect recombination.

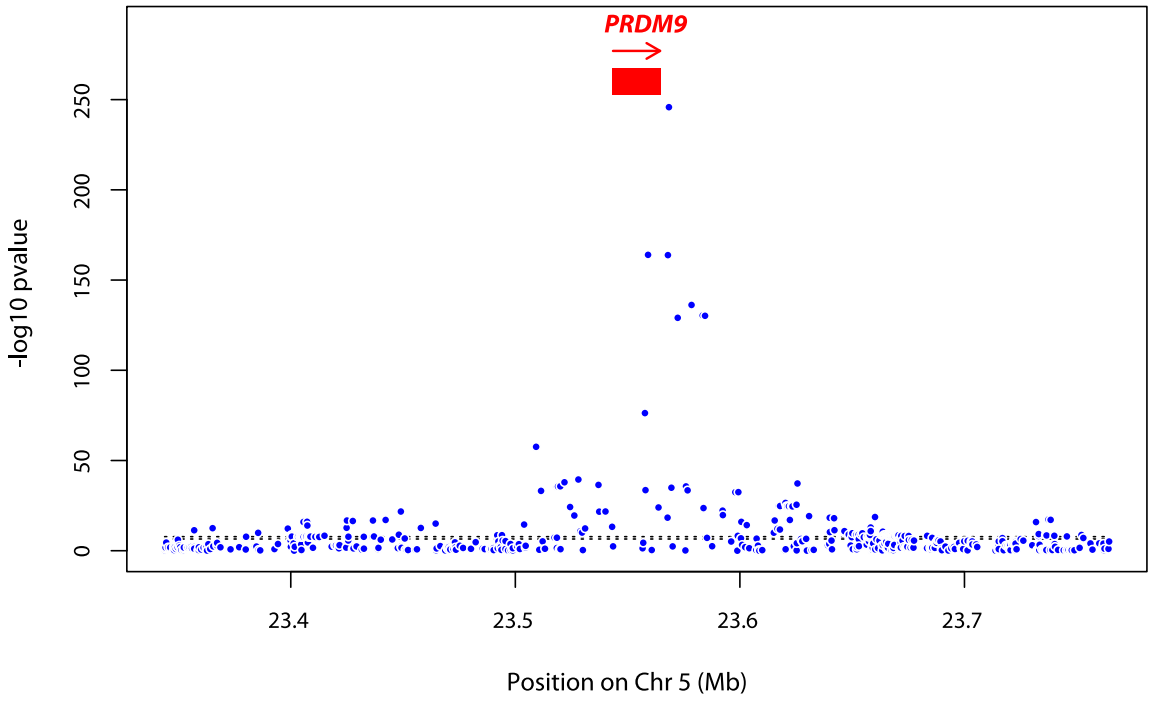


Figure 5.8: Association of *PRDM9* genetic variation with the African-enrichment phenotype. A single genome-wide significant peak is observed at *PRDM9* with rs6889665 the most associated SNP.

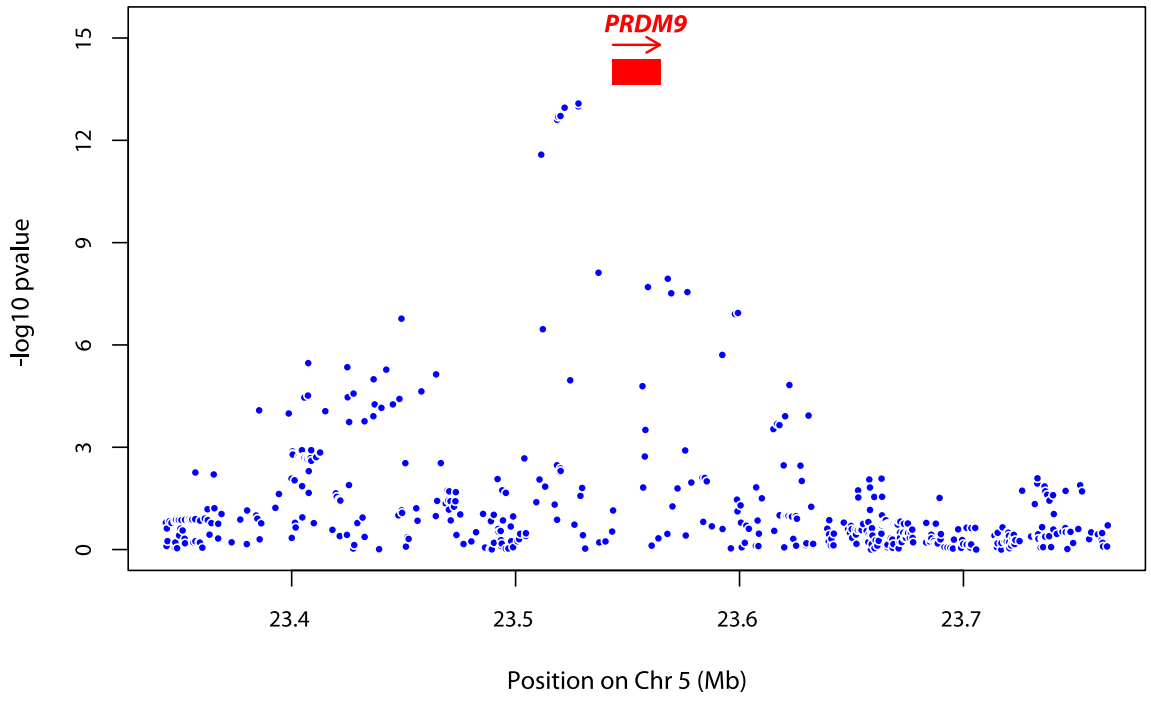


Figure 5.9: Association of *PRDM9* genetic variation with African-enrichment phenotype, conditional on rs6889665. Significant residual signals are observed in several SNPs. The strongest residual association is with rs10043097, upstream of the *PRDM9* transcription start site.

allele in males and 0.9 Morgans in females. Baudat et al. [2010] have previously noted that a particular PRDM9 allele, the B allele, increases total genetic map length in both males and females by an estimated 2.5 Morgans. This adds to evidence that PRDM9 can influence genome-wide rate, which is interesting as this is a phenotype on which natural selection may be active in humans [Stefansson et al., 2005].

As discussed in the introduction to this chapter, Kong et al. [2008] demonstrated that haplotypes in the gene *RNF212* are associated with genome-wide rate in both males and females in an Icelandic population. This finding has since been replicated in humans [Chowdhury et al., 2009; Fledel-Alon et al., 2011] and cattle [Sandor et al., 2012]. We do not, however, find a significant association for this gene. In both Europeans and West Africans, three haplotypes defined by SNPs rs3796619 and rs1670533 are found: [C,T], [T,C] and [T,T], however they segregate at very different frequencies. In females, Kong et al. found that the haplotype [T,C] is associated with higher genetic map length relative to the haplotypes [C,T] and [T,T]. This haplotype is present at 19.2% frequency in CEU but at only 1.7% frequency in YRI. Due to the very low allele frequency of this variant in African ancestry populations, we do not expect to see a significant association in females in our study. In males, Kong et al. observed that the haplotype [C,T] is associated with a strong increase in recombination rate relative to the haplotype [T,C] (effect size ≈ 0.78 M, $P = 7 \times 10^{-23}$ in 3,135 males) and a somewhat weaker effect relative to the haplotype [T,T]. As mentioned above, the haplotype [T,C] is nearly absent in West Africans, and therefore we are not likely to see an association between these haplotypes in males either. I believe that our inability to find an association is due both to the modest size of the effect and the small sample size in our testing (22 samples genotyped at rs3796619 and 80 imputed).

I do not find an association with the chromosome 17q21.31 inversion either [Stefansson et al., 2005]. I believe this is due to the modest size of the effect (0.47

crossovers per copy of the inversion in mothers) and the inversion being uncommon in African populations (frequency $\sim 6\%$).

5.4 Inferring the mechanism of PRDM9 action

Next we try to understand the biological differences between *PRDM9* alleles that are tagged by the top hit SNP rs6889665. The SNP is in the same LD block as *PRDM9* in both YRI and CEU populations (Figure 5.10), and is about 4 kb downstream of its 3' end. As discussed in Section 5.1, PRDM9 has a highly variable zinc finger array in humans, and the predicted binding motif of the dominant European allele, the A allele, is highly enriched in LD-based hotspots. This leads us to hypothesize that rs6889665 may be tagging one or more *PRDM9* alleles with different zinc finger arrays that mark African-specific hotspots. This hypothesis was tested by my colleagues, as I report briefly in Section 5.4.1. Next, Dr Simon Myers leveraged the Shared and African-specific maps to identify a motif that is strongly enriched in African-specific hotspots, which is discussed in brief in Section 5.4.2. Finally, in Section 5.4.3, I examine the role of further variants that were significant in the association testing above, even after conditioning on rs6889665, and the relationship between different PRDM9 alleles in heterozygous individuals in Section 5.4.4.

5.4.1 rs6889665 tags *PRDM9* alleles similar to the C allele

The work in this section, following up my finding that rs6889665 is significantly associated with recombination, was done primarily by my colleagues Yunli Song, Nadin Rohland and my supervisor Simon Myers.

Yunli Song (as part of his PhD work) and Simon Myers used short-read sequencing data from the pilot stage 1000 Genomes Project [1000 Genomes Project Consortium, 2010] to assemble and infer the *PRDM9* zinc finger array for each of 146 individuals

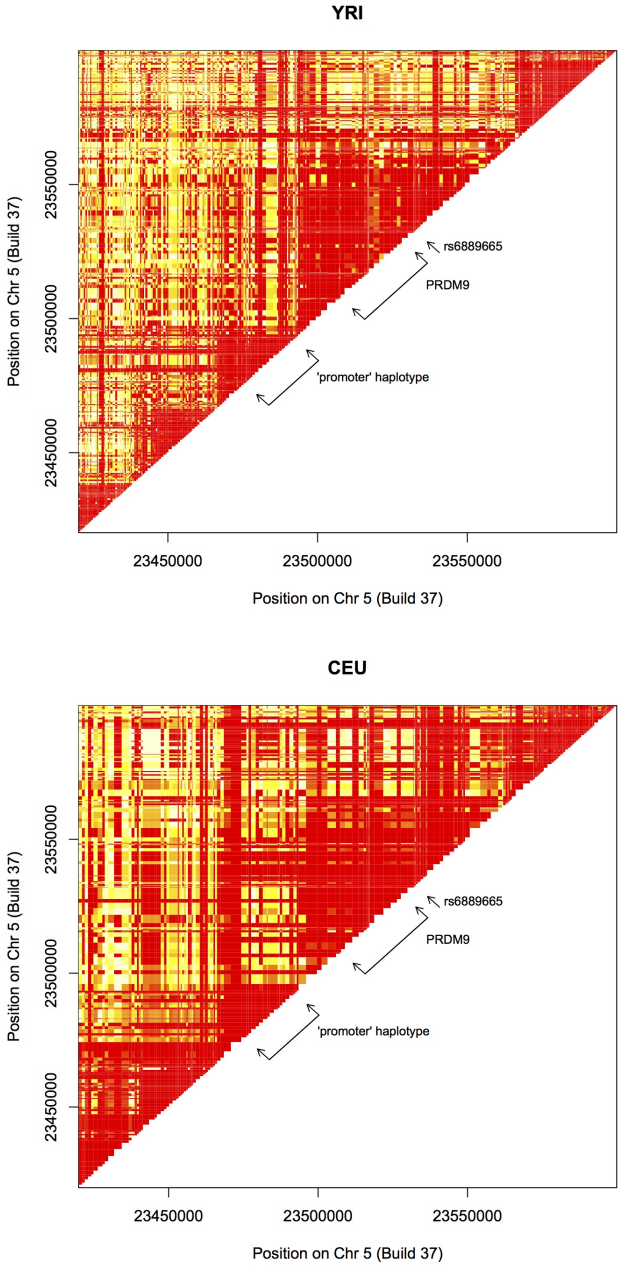


Figure 5.10: D' in the vicinity of the *PRDM9* gene. The top hit SNP rs6889665 is in the same LD block as *PRDM9* in both YRI and CEU populations. The top residual signal comprises a haplotype *(promoter haplotype)*, which includes the SNP rs10043097, and is in a separate LD block upstream of the gene. These plots are in Build 37 and utilize Phase I data from the 1000 Genomes project [1000 Genomes Project Consortium, 2012].

included in that study. The set of possible alleles was assumed to comprise the 29 *PRDM9* alleles previously identified via targeted sequencing by Berg et al. [2010]. Yunli and Simon designed a novel approach [Hinch et al., 2011] to type the complex and repetitive structure of the *PRDM9* gene, wherein they re-mapped all 1000 Genomes reads in the vicinity of the gene, and used a model-based approach to calculate the likelihood of observed reads conditional on pairs of *PRDM9* alleles drawn for the known set above. They also estimated the frequency distribution of each of the 29 alleles above in multiple populations to obtain the *maximum a posteriori* (MAP) estimator of the joint probability distribution of the pair of *PRDM9* alleles carried by each individual.

In addition to sequencing *PRDM9* alleles, Berg et al. [2010] made a bioinformatic prediction of the binding target of each allele [Persikov et al., 2009], and grouped alleles by how many bases the binding target matched the 13-bp Myers motif CCnCCnTnnCCnC. Since there are 8 non-degenerate bases in the Myers motif, the grouping can range from 1 to 8 matching bases in principle. Among West African, European and Indian samples analyzed [Berg et al., 2010, 2011], alleles predicted to bind motifs matching 4, 5, 6, 7 and 8 bases have been observed (Figure 5.2). For 46 HapMap2 CEU and 44 HapMap2 YRI individuals, Simon and Yunli use the MAP estimates above to integrate over the alleles in each group and calculate, for each individual, the expected number of their alleles that match 4, 5, 6, 7 or 8 bases of the Myers motif. The results are shown in Figure 5.11. There is an almost complete association between individuals' genotype of rs68869665 SNP and their *PRDM9* allele class, with the number of copies of the C variant of rs6889665 almost exactly tagging the number of alleles matching 5 bases of the Myers motif that they carry.

To refine our understanding of this association, Simon re-estimated the world-wide allele frequencies of each *PRDM9* allele as above, while this time permitting different allele frequencies, depending on the rs68869665 background. The estimated

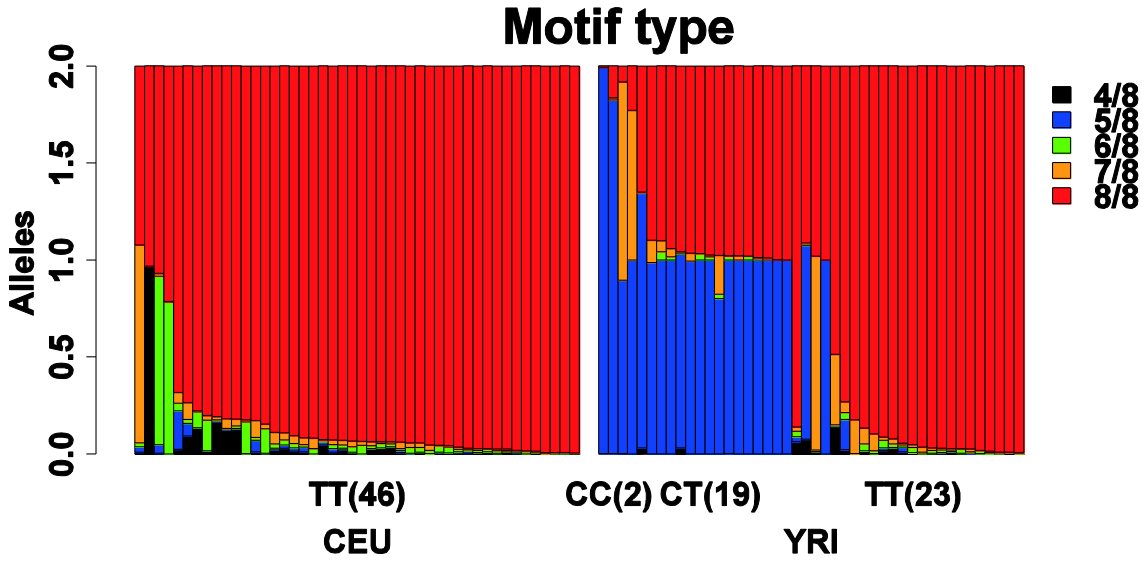


Figure 5.11: Variation of predicted *PRDM9* allele classes associates with individuals' genotype of rs6889665. Each column represents an individual from the CEU panel (left) or YRI panel (right). The y-axis shows, of each individual's two alleles, the expected number predicted to match different numbers of bases of the motif. Note that individuals carrying k copies of the C allele at rs6889665 are almost always predicted to have k *PRDM9* alleles with a predicted 5/8 motif match ($k \in \{0, 1, 2\}$).

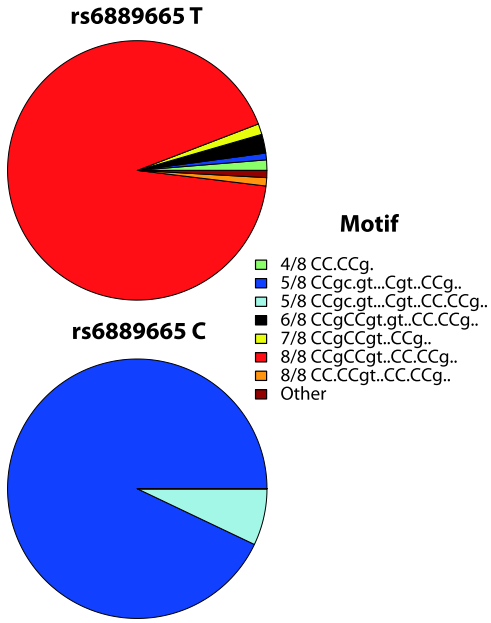


Figure 5.12: rs6889665 C allele tags *PRDM9* alleles with a 5-match to the Myers motif. This work refines the analysis in Figure 5.11 and shows the frequency distribution of *PRDM9* alleles co-occurring with variants of rs6889665.

allele frequencies conditional on the rs6889665 background are shown in Figure 5.12. This analysis demonstrates that the T variant (the ancestral allele) at rs6889665 is strongly correlated to alleles with an exact (8/8) match to the Myers motif, whereas the C variant (the derived allele) is almost perfectly correlated to a group of alleles, all predicted to bind a common, different 17-bp motif CCgCngtnnnCgtnnCC ¸which matches the 13-bp motif at only 5 bases. The most common *PRDM9* allele in YRI which has a 5-match to the motif is the C allele with frequency $\sim 13\%$ [Berg et al., 2011].

Another dimension of *PRDM9* variation in humans is variability is the number of zinc fingers within the array, and alleles containing 7 to 20 zinc fingers have been observed in humans. My colleague Nadin Rohland experimentally measured the number of zinc-finger domains in PRDM9 in 354 individuals including HapMap2 CEU and YRI samples and 166 African Americans from the pedigree study. She also genotyped rs6889665 in these individuals. Trios and family data were used to assign phase to the SNP as well as the zinc finger array. The results are shown in Figure 5.13. This shows, again, that rs6889665 differentiates *PRDM9* alleles into two different classes: 96% of haplotypes carrying the T variant of rs6889665 have 13 zinc fingers (compatible with the 8-match A and B alleles of *PRDM9*, each of which has 13 zinc fingers) and 93% of haplotypes carrying the C variant have 14 or more ZFs (compatible with a family of 5-match alleles including the C, L4, L6 and L14 alleles, amongst others).

To summarize, we demonstrate that the association of differences in the usage of the African-specific and European recombination maps with rs6889665 is caused by differences in the DNA binding properties of particular PRDM9 variants with different zinc finger arrays.

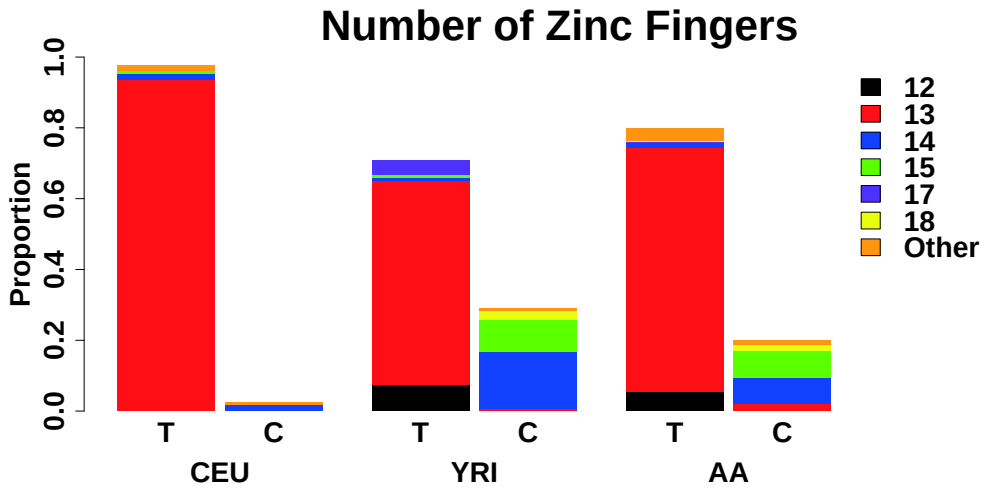


Figure 5.13: Variation of *PRDM9* zinc finger length with rs6889665. rs6889665 C variant is highly associated with *PRDM9* alleles containing 14 or 15 zinc fingers. The *PRDM9* C allele is the most common allele among YRI which contains 14 zinc fingers ($\sim$ 13% frequency). The most common allele containing 15 zinc fingers is L6 ($\sim$ 6% frequency). Both C and L6 alleles have predicted binding sequences that match 5/8 bases of the Myers motif.

5.4.2 Finding a motif underlying African-specific hotspots

Given the relationship between the African-specific *PRDM9* alleles with a 5/8 motif match and differential use of the African-specific map established in the previous section, we predicted that a new motif underlies African-specific hotspots. The search for this motif was performed by Simon Myers using the African-specific and Shared maps that I estimated in Section 5.3.1.

To identify African-enriched hotspot motifs, Simon first identified a more stringent set of candidate African-specific hotspots and a control set of Shared hotspots. To identify African-enriched hotspots, he used the two LD-based maps: YRI and CEU maps, and the African-specific and Shared maps. Information was combined from both pairs of maps to reduce noise and enrich for regions with genuine differences between the African and European populations. Specifically, he called African-specific hotspots as 2-kb intervals representing a peak in the African-specific map rate, and

where the estimated rate was at least 2 cM/Mb and at least twice the rate in the Shared map. In addition, the YRI map rate was also required to be at least 2 cM/Mb and at least double the CEU map rate. Simon took the resulting candidate hotspot set and defined hotspot boundaries by identifying the region flanking the 2 kb rate peak that had a rate at least 50% of the peak value in the African-specific map. Regions larger than 5 kb were discarded due to the greater noise in their hotspot localization.

Simon constructed a control set of Shared hotspots but modified the converse criteria given the lack of measurable hotspots present only in people of European ancestry (Section 5.2). Specifically, Shared hotspots were called in 2 kb regions where the Shared map had a rate peak and where both the Shared and CEU estimated rates were at least 2 cM/Mb, while the African-specific and YRI map rates were below those in the corresponding European maps. This resulted in 2,454 candidate African-specific hotspots with a control set of 7,328 shared hotspots likely to be more active in Europeans.

Simon tested all candidate motifs of 5 to 9 base pairs for enrichment in the African-specific hotspot set relative to the Shared hotspot set in both repeat and non-repeat backgrounds separately. For each motif, he counted its occurrences in the two hotspot sets, relative to the total number of motifs of the same length in each set, and tested for a frequency difference using a chi-squared test (1 degree of freedom). An overall P-value was calculated for each motif by converting repeat and non-repeat P-values to Z-scores, and then summing the Z scores to obtain a single overall P-value. Motifs were considered statistically significant only if they passed four criteria: (1) they were statistically significant after Bonferroni correction for the number of motifs tested; (2) they were over-represented in the African-specific hotspots; (3) they were statistically significant on both the repeat and non-repeat backgrounds independently; and (4) they were statistically significant after accounting for systematic differences in G/C

content in the two hotspot sets.

This testing revealed a unique 9-bp significant motif CCCCAGTGA. Simon extended the motif and explored its degeneracy by exploring whether flanking DNA around matches to this motif also had a role by testing whether bases at a given site relative to the motif were associated with difference in rates between African- and European-ancestry populations using the non-parametric Kruskal-Wallis test. Rates were evaluated in the 2 kb surrounding each motif occurrence. This lead to the identification of a 17-bp consensus African-specific motif CCCCaGTGAGCGTtgCc (Figure 5.14).

The 500 best matches to this motif have a 3-fold increase in average rate in the AA and YRI relative to the deCODE and CEU maps (Figure 5.14). Comparison of this 17-bp consensus to the binding motif predicted for the class of *PRDM9* alleles, which match 5 out of 8 bases in the Myers motif, reveals a very close concordance. ($P = 8.1 \times 10^{-6}$).

5.4.3 Understanding additional signals at *PRDM9*

I tested the association of African-enrichment with SNPs conditional on rs6889665 to analyse if there are further unexplained differences in the recombination landscape between individuals. Several SNPs near *PRDM9* continue to have highly significant signals, as shown in Figure 5.9 and listed in Table 5.2. These SNPs fall into two apparently separate LD blocks (Figure 5.10), and I analyze them in turn.

I first examined the role of the upstream haplotype that is genome-wide significantly associated with AE, even after conditioning on rs688966. Since this haplotype is in linkage equilibrium with rs6889665 (r^2 between rs10043097 and rs6889665 is ~ 0.001), I hypothesized that it is unlikely to tag variants of the *PRDM9* zinc finger array. Given its location upstream of the gene, I speculated that SNPs tagged by this haplotype may influence transcription factor binding. Two SNPs on this haplo-

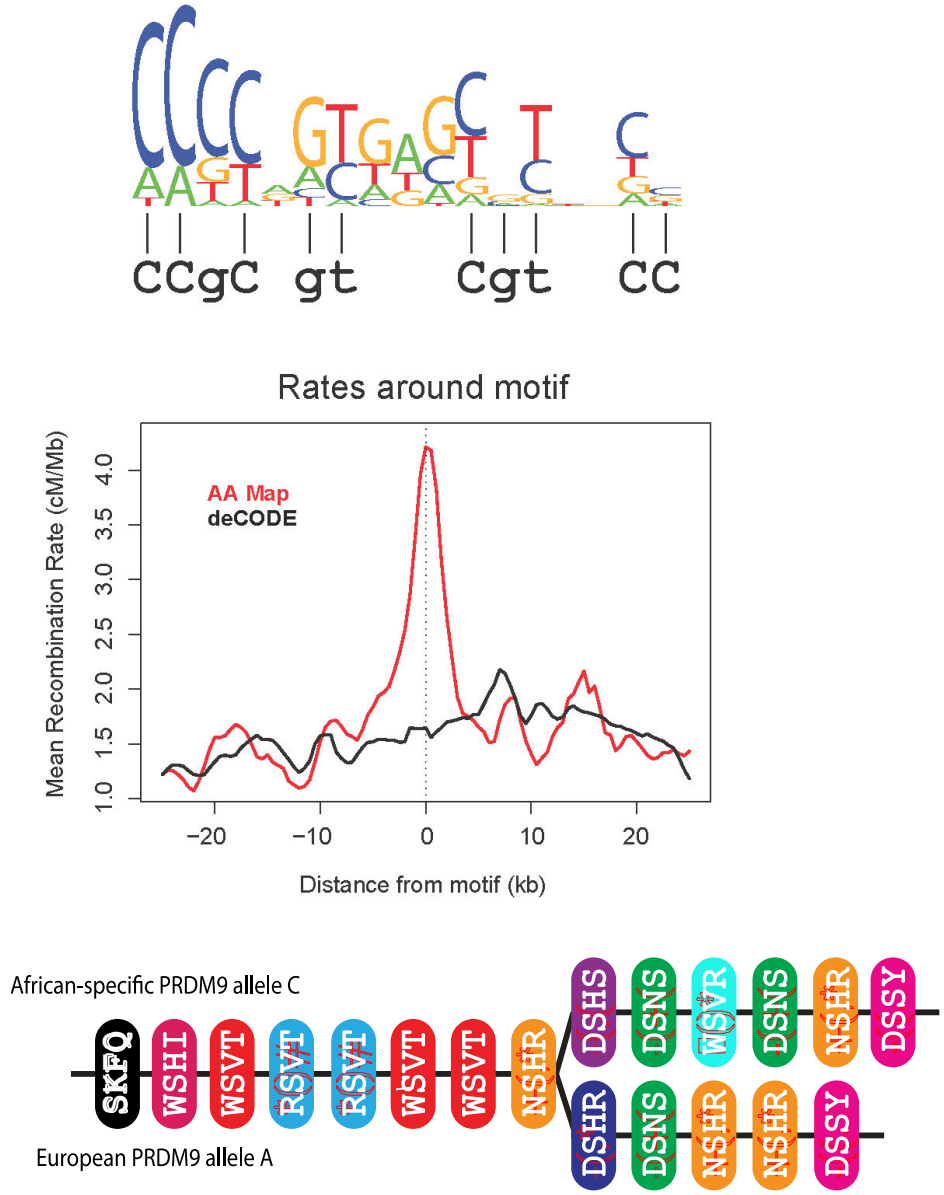


Figure 5.14: A 17-bp degenerate hotspot motif is highly enriched in African-specific hotspots. (Top) Logo plot showing the 17-bp motif, with stack height proportional to -log P-value of association with the specific base position, and relative letter height proportional to the mean crossover rate increase given each base. In black is the bioinformatic *PRDM9* binding prediction for the class of *PRDM9* alleles with a 5-match to the Myers motif. The two motifs match at 10 of 11 bases (vertical lines). (Middle) Average crossover rate in the AA (red line) and deCODE (black line) maps surrounding the 500 strongest motif matches (in 2-kb sliding windows). This work was done by Simon Myers. (Bottom) *PRDM9* A and C alleles have identical zinc fingers 1-8, while differing in downstream fingers, which is the region of the array that is predicted to bind the newly identified African-specific motif above.

SNP	P-value	YRI allele frequency*	CEU allele frequency*	Chr: Hg18 position
Upstream LD block				
rs13159483	2.6×10^{-12}	0.71	0.21	5:23,511,449
rs10039524	2.5×10^{-13}	0.71	0.22	5:23,518,536
rs10462242	2.0×10^{-13}	0.71	0.21	5:23,519,019
rs6872288	1.9×10^{-13}	0.71	0.22	5:23,520,040
rs2115260	1.1×10^{-13}	0.73	0.24	5:23,521,882
rs10043071	1.0×10^{-13}	0.68	0.18	5:23,527,983
rs10043097	8.3×10^{-14}	0.68	0.19	5:23,528,077
PRDM9 LD block				
rs955300	7.6×10^{-9}	0.97	0.30	5:23,537,028
rs1874165	2.0×10^{-8}	0.45	0.06	5:23,559,105
rs1603084	1.1×10^{-8}	0.45	0.06	5:23,567,951
rs12153202	3.1×10^{-8}	0.96	0.30	5:23,569,503
rs11746007	2.8×10^{-8}	0.97	0.31	5:23,576,715

Table 5.2: Residual signals of association with African-enrichment conditional on the genotype at rs6889665 with P-value $< 10^{-7}$. Of the five hits in the PRDM9 LD-block, rs955300, rs12153202, rs11746007 are in tight LD ($r^2 \approx 1$), as are rs1874165 and rs1603084. All SNPs in the upstream LD-block represent approximately one haplotype with nearly perfect pairwise LD ($r^2 \approx 1$).

*The frequency of the allele that confers a higher value for the African-enrichment phenotype is quoted.

type, the top SNP rs10043097 and rs10043071 both lie within a binding site of the transcription factor INI1 as determined by CHIP-seq (ENCODE Regulation track in Genome Browser Build 36). INI1, also known as SMARCB1 or SNF5, is a transcriptional activator and works as part of a complex to relieve repressive chromatin structures (Entrez Gene ID 6598 [Maglott et al., 2011]). Since this haplotype is in low LD with variants tagging the zinc finger array, however, change in AE implies that the haplotype must be controlling transcription in an allele-specific manner (if both alleles are equally influenced, the relative use of Shared and African-specific hotspots seems unlikely to be changed). This makes the hypothesis fairly complex, though still plausible.

Another possibility is that, even though it seems unlikely *a priori*, the haplotype is tagging variants of the *PRDM9* zinc finger array, and by chance, no SNP genotyped or imputed in our data, which in the same LD block as the zinc finger array, is a better proxy for it. To test this hypothesis, I used pedigree parents for whom we have genotype data for rs10043097 as well as the number of zinc fingers in their *PRDM9* arrays. Since rs6889665 is a nearly perfect tag for C-like zinc fingers, I restricted this analysis to individuals who are homozygous for the ancestral, major allele T for rs6889665. As shown in Figure 5.13, the most common allele associated with the T allele of rs6889665 has 13 zinc fingers, among whom are the A and B alleles with a combined allele frequency of $86\% - 96\%$ in European ancestry populations. To circumvent difficulties with phasing, I further restricted to individuals who are homozygous in rs10043097. I ask if we can reject the null hypothesis that the frequencies of length 13 zinc finger arrays are the same in individuals with 0 or 2 copies of the ancestral C allele of rs10043097. Table 5.3 presents the contingency table for this test. I perform Fisher's exact test and reject the null hypothesis (P-value $= 0.0004$). In addition to length 13 ZF arrays, the C variant of rs10043097 is found with alleles of length 10, 11, 12, 14, 17 and 18, of which 12 is the most common. I specifically

	rs10043097 = TT	rs10043097 = CC
Number of length 13 ZF arrays	51	48
Number of ZF arrays with length $\neq 13$	1	16

Table 5.3: Contingency table for testing allele frequency of length 13 ZF arrays in individuals carrying different genotypes of rs10043097

	rs10043097 = TT	rs10043097 = CC
Number of length 12 ZF arrays	0	9
Number of ZF arrays with length $\neq 12$	52	55

Table 5.4: Contingency table for testing allele frequency of length 12 ZF arrays in individuals carrying different genotypes of rs10043097

test for association of rs10043097 with arrays of length 12 and the contingency table is shown in Table 5.4. Again, I perform Fisher's exact test and reject the null hypothesis of no association (P-value 0.004). This analysis suggests that, despite being separated from the *PRDM9* ZF array by a hotspot, rs10043097 is tagging *PRDM9* alleles with 12 zinc fingers (and possibly others). This conclusion is further supported by the data in Figure 5.13, where we note that the T allele of rs6889665 is often associated with length 12 zinc finger array alleles in the YRI population but not in the CEU population. Known alleles with 12 zinc finger arrays and allele frequency 1% in any population are L7 and L11 with allele frequencies of 2% and 3% respectively in Africans and absent in Europeans. Both these alleles match 7/8 bases of the Myers motif. They were rarely inferred in 1000 Genomes Pilot (only one instance of L11 and none of L7, so association of these alleles with rs10043097 could not be explicitly tested). Other SNPs in the upstream LD block (Table 5.2) are no longer significant after conditioning on both rs6889665 and rs10043097.

I now examined the LD-block containing *PRDM9*. The top SNP in this block, rs955300 (conditional on rs6889665, Table 5.2) is still significantly associated with African-enrichment after conditioning on rs10043097 (P-value $< 10^{-5}$). After condi-

tioning on rs955300, rs1603084 also continues to be significantly associated with the phenotype, although the SNPs rs955300 and rs1603084 are highly correlated. The direction of the signal for both SNPs remained unchanged when the other SNP was included in the model.

To explore the history of the *PRDM9* zinc finger array and to place SNPs showing association with African-enrichment in this context, we identified 19 SNPs from HapMap2 that form a maximal block of SNPs where there is almost no evidence of recombination: $D' = 1$ for all pairs of SNPs in the data after removing 2 of 120 YRI and 1 of 120 CEU haplotypes. We used these haplotypes to construct a gene tree using the software *Genetree* [Griffiths and Tavaré, 1995], using the chimpanzee genome to define the ancestral allele. Genetree assumes a coalescent prior on genealogies and approximately infers branch lengths and ages of the mutations[3]. Figure 5.15 illustrates the gene tree.

I note that the SNP rs955300 (together with SNPs rs12153202 and rs11746007), which is on a branch under rs1603084, tags a pair of haplotypes with highly divergent allele frequencies in different populations. The derived haplotypes appear to have increased sharply in frequency among Europeans (69%) relative to African (3%) and Asian (41%) populations. At least a subset of the ancestral *PRDM9* A alleles must be on this haplotypic background because of the high frequency of the A allele among Europeans (86%). One possibility is that these haplotypes carry an allele such as L20, which has a distinct motif prediction (7/8 match with the Myers motif) but also contains the 13 zinc fingers. It is present in the European population at 4% frequency, and has not been observed in Africans [Berg et al., 2011]. However, it has

[3]The tree building uses data from HapMap2 array genotyping and thus potentially suffers from ascertainment bias in the choice of SNPs. In an ideal scenario, we would use sequence data for this construction. However, I discovered several genotyping or imputation errors in this LD-block when using low coverage 1000 Genomes Phase I data, as evidenced from highly discordant allele frequencies between YRI, CEU and African-American individuals. Since this analysis is very sensitive to even small numbers of genotyping or phasing errors, I report the analysis with HapMap2 data where the use of trios assured highly confident haplotypes.

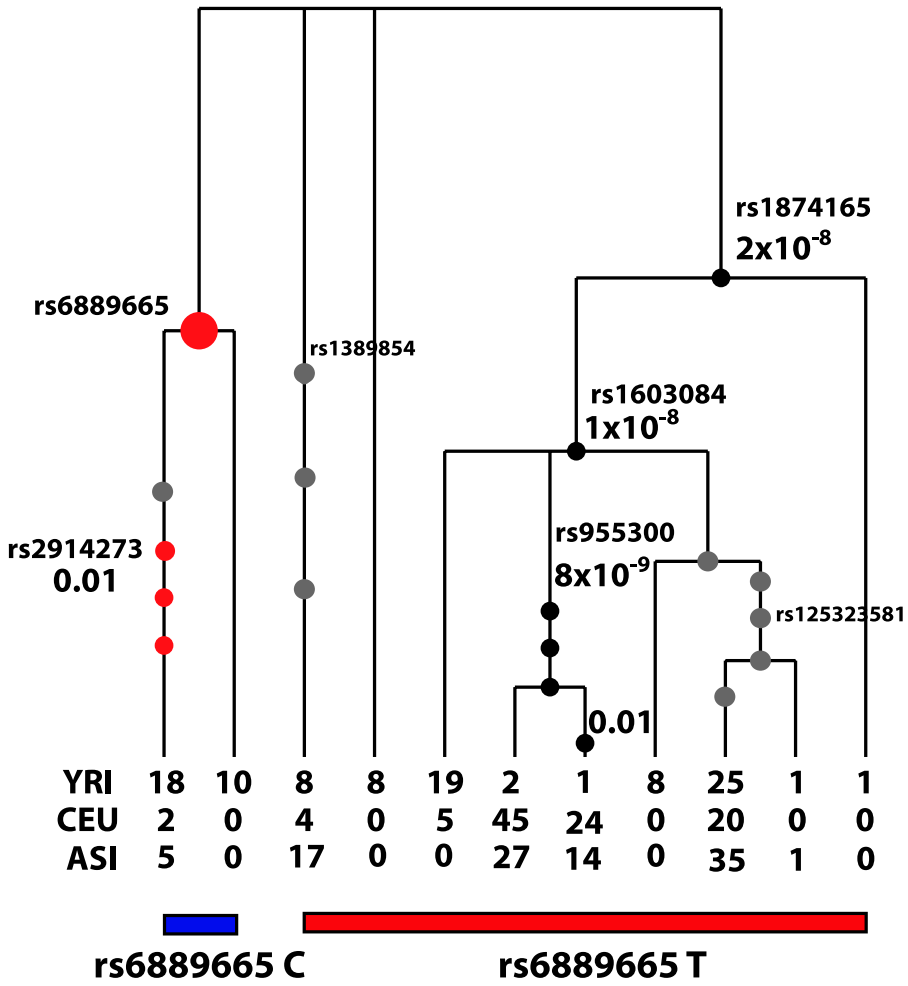


Figure 5.15: Gene tree of SNPs in the *PRDM9* LD block. SNPs significantly associated with African-enrichment conditional on rs6889665 are shown with black circles, along with their P-values. Only one SNP per haplotype is named for clarity. rs955300 tags two haplotypes which are present at a much higher frequency in Europeans than in Africans and Asians (details in text). rs2914273 tags alleles C and L4, while it places L6 and L14 under the other branch under rs6889665 (detailes in text).

not been possible to test this hypothesis because of the rarity of the L20 allele in the 1000 Genomes Pilot Data (no samples were inferred to carry this allele). Another possibility is that these haplotypes tag the B-allele whose frequency estimates vary in the literature between 3% and 5% in both African and European populations [Baudat et al., 2010; Berg et al., 2011]. The B-allele has 13 zinc fingers and a predicted 8-match to the motif, however previous studies have found significant differences in the hotspot usage phenotype between AA and AB individuals [Baudat et al., 2010], with AB individuals showing greater hotspot usage on average than AA individuals. However, this does not appear to be the cause for either signal rs955300 or rs1603084. I tested the correlation of these SNPs with inferred alleles in 1000 Genomes Pilot data. The more common allele of rs1603084, which is associated with more European recombination, has strong evidence of *not* being in high LD with the B allele ($P< 10^{-10}$), and being more strongly associated with the A allele (directionality is the same for rs955300). This suggests that the effect may be signalling one of the branches not under either of rs6889665 and rs1874165 (Figure 5.15). And indeed, conditional on rs6889665 and rs1874165, the branch under rs1389854 is weakly associated with more European-like recombination (P=0.03). This suggests, by elimination, that the final branch, not well tagged by any SNPs in HapMap2 and at 8% frequency in YRI, but absent in the other populations, may bear *PRDM9* alleles that have more African-specific recombination.

Finally, I analyzed the signal at rs2914273. I tested the association of this SNP with predicted *PRDM9* alleles in 1000 Genomes Pilot Data, conditional on rs6889665. rs2914273 is positively correlated with the C allele ($P < 10^{-7}$), and the L4 allele ($P < 10^{-3}$) and negatively correlated with the presence of the L6 allele ($P < 10^{-16}$), or the L14 allele ($P < 10^{-16}$). Alleles C, L4, L6 and L14 are all tagged by rs6889665 and have a 5/8 match with the Myers motif. The terminal six zinc fingers (Figure 5.2) differ in one zinc finger, however, with the L6 and L14 alleles sharing the

change. This suggests that even subtle changes in the array, although they are not bioinformatically predicted to change the binding preferences of PRDM9, do in fact alter the recombination landscape.

These analyses confirm that the AE phenotype is extremely sensitive and is able to pick up the impact of even rare alleles on the recombination landscape of African Americans.

5.4.4 Dominance relationship between *PRDM9* alleles

Carriers of the *PRDM9* I allele, in heterozygous state with the A allele, have been observed to have a 70% drop in the fraction of recombination in LD-based hotspots relative to individuals carrying the AA alleles [Baudat et al., 2010]. The strong decrease suggests that the I allele is out-competing the A allele in determining crossover locations.

A similar result has been observed in F1 mice crossed between inbred strains 9R and 13R, which have different *Prdm9* alleles [Brick et al., 2012]. They found that although the vast majority of hotspots in the F1 mice coincided with hotspots in the parental strains, 75% corresponded to 13R hotspots, while 22% corresponded to those of 9R. They also noted that hotspots in the F1 pool which were derived from the 13R strain were also significantly stronger than those of 9R origin.

Ségurel et al. [2011] re-analysed differences in the activity of several hotspots in men homozygous and heterozygous for A-like alleles (8/8 match to the motif) and C-like alleles (5/8 match to the motif), i.e, in men who were AA, AC, or CC at *PRDM9*. Crossovers in these hotspots were originally identified and quantified via sperm typing [Berg et al., 2010, 2011]. Ségurel et al. [2011] suggested that C-like alleles were partially dominant over A-like alleles. However, we note that significant inter-individual and inter-locus differences are evident in the dominance behavior in that analysis, suggesting a potential role for the local DNA context in stabilizing and

promoting PRDM9 binding.

I analyzed the AE and hotspot usage phenotypes relative to *PRDM9* motif classes A-like and C-like respectively by using the rs6889665 as a proxy. Our data are also suggestive of a partial dominance effect of C-type alleles over A-type alleles (when measured using the AE phenotype), even though the noisier hotspot usage phenotype is compatible with a model of co-dominant alleles (Figure 5.16).

5.5 Assessing the impact of *PRDM9* variation on recombination

The absence of any variants other than those near *PRDM9* in the association scan for AE (Figure 5.7) despite considerable power to detect correlations suggests that *PRDM9* plays a predominant role in localizing hotspots in humans. I quantitatively assess how much of the variation in crossover localization in families is explained by *PRDM9* variation. I exclude unrelated individuals in this analysis due to large statistical noise in the assessment of their phenotype, as discussed in Section 5.3.2.1. I further restrict to those individuals in families whose alleles at rs6889665 had been genotyped (as opposed to imputed), leaving 158 pedigree parents.

I implemented a linear regression model with three variables, allowing for different mean AE estimates for individuals carrying 0, 1, and 2 copies of the derived allele of rs6889665. This model explained approximately 66% of the variation in the AE phenotype. What is of greater biological interest, however, is the fraction of systematic variation explained, i.e., after excluding the noise in estimating the AE phenotype. Therefore, I estimated the noise in measuring the AE phenotype using jackknife on the set of crossovers for each individual. I used the mean of the variances in the AE phenotype across individuals as the estimate of variance solely due to measurement noise. When I subtract variance due to noise from both the total AE variance, and

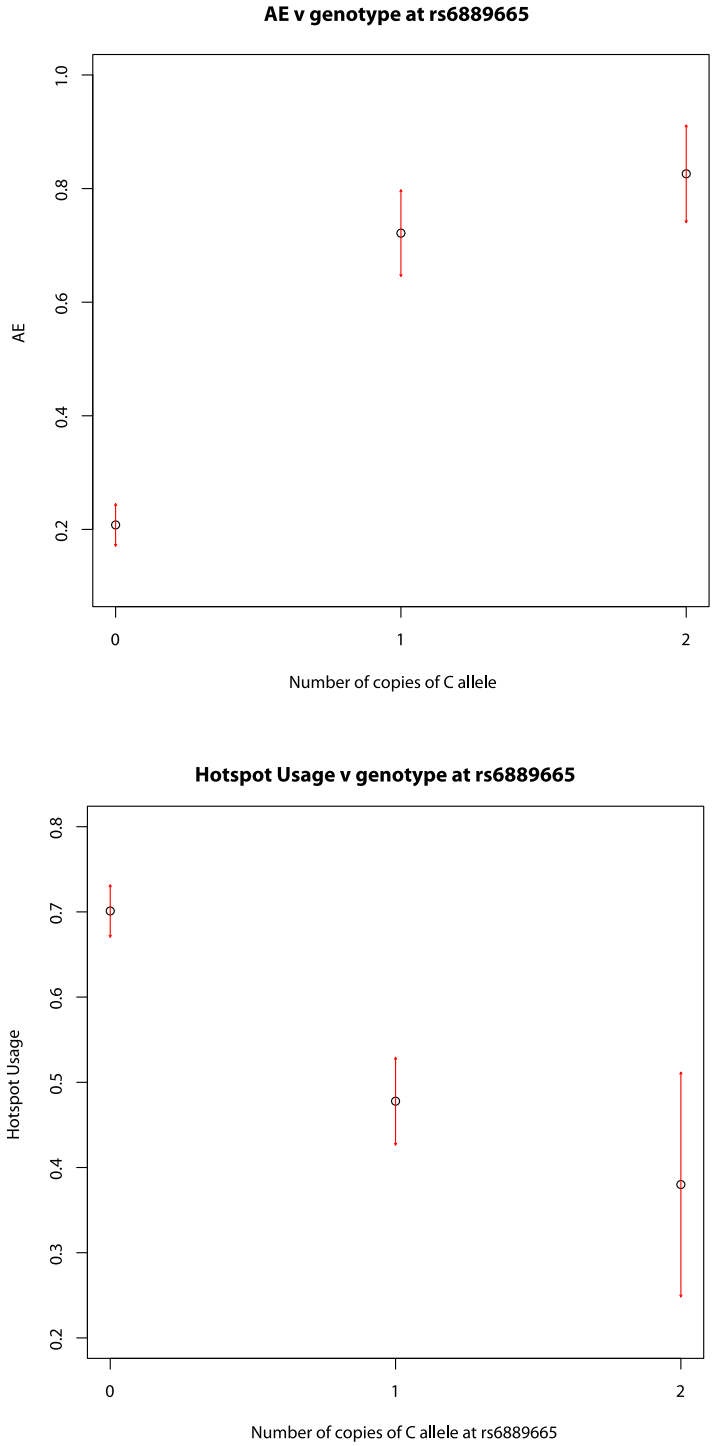


Figure 5.16: Dominance relationship between A-like and C-like *PRDM9* alleles in African-American pedigrees. The points show the mean phenotype across all pedigree parents genotyped at rs6888665, with the arrows showing two standard deviations around the mean. Due to the greater vulnerability of the hotspot usage phenotype to crossover resolution, only events resolved within 18kb were used for that phenotype. AE phenotype suggests partial dominance of C-type alleles over A-type alleles. While consistent with partial dominance, the hotspot usage phenotype is also compatible with a model of co-dominance.

from the residual variance after the effect of rs6889665 is taken into account, I estimate that rs6889665 alone accounts for 82% of the underlying phenotypic variability. This increases to 87% when both rs6889665 and rs10043097 are included in the model. There are further influential *PRDM9* variants (Section 5.4.3), which suggests that this gene may explain almost all differences in hotspot-scale recombination between African and European populations.

To understand this further, I examine if individuals who are homozygous in the derived allele at rs6889665 (CC) show any evidence of hotspots that are active in Europeans. The C allele is rare in Europeans ($f = 0.02$). The presence of European hotspots in CC individuals would therefore show that hotspots are shared between individuals with very different *PRDM9* alleles, suggesting additional factors may position hotspots. I examined rates around narrowly defined crossovers (< 10 kb) in 7 CC, 51 CT and 102 TT individuals. I find that there is no evidence of hotspots at crossover loci (encompassing 82 sites spread across the genome) in CC individuals, in either the deCODE or CEU maps (Figure 5.17), in contrast to crossovers in individuals carrying a T allele at rs6889665. Thus, crossover positions in individuals who are homozygous for the derived allele at rs6889665 are consistent with an entirely different hotspot landscape, implying that *PRDM9* controls all hotspots.

5.6 Discussion

In this chapter, I used the AA map to identify hotspots that are active in African ancestry populations, but not in Europeans. I developed a genome-wide association testing approach to search for variants influencing the usage of hotspots genome-wide in different individuals. Together with my colleagues, I performed follow-up studies to understand the biological causes of these differences. Here, I summarize briefly the findings in this chapter, in the context of the questions raised in the introduction to

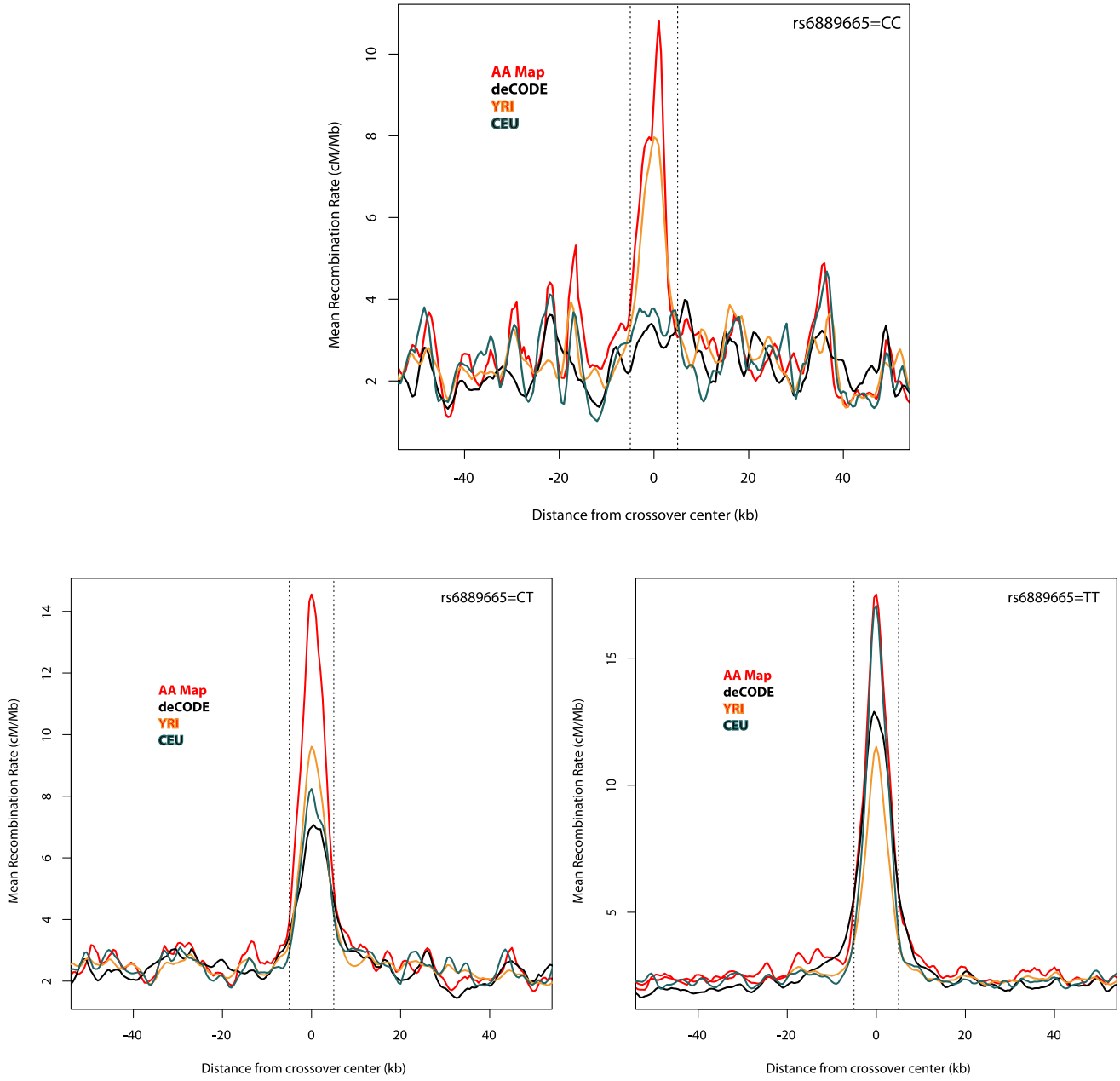


Figure 5.17: PRDM9 controls all recombination hotspots. (Top) In seven rs6889665 CC individuals I localized 82 crossovers to within 10 kb, and plotted average AA, YRI, deCODE and CEU map rates. There is no strong peak above local background in the deCODE or CEU maps. (Bottom left) Rates are plotted around 544 recombination events from 51 individuals heterozygous at rs6889665. (Bottom right) Rates around 1,009 events from 102 individuals who are homozygous in the major allele T. Rates in the European maps of deCODE and CEU show a significantly stronger peak in TT individuals than in CT individuals.

this chapter (Section 5.1):

- *The fraction of hotspots controlled by PRDM9.* I used crossovers I identified in African American pedigrees to ask if individuals homozygous for *PRDM9* alleles activating different recombination maps, also use hotspots from the *other* map. I find no evidence, using scores of finely resolved crossovers, that individuals carrying African-specific alleles with a 5/8 match to the Myers motif, have crossovers in hotspots that are active in Europeans. This indicates that *PRDM9* controls nearly all hotspots, at least in the autosomes.

- *Relationship between PRDM9 allelic types and the DNA sequences they bind.* We find a degenerate motif that is enriched in African-specific hotspots. This motif closely matches the predicted binding motif for the most common African-specific allele (allele C). This indicates, for C-like alleles, as had been shown before for the A allele, that PRDM9 acts via binding preferred DNA motifs.

  I find that *PRDM9* activity in positioning crossovers is highly sensitive to its zinc finger array, with multiple variants, often fairly rare variants, leaving an impact on the recombination landscape. Further, I find that, even subtle changes in the zinc finger array, *within* an allelic class with the same motif prediction, often change at least a subset of hotspots.

- *Penetrance of motifs and the role of genomic context.* We find that, unlike the Myers motif, the African-specific motif does not have elevated rates in THE1 or L2 elements. However, it has elevated activity in other repeat elements, such as L1PA10 and L1PA13 elements (data not shown in this report. Details can be found in Hinch et al. [2011]).

  Why both motifs show strongly elevated recombination rate in specific repeat elements, despite increased risk of deleterious non-allelic deletions and duplications therein, remains an open research question. The hypothesis that repeat

elements gain an advantage in spreading through the genome by piggybacking on recombination machinery seems unlikely [McVean, 2010]. That is because the strongest increase in rate in the presence of the Myers motif occurs in THE1B elements, which are inactive, while LINE1 elements show suppression of rate with the (older) Myers motif, and an elevation of rate (in some cases) with the African-specific motif. A potentially more likely explanation, proposed by McVean [2010] is that the evolution of new PRDM9 binding motifs (Section 1.4.5) may be strongly constrained, for example, to motifs that are abundant, spread fairly widely throughout the genome, and do not have critical functions, such as being transcription factor binding sites. Short repeat sequences, including families of now inactive transposable elements, may provide a rich source for such motifs.

Further, the differential increase in recombination rate in particular repeat elements by the Myers and African-specific motifs suggests that there may be an element of chance in the particular choice of repeat family that happens to be enriched for, and is made hot by particular motifs.

- *Interaction of PRDM9 variants with each other.* Our data suggest partial dominance of C-like alleles over the A-allele. It is worth noting that the C-like alleles are also associated with an increase in genome-wide recombination rate. In other words, an individual heterozygous for A and C-like alleles may have more crossovers in African-specific hotspots simply because C-like alleles may have a greater propensity to lead to crossovers. In that case, the overall landscape of recombination will be biased towards the African-specific map. Therefore, it is a possibility that, the "dominance" effect is simply due to the effect on genome-wide rate, and will be a subject for future work (with more pedigree data).

- *Conservation and evolution of the PRDM9 gene. PRDM9* alleles with a predicted '5' match to the Myers motif, are a set of at least diverse 10 alleles, containing anywhere between 13 and 18 zinc fingers. Nevertheless, they are well tagged by a single SNP rs6889665. This implies a common evolutionary origin for these alleles, with subsequent mutations to form multiple alleles. It is surprising and puzzling that, despite these mutations and rearrangements, only subtle differences have occurred in their binding preferences, potentially suggesting a selective constraint in their evolution.

Chapter 6

Broad-scale rates and the pseudoautosomal region

In Chapter 5, I discussed our recent findings about the control of crossover positioning at the scale of hotspots, i.e., a few kilobases. Less is known about the biological determinants of rates at broader scales, such as a few megabases. In this chapter, I investigate whether there are differences in rates in human populations genome-wide at a broad-scale and search for genetic variants that may influence them. I also investigate recombination in the human pseudoautosomal region PAR1, which is the region of homology between the X and Y chromosomes. PAR1, in males, must undergo an obligatory crossover for successful meiosis, and therefore, has a far higher broad-scale rate than anywhere else in the genome.

I present an overview of factors known to correlate with or influence broad-scale rates in Section 6.1.1. In Section 6.1.2, I discuss how PAR1 differs from the rest of the genome in its requirements for recombination. I present our current knowledge of how the obligatory crossover between X and Y chromosomes is achieved in males, and discuss a recent finding that Prdm9 may not be responsible for specifying recombination in this region in male mice [Brick et al., 2012]. In Section 6.2, I compare

rates at a broad scale (3 Mb) in the African-specific and Europeans maps estimated in Chapter 5 and report, for the first time, that recombination rates are highly conserved between populations at this scale despite large differences in hotspot locations. I also search for genetic variants that underly differences between the maps at multiple scales from 1 kb to 10 Mb, in Section 6.2. In Section 6.3, I build pedigree-based genetic maps in both males and females, and use them to investigate if PRDM9 plays a role in PAR1 recombination.

6.1 Introduction

6.1.1 Broad-scale Rates

Research in organisms from yeast to humans has shown that recombination is under many levels of control [Allers and Lichten, 2001; Paigen et al., 2008; Paigen and Petkov, 2010; Kong et al., 2010; Pan et al., 2011]. For instance, comparison of maps between HS mice and B6×CAST mice suggests megabase scale conservation of rates despite little, if any, sharing of hotspots [Paigen et al., 2008]. At the same time, the distribution of crossovers is very different between males and females at these scales, despite the fact that completely sex-specific hotspots are rare [Kong et al., 2010]. In humans, males have very elevated recombination rates in the telomeric regions relative to females while the reverse is true near the centromeres [Broman et al., 1998; Kong et al., 2002; Paigen et al., 2008].

One of the clearest influences on rates at the chromosomal scale is chromatin organization during meiosis, as evidenced by the relationship between physical size of chromosomal axes and the crossover rate (Section 1.2.4). An epigenetic mechanism that may explain greater telomeric rates in males is the existence of meiosis-specific, testis-specific histone variants that bind telomeric regions [Govin et al., 2004]. Testis-specific variants have been identified for H3, H3A, H2B and H1, of which certain H2B

variants bind specifically to telomeres in a sequence-specific manner [Tost, 2008]. The testis-specific linker histone H1t has the lowest condensing capacity relative to other H1 variants and regions with abundant H1t correlate with those most accessible to DNase I [Tost, 2008]. However, it is not known whether this mechanism has evolved to enhance recombination in male telomeres (possibly to achieve the obligate crossover in the PAR, see below), or if elevated recombination is simply a consequence of more open chromatin and longer axes that may have evolved for entirely separate reasons.

A likely connection between crossover rate and chromatin organization is also visible in a region of constitutively different chromatin in humans and chimpanzees [Auton et al., 2012]. Chimpanzees and humans have a strong correlation in rates at megabase scales, with the striking exception of chromosome 2. Chromosome 2 arose from a telomeric fusion event in the human lineage and exists as two separate chromosomes 2a and 2b in chimpanzees [IJdo et al., 1991]. Auton et al. [2012] found that while the sub- telomeric regions of chromosomes 2a and 2b in chimpanzees show high recombination rates, the rate over the syntenic region in humans is suppressed nearly three-fold. This results in a 20% reduction in the crossover rate of the human chromosome 2 as a whole.

Auton et al. [2012] further note that structural variants can influence broad-scale rates. Inverted regions between humans and chimpanzees show a lower correlation in rate than non-inverted regions, despite there being no systematic change in mean rate in those regions. The relationship between inversions and recombination has been documented extensively in *Drosophila*, while the behaviour in mammals is less well understood. Nevertheless, genetic maps compiled from the offspring of heterozygotes for certain inversions show large changes in *S. cerevisiae*, *Drosophila* and mice [Sturtevant and Beadle, 1936; Dresser et al., 1994; Gorlov and Borodin, 1995]. These changes include dramatic reduction in the crossover rate near and within inversions, sometimes accompanied by a large increase in the frequency of distal exchanges [Koehler

et al., 2004; Stevison et al., 2011]. The reduction in recombination in inversions is due to at least two reasons. First, chiasmata are partially inhibited by the inversion loop required for homologous pairing. The second effect, which is thought to be the dominant one, is the failure to recover many recombinant chromosomes. This is because crossover in both paracentric and pericentric inversions generally (but not always) leads to non-transmissible or lethal changes [Sturtevant and Beadle, 1936; Koehler et al., 2002b]. Reduction in recombination is not complete, however, due to double crossovers in large inversions [Levine, 1956] and due to gene conversion [Chovnick, 1973].

Recombination rates are correlated with several sequence parameters, notably, CpG motif fraction (positive correlation), GC content (positive correlation)[1] and poly(A)/poly(T) tract fraction (negative correlation) [Kong et al., 2002; Blat et al., 2002; Fullerton et al., 2001; Spencer et al., 2006; Pan et al., 2011; Auton et al., 2012]. While recombination rates are elevated in gene rich regions in humans, they are locally reduced in transcribed domains [Myers et al., 2005; Kong et al., 2010]. Recombination rates are also positively correlated with sequence diversity in humans [Nachman, 2001] and with divergence between human and chimpanzee [Hellmann et al., 2003]. A plausible explanation is the correlation of both mutation and recombination with GC content in this way: Recombination→increased GC content→mutation (this has not been experimentally verified and remains controversial) [Fullerton et al., 2001; Webster and Smith, 2004]. Human and chimp polymorphism data suggest that the link between recombination and GC content may be a bias towards GC bases in the mismatch repair process of DSBs [Meunier and Duret, 2004; Auton et al., 2012], while the greater mutability of CpG dinucleotides may explain the elevated mutation rate in high GC regions [Nachman and Crowell, 2000].

[1]Kong et al. [2002] suggest that the positive correlation between recombination and GC content is driven entirely by CpG fraction. When both CpG fraction and GC content were included as regressors in a multiple regression model to predict recombination, GC content was negatively associated with recombination.

6.1.2 Pseudoautosomal Recombination

The region of the human genome with the highest broad-scale rate by far is the Pseudoautosomal Region (PAR1). This is because successful progression through meiosis is dependent upon synapsis and recombination between homologous chromosomes in most sexually reproducing organisms (Chapter 1). In the heterogametic sex of most vertebrates, the non-homologous sex chromosomes must synapse and recombine to ensure proper disjunction. This obligatory crossover takes place in the Pseudoautosomal Regions, which are short regions of homology between the X and Y (or Z and W) chromosomes.

In humans, the PAR is comprised of two regions, PAR1 and PAR2, which are at the tips of Xp/Yp and Xq/Yq respectively. PAR1 is about 2.7 Mb long and is the site of the obligate male crossover. This leads to PAR1 having a crossover rate about 20 times the genome average in males, and it is over four times hotter than any other region of comparable size, while having a rate comparable to the genome average in females [Page et al., 1987; Henke et al., 1993; Flaquer et al., 2009]. Human PAR1 is homologous to the pseudoautosomal regions in great apes and Old World monkeys [Blaschke and Rappold, 2006], even though the gene content is variable even among them [Graves, 1998]. Mouse and human PARs are completely non-homologous: rodents appear to have lost the distal 9 Mb portion of Xp, and instead have a different, considerably shorter PAR1, which in mouse spans only 720 kb [Blaschke and Rappold, 2006].

PAR2 is much smaller at approximately 330 kb, and is specific to the human lineage [Kvaløy et al., 1994]. Crossovers in PAR2 are rare in both sexes (at a rate consistent with the genome average) and there do not appear to be any significant gender differences in the crossover rate [Flaquer et al., 2009]. We restrict our attention to the more evolutionarily significant PAR1 region in the rest of this work.

Despite the importance of PAR1 in male fertility, it is not yet understood how its

high rate of recombination is achieved biologically. X-Y pairing is more challenging than autosomal pairing as it cannot be mediated by DNA interactions throughout the length of the chromosomes. The unsynapsed regions outside the PARs undergo transcriptional inactivation, MSCI (meiotic sex chromosome inactivation). This is manifest in the remodelling of the unpaired XY chromatin into heterochromatin, forming what is known as the XY body, or sex body, as a visibly distinct domain within spermatocytes in pachytene during Prophase I.

Recent research in mice [Kauppi et al., 2011] suggests that recombination machinery specific to the PAR may be active in males. They found that synapsis and pairing-facilitating DSBs occur significantly later in the PAR than in the autosomes. PAR axes are disproportionately long relative to DNA length, which may facilitate DSB formation (Section 1.2.4). Most significantly, they noted that a different mRNA splicing isoform of Spo11 may be responsible for DSBs on the PAR relative to the autosomes. $Spo11\beta$ is expressed early in meiosis when most DSBs are formed and appears to be sufficient for successful pairing, crossing-over and synapsis in the autosomes in males and for all chromosomes and full meiotic progression in females. $Spo11\alpha$ isoform lacks exon 2 and is expressed later in meiosis, concomitant with PAR DSBs. In the absence of $Spo11\alpha$, approximately 70% of spermatocytes failed to synapse between the X and Y chromosomes and few post-meiotic cells are formed.

Working with PAR1 is challenging. Although less than 1% of the euchromatic sequence of the X chromosome remains to be determined, only about 85% of the PAR sequence is available to date (estimate based on Feb 2009 Human reference sequence GRCh37). This is caused by five gaps in the assembly, and is likely caused by a variety of structural properties that distinguish PAR1 from the rest of the X chromosome [Ross et al., 2005]. PAR1 has a significantly higher GC content than the rest of the X chromosome (48% versus 39%), an excess that is more pronounced than similar patterns noted in other sub-telomeric regions [Blaschke and Rappold,

2006]. In addition to this, PAR1 has a high proportion of minisatellite and short tandem repeats and various other duplicated structures. There is a particular excess of Alu repeats (29% vs 8%) though the pattern is reversed for L1 repeats (7% vs 29%) [Blaschke and Rappold, 2006].

Several low resolution genetic maps have been built to date in PAR1. However, the use of small sample sizes and sometimes very small (< 10) number of markers means that they often disagree on the PAR1 recombinational landscape [Flaquer et al., 2009]. The maps published so far include pedigree-based analyses such as Flaquer et al. [2009] including 245 meioses typed at 22 polymorphic markers, Kong et al. [2004b] with 2000 meioses typed at 6 markers (Rutgers map), sperm typing studies such as Lien et al. [2000] including 1912 meioses from 4 men typed at 9 markers, Schmitt et al. [1994] including 903 meioses from 2 men typed at 4 markers, and the HapMap Phase 2 LD-based map including samples from European, West African and Asian samples. Apart from the HapMap2 map, all studies that reported the ancestry of their samples were based on individuals of European ancestry. These studies report an average 50-55 cM genetic length for the male PAR and approximately 4 cM for the female PAR. Lien et al. [2000] discovered significant inter-individual differences in regional rate among the 4 men tested. The HapMap2 combined LD-based map is by far the most detailed map available for this region. Its use has been complicated by concerns that, due to the very high rate of recombination in this region, rate estimates may be inaccurate due to saturation from complete LD breakdown. We address this concern below in Section 6.3.

Prior to May et al. [2002], it was not known whether the PAR exhibited the autosomal pattern of intense hotspots interspersed between recombinationally inert regions. Using sperm-typing, they identified the *SHOX* hotspot, which was the hottest hotspot identified at that time, and established that it was similar to autosomal hotspots in terms of size, hotspot width and symmetry of crossovers around the centre of the

hotspot. The HapMap LD-based maps also confirms the presence of extremely hot hotspots throughout the PAR. However, only a few of loci have an estimated recombination rate that is lower than the genome-wide average rate, with little evidence for truly cold regions anywhere in the PAR. That this is possibly a genuine feature of PAR recombination, and not an artefact of HapMap2 SNP resolution, is suggested by Sarbajna [2012]. They genotype three regions 5-10 kb away from approximate boundaries of the *SHOX* hotspot at a high resolution (mean interval between SNPs $\approx$ 225 bp), and observe that recombination rates estimated by sperm typing in those intervals vary between 5 cM/Mb and 12 cM/Mb. This is compared with the 0.04 cM/Mb rate estimated by the same laboratory for comparable inter-hotspot intervals on the autosomes.

A particular mystery concerning recombination in PAR1 is the role of the gene *PRDM9*, if any. The work of Kauppi et al. [2011] regarding Spo11-mediated programmed DSBs in the PAR (discussed above) raises the possibility of entirely different recombination machinery operating in the PAR, evolved to ensure the obligate crossover. If *PRDM9* is responsible for PAR hotspots, it is important to understand how recombinogenic regions are maintained in the face of motif destruction due to biased gene conversion (Section 1.4.5). Recent work in mice [Brick et al., 2012] suggests that hotspots independent of *PRDM9* may be active in the PAR, unlike the autosomes. However, only about 40 kb of the mouse PAR is sequenced, and most of it was covered by a cluster of intense and interdigitated hotspots in all the mouse strains tested in that study, making it difficult to be certain that they are genuinely shared, as opposed to possibly overlapping.

Further, *Prdm9* exhibits complex epistatic interactions resulting in F1 hybrid male sterility in mice carrying particular combinations of *Prdm9* allele [Mihola et al., 2009]. Both sterile hybrids (in their most severe form, with all epistatically interacting loci) and $Prdm9^{-/-}$ mice exhibit failure in sex-body formation [Mihola et al., 2009]. In

fact, *Prdm9* is the only known speciation gene in vertebrates. Other known speciation loci in mice, which are not fully understood, are at least two loci in or near the PAR itself (*Sxa* or sex chromosome association locus and the *Hst3* locus) and three tightly linked loci in the meiotic drive t-haplotype[2] [Forejt, 1996; White et al., 2012]. However, whether there is any direct connection between *Prdm9* hybrid sterility and recombination in the PAR is not known.

6.2 Broad-scale rates

In Chapter 4, I noted that broad-scale rates are highly correlated between populations despite significant fine-scale differences. However, since the majority of *PRDM9* alleles in African populations are A-like, similar to Europeans (Figure 5.3), we may wonder how much of the broad-scale conservation is simply due to shared alleles. Therefore, here I compare broad-scale rates in the African-specific and Shared maps estimated in Section 5.3.1. The MCMC procedure sampled these two maps, a map that represents crossover locations in individuals from European populations (Shared), and one that represents crossovers unlikely in a European map (African-specific).

To assess how well the two maps are separated, I calculate the correlation of the maps at the 10kb scale, which is approximately the resolution of these maps. The correlation is 0.58 which, as expected, is significantly lower than the correlation between the deCODE and AA maps (0.75). Further, I note from the analysis in Section 5.4.3 that the African-specific map is able to pick up even rare crossovers that are unlikely among Europeans. However, my analysis suggests that the converse is not entirely true, and that the African-specific map contains some hotspots that are common to both populations. I suspect that this is because our samples come from an admixed population with few, if any, samples that present truly extreme cases of

[2] As it happens, the t-haplotype includes the *Prdm9* gene. Whether, there is a connection between the meiotic drive exhibited by the t-haplotype and *Prdm9* is not known.

all or most crossovers arising from an African-specific recombination landscape.

Nevertheless, I note that the correlation at 3Mb is only slightly reduced to 0.96 from 0.98 (Figure 6.1), even though the fine-scale correlation was lowered from 0.75 to 0.58 after separating the maps. This suggests that rates are conserved at this scale (about fifty to a hundred hotspots), which is quite small considering that there are thousands of hotspots per chromosome. At this stage, it is not clear if this conservation operates at the stage of programmed DSBs or the choice of recombination resolution pathway. DSB hotspots reported recently in strains of mice with different *PRDM9* alleles [Brick et al., 2012] may help illuminate this question.

6.2.1 Genome-wide association testing for broad-scale African-enrichment phenotypes

I next test for association between genetic variants and differences in broad-scale rates. Even though broad-scale rates are visually highly conserved genome-wide (Figure 6.1), I hypothesized that there may be subtle differences that may illuminate the underlying biology of broad-scale control. I repeated the genome-wide association analysis for the African-enrichment phenotype described in Section 5.3.2, but with the following differences:

- I performed association testing at four fixed scales ranging from 2 Mb to 10 Mb. New recombination phenotypes were estimated for each unrelated individual at each scale as follows:
  - the African-specific and European maps and the PDF for each crossover were linearly interpolated to the corresponding scale before estimating the phenotype, as described in Section 5.3.2.
  - Separate phenotypes were calculated for each chromosome for each individual, so that when I tested a SNP on chromosome 1, the phenotype

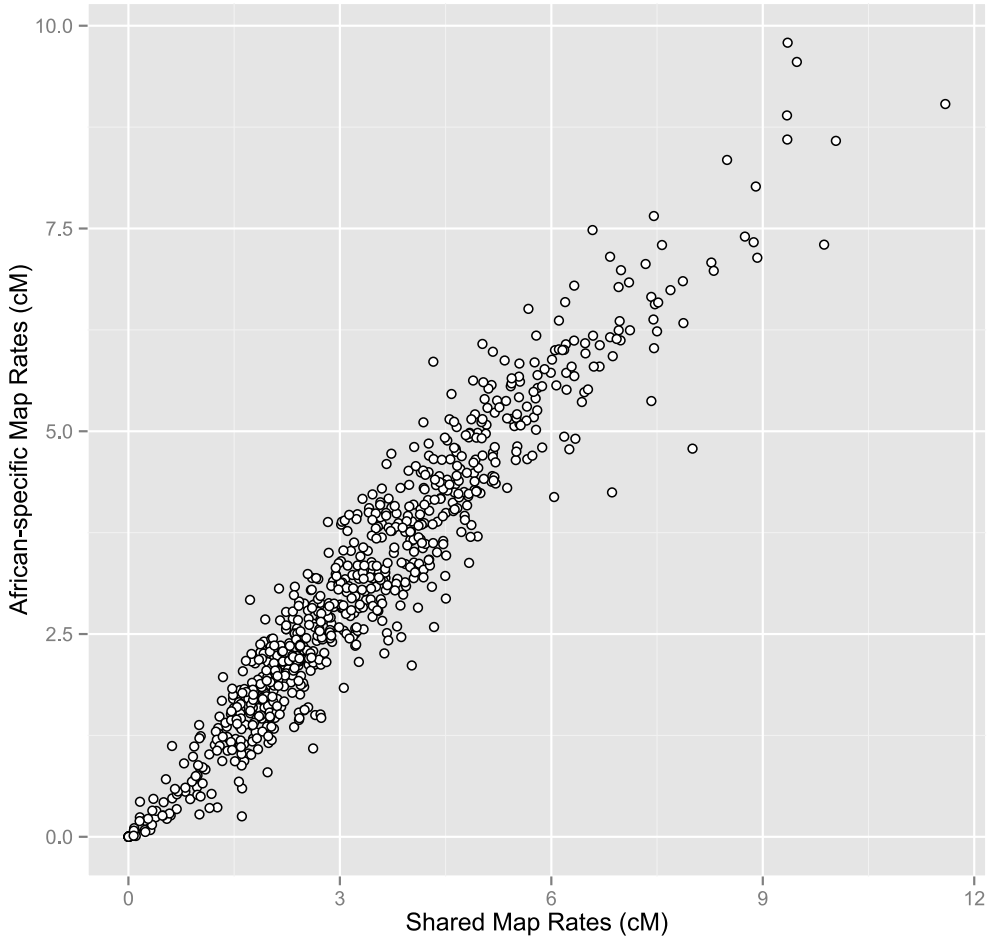


Figure 6.1: Rates in African-specific and Shared Maps at the 3Mb scale are in close agreement. Rates in the 6 Mb nearest to the telomeres were not included due to insufficient data for accurate rate estimation in those regions in both the underlying maps (the AA and deCODE maps). (Pearson correlation coefficient = 0.96)

for each individual included crossovers on all chromosomes *except* chromosome 1 itself. In other words, I restricted the testing to variants that would influence the phenotype in *trans*.

The testing of effects in *cis* is more challenging, partly due to decreased power since fewer ancestry switches are available in any particular region to estimate phenotypes of interest. A further subtlety is that the size of African and European ancestry blocks is systematically different in African-American individuals, since they have approximately 4 times more African than European ancestry on average (Figure 2.7). This is likely to induce correlations between population allele frequencies of SNPs and the ability to detect ancestry switch crossovers (for example, a SNP which has a high frequency in Africans and a low frequency in Europeans, would be correlated with a smaller number of crossovers, all else being equal. This is because a crossover in the vicinity of this SNP is more likely to be between two blocks of African ancestry, which would render the crossover invisible to our detection method). This would, in turn, lead to false-positive associations if not accounted for carefully. Finally, positive associations with *cis*-testing may be enriched for difficult to analyse regions of the genomes or regions where the HAPMIX model assumptions deviate significantly from the truth, such as long tandem repeats and inversions. Testing for *cis*-effects is something I would like to address in future work.

- When testing a SNP, both genome-wide ancestry and the mean ancestry genome-wide using all chromosomes except the chromosome the SNP is on are used as regressors. As discussed in detail in Section 5.3.2, two separate effects lead to a correlation between the phenotype and ancestry. I recapitulate those points in brief here: genome-wide ancestry may be correlated with the causal SNP in ancestors in whom the crossovers happened, and thus be informative about the

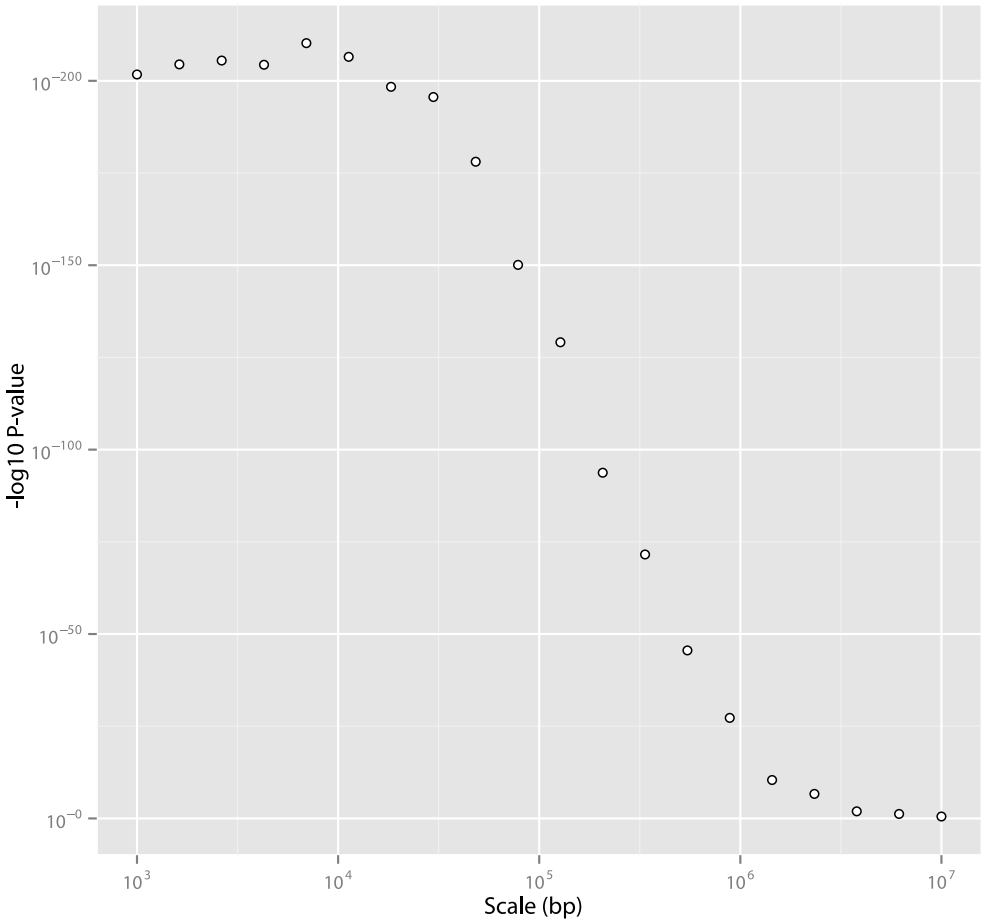
phenotype. This requires genome-wide ancestry to be included in the testing. Second, the phenotype is more difficult to estimate in individuals with high levels of African ancestry, possibly due to greater challenges in phasing. Therefore, the phenotype may be artificially correlated with the ancestry of chromosomes used to estimate the phenotype. I found that including the ancestry of chromosomes in *trans* with the tested SNP was in fact necessary to avoid P-value inflation.

Figure 6.2 shows the results of testing the association of the previous top hit SNP, rs6889665, with AE phenotypes ranging from 1 kb to 10 Mb. As reported in Section 5.4, this SNP tags C-like alleles of *PRDM9*. The P-value peaks at approximately 7 kb, which is close to the resolution of African-specific and Shared maps. This analysis implies that the impact of *PRDM9* in hotspot positioning drops off rapidly at megabase scales, but remains significant at scales as large as 4 Mb, possibly due to subtle clustering of hotspots in certain genomic regions.

I next performed a genome-wide scan at genotyped and imputed SNPs at four size scales, of approximately 2Mb, 4Mb, 6Mb and 10Mb. This constituted approximately 12 million tests, with a Bonferroni-corrected suggestive P-value of 8.3×10^{-8} (fewer than one false-positive test expected by chance) and significant P-value of 8.3×10^{-10} (FDR=0.01). Excluding SNPs within 20 Mb of rs6889665, only one SNP passed the suggestive P-value threshold: rs10516988 on chromosome 4q22.3 with P-value=1.3×10^{-8} at the 10Mb scale. The nearest gene PDHA2 is more than 250 kb away and is expressed in the mitochondrial matrix of post-meiotic spermatogenic cells. It seems not unreasonable to conclude that this is a null scan.

There may be several reasons for our inability to find broad-scale associations:

- Broad-scale rates are highly conserved, as noted above, and we may be underpowered to pick up variants that lead to subtle biological differences, if any.



7

Figure 6.2: The association of *PRDM9* variants with African-enrichment at various size scales. The association with rs6889665 is significant at the 0.01 level (after multiple-testing correction for twenty tests) at scales up to ∼4 Mb.

- This procedure relies on testing sex-averaged rates. The differences between male and female rates are large, and if genetic variants affect only one of the two sexes or act in opposite directions in different sexes, it will make the signal more difficult to identify.

- This procedure explicitly avoids testing for variants that may influence rates in *cis*, such as inversions. This will be addressed in future work.

6.3 Recombination in the Pseudoautosomal Region PAR1

In this work, I have started to build a fine-scale genetic map for PAR1, which is necessary to understand the pattern and control of recombination in this region. I first build a pedigree-based map which contains over an order of magnitude more markers than the best available map of contemporary recombination to-date [Flaquer et al., 2009]. I use it to validate the LD-based map previously built for PAR1 using HapMap2 variation data [International HapMap Consortium, 2007]. I find that these data do not yet have enough power to ascertain the role of *PRDM9* in this region, though they do give suggestive results. To increase our sample size, I am working towards building a genetic map using ancestry switches in unrelated admixed individuals, as I did for the autosomes in Chapter 2.

6.3.1 Pedigree-based genetic map for PAR1

To build this map, I use the genotype data for 220 pseudoautosomal markers from 135 nuclear families from the CARe dataset comprising 339 male and 337 female meioses (as in Chapter 3). The algorithm used for calling recombination events was similar to the algorithm in Chapter 3, but with the following differences:

- Due to the lower SNP density in the PAR, together with a much greater degree of LD breakdown, I found the approximation of unlinked loci to be adequate for this region. Therefore, I did not need to thin loci as required for Step 1 in Section 3.4. Subsequently, therefore, I did not have to perform Step 3 for finer mapping of the crossovers.

- Due to insufficient markers, we lose ability to detect crossovers in the ~ 250 kb on each end of the chromosome under consideration. There is no way to avoid

this problem in the sub-telomeric region of any chromosome, however, in this case we can partially solve the problem near the Pseudoautosomal boundary (PAB) by including genotype data from the X chromosome in the region closest to the PAB. The inheritance pattern of the X is obviously different from the PAR, and I made changes to the algorithm to reflect this difference.

138 paternal and 17 maternal crossovers were detected in total. The maps thus constructed for males and females are shown in Figure 6.3 and Figure 6.4 respectively. I note an increase in rate near the PAB, present in both males and females. The four hotspots previously identified by sperm typing also occur in regions of elevated rate in our samples [May et al., 2002; Sarbajna et al., 2012].

I next compare the rates in the male genetic map with re-scaled rates in the HapMap2 LD-based map for PAR1 [International HapMap Consortium, 2007]. The latest publicly available LD-based map for this region was published in release#21 in NCBI Build 35. I used the liftOver tool [Hinrichs et al., 2006] to transform the map to NCBI Build 36 coordinates, and make it comparable with our map. Despite the expectation of saturation of rate estimates in the LD-based map and the relatively low resolution of our pedigree-based map, the rates show reasonably good agreement (Figure 6.5). The rank correlations are 0.66 and 0.71 at scales of 30 kb and 100 kb respectively. This validates the use of the LD-based map to aid further fine-scale analysis.

To explore the role of *PRDM9* in recombination in PAR1, I plot the average rate in the LD-based map (the resolution of the pedigree-based map is not fine enough for this analysis) around all copies of the degenerate 13-mer Myers motif CCnCCnTnnCCnC, including both repeat and unique DNA. If PRDM9 is marking hotspots, we would expect to see a peak centered around the motif, similar to Myers et al. [2008]. Figure 6.6 plots the rate around 655 copies of the motif within PAR1. Myers et al. [2008] noted that the penetrance of the motif was greatly increased on THE1

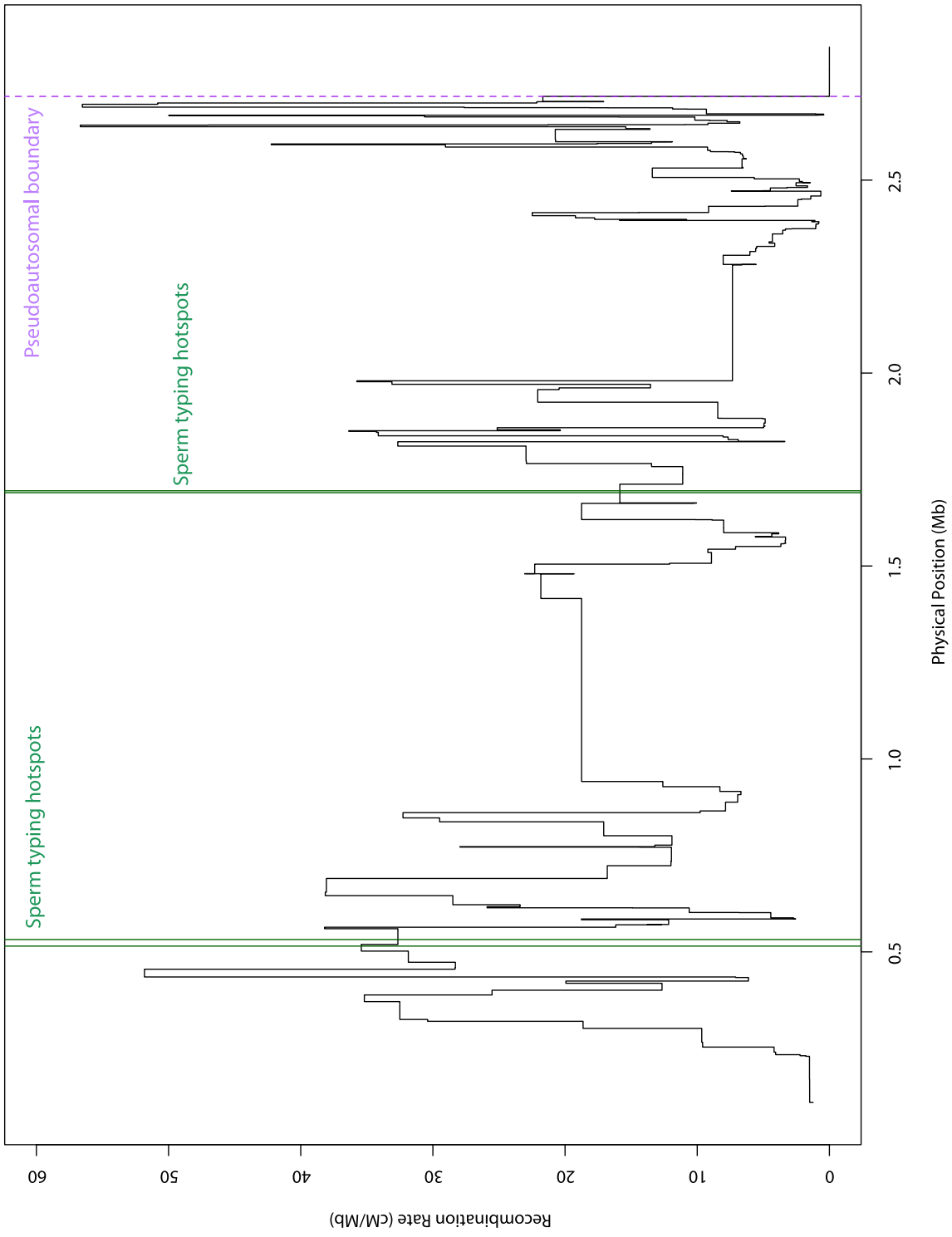


Figure 6.3: Estimated male genetic map for PAR1

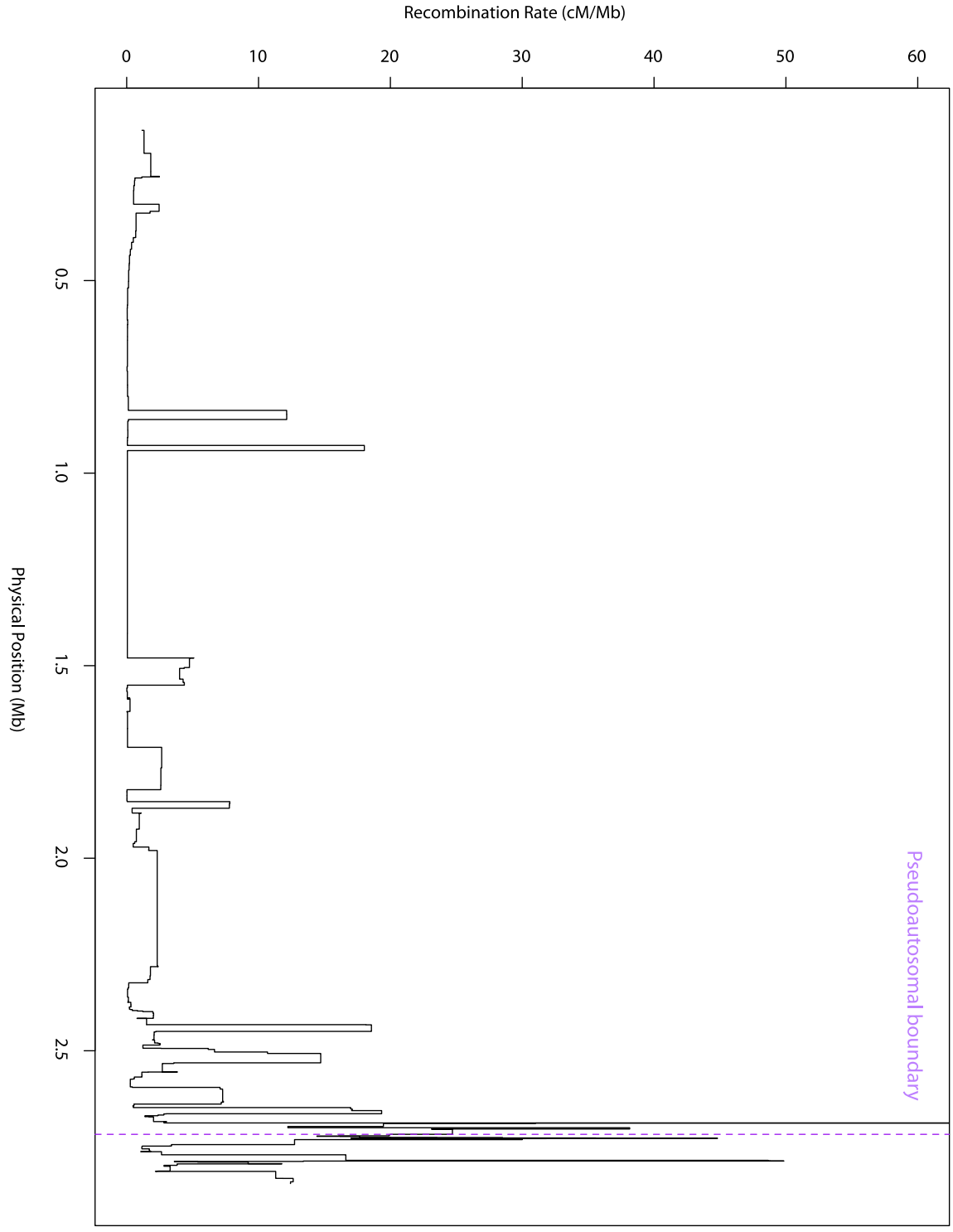


Figure 6.4: Estimated female genetic map for PAR1

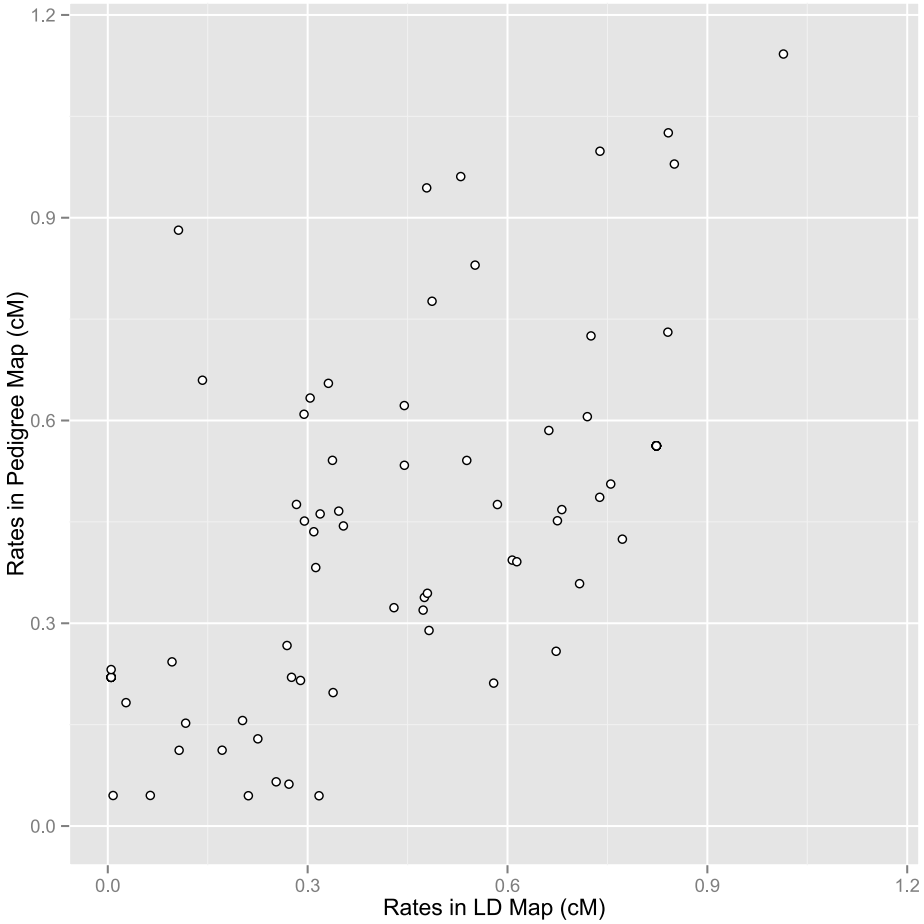


Figure 6.5: Comparison of the LD and pedigree maps at the 30 kb scale. The Pearson correlation is 0.62 and Spearman rank correlation is 0.66.

and L2 repeat backgrounds. We only have rate estimates around one copy of the motif in a THE1 element, and for 4 copies of the motif in L2 elements. Figure 6.6 also shows rate elevation around the L2 copies. The distinct peaks in both these plots close to the motif position suggest that at least some fraction of the recombination in the PAR may be due to *PRDM9* binding to its canonical target motif.

I next test whether the map of pedigree fathers with two A-type *PRDM9* alleles is statistically different than those who have at least one C-type allele, using rs6889665 as a proxy for *PRDM9* allele type. rs6889665 alleles were genotyped or could be inferred in only a subset of fathers, so this test compared the maps of 39 men homozygous for rs6889665 with 22 men who had at least one copy of the minor allele of rs6889665 (all but 1 of them were heterozygous). I asked what fraction of the crossovers in homozygous AA men occurred in the hottest 10% of PAR1 according to the map of heterozygous men and vice versa. I performed a permutation test with 1000 runs, such that for each run I permute the genotypes of the individuals, and calculate the overlaps of the two new maps based on the new genotype assignments. I ask how extreme the true overlap of the maps is relative to the permuted set to obtained P-value. In both cases, I failed to reject the null hypothesis that the two maps are the same (P-value=0.75 and P-value=0.19 respectively).

At this stage, it is not possible to say whether, given our small sample sizes, I do not have the power and resolution to distinguish between the maps, or if the maps genuinely have a very high degree of overlap. To address this issue, I am working towards an admixture based map using 18,000 of the 30,000 individuals in our dataset who have a good coverage of SNPs in PAR1. So far, this has proved extremely challenging due to relatively poor quality of data for the PAR both in our samples and the reference panels, as well as the fundamentally greater challenge of identifying ancestry blocks in a short region at the end of a chromosome, and with high LD breakdown. Finding solutions to these problems is a subject for future work.

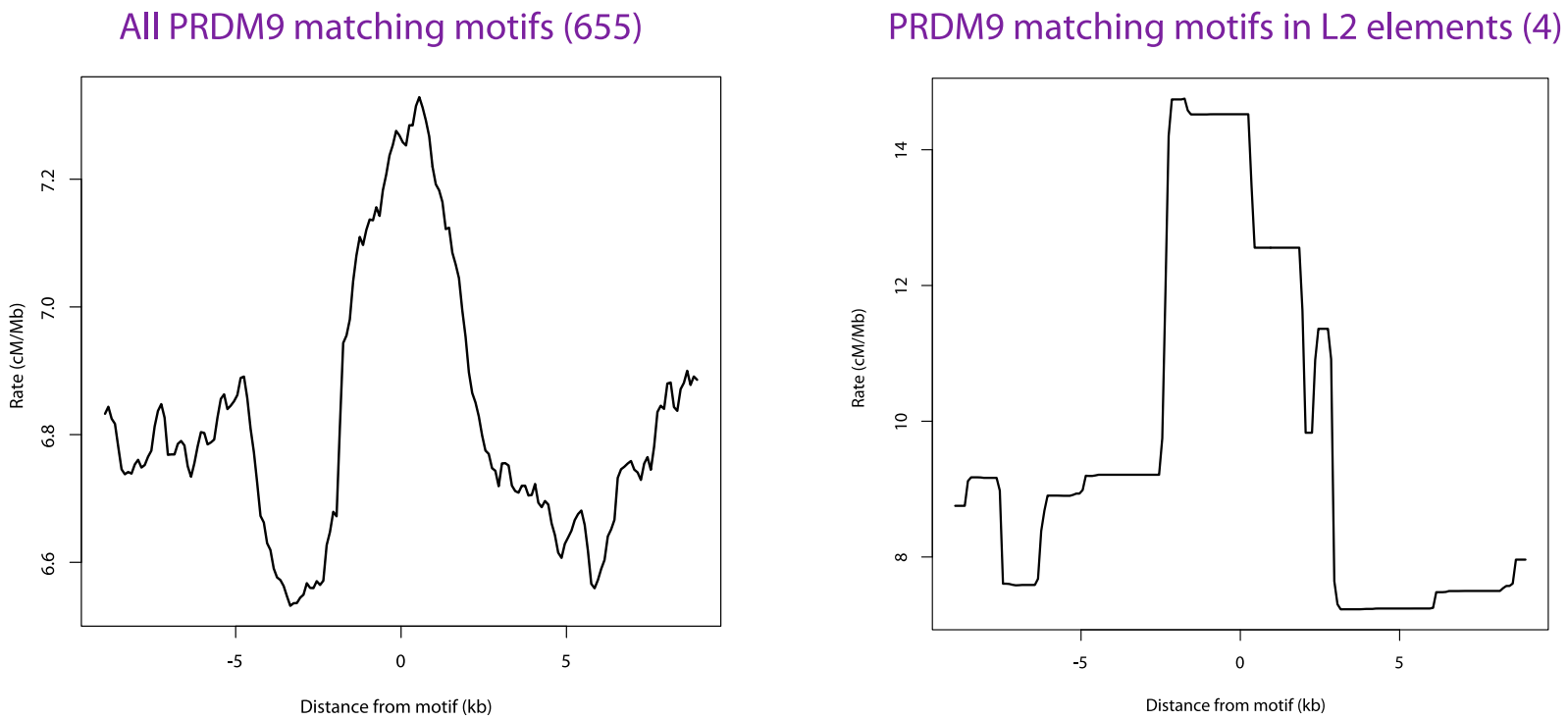


Figure 6.6: Average recombination rate in the combined HapMap2 LD-based map around (left) all occurrences of the Myers motif in repeat and unique DNA (655 copies) and (right) occurrences with L2 elements.

Chapter 7

Conclusions and Future Work

In this work, I have leveraged the particular structure of long-range LD in a recently admixed population to build a map of recombination in an understudied population. This is the first time recombination has been studied using admixture patterns and this approach presents a powerful new tool for studying crossovers in contemporary populations using only unrelated individuals. This work presents an example of a study that can be performed using only genetic or genomic data, e.g., without access to information about individuals' phenotypes or locations in case-control and epidemiological studies.

Prior to the start of my project in 2008, genetic maps were available for Europeans, using exquisite pedigree data that has been collected in the Icelandic [Kong et al., 2002] and the Hutterite [Coop et al., 2008] peoples. Comparison of recombination patterns between humans and chimpanzees has demonstrated extremely rapid evolution of the recombination landscape in the last 5-8 million years or so [Ptak et al., 2005; Winckler et al., 2005], raising the possibility that recombination may have evolved within the human lineage and that the European genetic map may not adequately represent the human spectrum. Although LD-based maps were available for both European and Yoruba populations [International HapMap Consortium, 2007], inter-

pretation of differences between the two maps was complicated by inherent noise in the estimates, plus the difference in the demographic histories and the role of natural selection in these populations, both of which are factors that are likely to influence LD patterns. This highlighted the need for a map that was largely independent of ancestral influences, and that represented a substantially differentiated human population. Building a map in African Americans, who derive the lion's share of their ancestry from West African populations, provided such a map.

The principal aim of this work was to use this map to learn more about the biology of recombination. While I have validated the properties of the AA map extensively relative to other genetic maps (Chapter 4), I did not perform detailed simulations to assess how the map building algorithm would respond to changes, for instance, in the number of samples, the size of the underlying population(s) or the number of generations since admixture. I note, in Chapter 2, that these factors are unlikely to differentially bias rate estimates. I also note that the discovery of systematic differences in human populations caused by the gene *PRDM9*, and the *de novo* identification of an African-specific motif, both speak to the precision and accuracy of the map. Nevertheless, a detailed simulation-based validation of a very similar technique can be seen in Wegmann et al. [2011], whose work was independent of and concurrent with ours. We note that Wegmann et al. [2011] also find evidence of population differentiation in recombination.

Before moving on to the biological questions addressed in this thesis, I would like to put the AA map in context with other approaches used to study recombination. There are many desirable qualities in genetic maps, not all of which are achievable at the same time. Further, depending on the goals of a study, a different choice of map may be appropriate. For instance, when picking SNPs in the design of a genotyping chip, the use of an LD-based map may be ideal, but perhaps not when comparing the likelihood of crossover inside and outside putatively functional regions in the genome.

Table 7.1 compares different approaches on several key parameters of interest.

There are several ways in which the AA map may contribute to genetics research, in addition to the fact that it is the most high resolution genome-wide crossover map in a contemporary human population. First and most obviously, the AA map highlights hotspots that are rare or absent in Europeans. The role of recombination in genome instability, especially minisatellite variability, leading to diseases is now well-established [Myers et al., 2008]. This work predicts the existence of hitherto unknown diseases in individuals of African descent due to African-specific hotspots. It provides a testable hypothesis for congenital diseases of presently unknown cause as well as motivates the search for diseases specific to non-Europeans caused by non-allelic and ectopic recombination. It is heartening to see that this is already being done [Borel et al., 2012]. Second, the map-building procedure provides estimates of uncertainty in the map, which may be useful in future population genetics research. Finally, errors in the AA map are expected to be largely uncorrelated with errors in pedigree-based and LD-based maps and therefore provide another point of reference when interpreting inferences dependent on recombination rates.

This work shows the existence of African-specific hotspots, and that differences between the recombination landscape of individuals are due to differences in their *PRDM9* zinc finger arrays. This work also shows that, at least in the autosomes, there is no evidence of shared hotspots in individuals with *PRDM9* alleles belonging to the different A-like and C-like 'motif classes' (based on number of bases matching the degenerate 13 bp Myers motif). Further, this work indicates, in agreement with sperm-typing data [Berg et al., 2011], that the motif class is not the whole story, and that individuals with subtly different *PRDM9* arrays show different propensity for recombination in at least a subset of hotspots. Despite high power to detect variants strongly influencing recombination at fine-scales, no compelling signals other than *PRDM9* were found. This suggests that either no other factors have a significant

	Sperm typing	Chip-seq assay	Pedigree map	LD-based maps	AA map
Resolution	◯	◯	∘	∘	∘
Number of meioses	◯	∘	∘	◯	◯
Genome-wide data	·	◯	◯	◯	◯
Informative about female crossovers	·	·	◯	∘	∘
Individual-level information	◯	∘	◯	·	∘
Ease/ sample availability	·	·	·	◯	∘
Recent events	◯	◯	◯	·	◯
Individual events	◯	·	◯	·	∘

Table 7.1: A qualitative comparison of the strengths and weaknesses of different approaches to learn about recombination. Big circles represent a high score, small circles a medium score and dots a low score. Methods compared are: Sperm typing [Jeffreys et al., 1998, 2001; Jeffreys and Neumann, 2002; Neumann and Jeffreys, 2006; Berg et al., 2010, 2011], Chip-seq of Rad51 and Dmc1 [Smagulova et al., 2011; Brick et al., 2012] (human data not yet published), European pedigree-based map [Coop et al., 2008; Kong et al., 2010], HapMap2 Ld-based maps [International HapMap Consortium, 2007] and this work.
Sperm-typing and Chip-seq assays give base-pair resolution of crossovers, while the other approaches have a resolution of several kilobases. All methods, except sperm typing, can be used genome-wide. Sperm-typing and Chip-seq are based on analyzing sperm or spermatogonial cells at present and are therefore only informative about males. LD-based maps and the AA map are sex-averaged, while pedigree-based approaches are informative about male and female recombination separately. LD-based maps are, by definition, population averaged and cannot be used to map individual variation in recombination phenotypes. On the other hand, they require a small number of unrelated individuals and are easiest to extend to new populations or species. Other than the Chip-seq assay and LD-based maps, events in other approaches can be directly and quantitatively interpreted as individual crossovers, although in the AA map they are not observed in a same person in whom they occurred. LD-based maps may be influenced by demography and natural selection, while the Chip-seq assay is likely to be influenced by sequencing or mapping biases and may not be interpretable quantitatively.

genome-wide effect on hotspots, or that any other factors act in a manner that is local and/or sex-specific.

Several questions regarding the evolution of *PRDM9* itself continue to puzzle. The *PRDM9* gene as a whole appears to be prone to duplication and pseudogenization in mammals, with different numbers of paralogs in mouse, cattle, and primates, and apparently only one pseudogenized copy in dogs [Oliver et al., 2009; Sandor et al., 2012]. It is the only known speciation gene in vertebrates, with crosses between particular strains of *Mus musculus musculus* and *Mus musculus domesticus* experiencing asymmetric hybrid male sterility [Mihola et al., 2009]. The mammalian *PRDM9* gene appears to be chimeric, formed from one or more gene fusion events [Pringle, 2012]. There is no apparent ortholog of the full length gene in birds, lizards or frogs, though a gene containing both the SET domain and a C2H2 zinc finger array does exist in zebrafish, possibly due to an independent gene fusion event [Pringle, 2012]. Within most mammals carrying *PRDM9* orthologs, there is strong evidence of selection for residues that determine DNA-binding properties [Oliver et al., 2009]. Within the human lineage, there is considerable variation in the number and types of zinc fingers, particularly in the C-terminal and central parts of the zinc finger array, possibly due to mismatched gene conversion events[1], while the upstream fingers are much less polymorphic. Nevertheless, this variation, particularly among alleles in the "5/8 motif match class" appears to be well-tagged by a small number of SNPs. This implies a common evolutionary origin for the 5/8 motif match alleles, with subsequent mutations to form diverse alleles.

No existing model effectively captures the constraints or the selective pressures on hotspot or *PRDM9* evolution. Loss of hotspots in the PAR or elsewhere in the genome may provide the kind of evolutionary pressure needed for rapid evolution, however, the link between crossovers and the speciation phenotype, if any, is opaque. From a

[1]There is no evidence, however, of a crossover hotspot within the gene based on our or other published maps

solely recombination initiation point of view, one might imagine that highly permissive binding would be favourable, since it would increase the number of possible binding targets and resist hotspot extinction (and creation, for that matter). That this is manifestly not the case, and that binding is both highly specific in general and fast evolving suggests that PRDM9 has an additional role, which is unknown. Another possibility is that short, less specific zinc finger arrays lead to insufficiently stable binding, and I hope to learn more about this in due course. A role for recombination has also been proposed in the control or modification of meiotic drive in female meiosis, possibly by targeting centromeric repeats [Oliver et al., 2009; Brandvain and Coop, 2012]. Finally, it remains a strong possibility that *PRDM9* has a role as a regulator of the meiotic programme since *Prdm9*$^{-/-}$ mice are infertile in both sexes, despite making adequate numbers of DSBs [Brick et al., 2012]. However, the accelerated evolution of the *PRDM9* binding sequence argues against its directly regulating more than a few genes. Nevertheless, it could explain our finding that a family of diverse alleles all have a common motif, i.e., the observation could be due to purifying selection. However, it is difficult to test this formally without more knowledge of the mutational processes occurring in the minisatellite region of *PRDM9* or the selective constraints involved.

Other than ongoing work by my colleagues and others on *PRDM9* or *Rad51* binding in cell lines and *in vivo*, sequence features across diverse taxa and the impact of recombination and gene conversion on sequence evolution within pedigrees and populations continue to provide the strongest purchase for understanding this puzzling gene. I hope to use a combination of these techniques, specifically in the PAR, to continue examining this question going forward. Sperm-based male maps and pedigree-based male maps, once the techniques are refined enough to make them comparable, may together also provide further information on the extent to which recombination influences fertility.

A limitation of this work is that the AA map is sex-averaged. Despite the large differences between male and female broad-scale patterns, the causes of these differences are almost entirely unknown. Since broad-scale rates are conserved across populations, we could leverage the sex-specific deCODE genetic maps and ask if there are any genetic variants that systematically make an individual's map more like the male map or the female map. This is a direction for future work. Mice, like humans, also appear to have greater sub-telomeric recombination in males and a more uniform distribution in females [Paigen et al., 2008]. It would be interesting to explore how this pattern manifests in mammals lacking *PRDM9* (canines) and any relationship between this pattern and the necessity to segregate faithfully after extended meiotic arrest during oogenesis. One possibility is to look in species that do not have a requirement to maintain chiasmata or sister cohesion over an extended period of time.

The strong conservation of broad-scale rates despite extreme stochasticity in individual usage of hotspots and population-level differences in hotspot locations is a particularly interesting conundrum. Since the search for genome-wide modifiers of broad-scale recombination suggested a lack of such variants commonly segregating in humans, a natural next step is to look at variants that affect rates more locally at a variety of size scales. This may help build up an understanding of how, for example, barriers to recombination are created and compensated for.

Chromatin modifications, and their role in hotspots, are now being examined by several research groups. This is, in turn, related to the complex relationship between transcription and recombination. On one hand, recombination is associated with the H3K4me2, H3K4me3 and H3K9ac marks, all of which associate with transcriptionally active domains [Buard et al., 2009; Brick et al., 2012]. Equally, research in mice has shown that *Prdm9* is directly or indirectly responsible for directing programmed double-strand breaks away from gene promoters [Brick et al., 2012]. Access to meiotic

tissue is a significant hurdle in exploring this question and a non-mammalian system with putatively functional *PRDM9* such as zebrafish may prove more amenable for exploring this question further. Another question that I think is worth exploring is the role of the 3-dimensional structure of chromosomes on recombination. Specifically, the role of the nuclear envelope in anchoring chromosomes as prelude to synapsis is known in *C. elegans* [Hawley and Gilliland, 2009] and lamina-associated domains are correlated with GC-content and transcriptional domains in humans [Guelen et al., 2008]. It would be interesting to test whether broad-scale rates may in turn be directly correlated with these domains.

Appendix A

Recombination rate estimated from linkage disequilibrium is influenced by the age of the recombination event

Mikkel Schierup, Thomas Mailund and Carston Wiuf

Unpublished result reproduced with permission from Mikkel Schierup

<mheide@birc.au.dk>

Schierup et al. investigated the effect of a single recombination event on Hudson's composite likelihood estimate [Hudson, 2001] of the scaled recombination rate as a function of the age of the event, the number of lineages remaining at the time the event occurred, on the demographic history of the population, and on properties of the sample, such as its size and number of SNPs. They simulate a large number of genealogies forward in time using a fixed parameter ρ, which is picked such that the expected number of recombination events in the simulated genealogy is one. Genealogies with exactly one recombination event were obtained using rejection sampling. Hudson's composite likelihood estimator is then calculated.

Figure A.1 shows that the estimated rate decreases monotonically back in time, in a model with no population growth, as well as in one with a population-scaled

Recombination rate estimated from linkage disequilibrium is influenced by the age of the recombination event

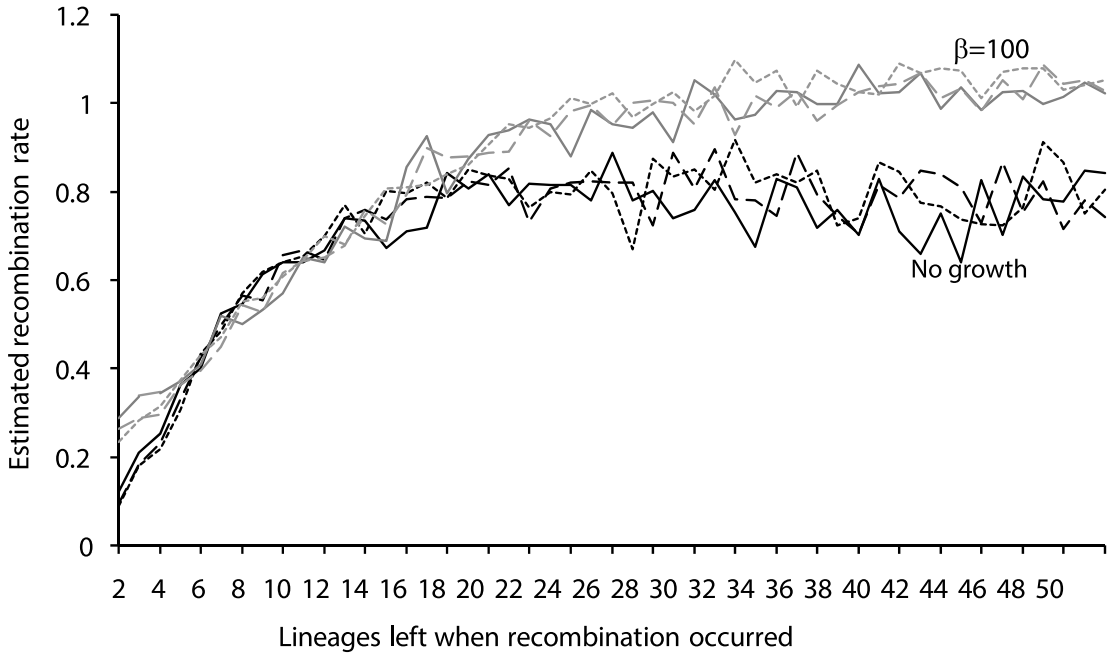


Figure A.1: Hudson's composite likelihood estimate of recombination rate as a function of the number of lineages left when the recombination event occurred in the simulated genealogy.

growth rate $\beta = N_e b = 100$.

Appendix B

Fraction of ancestry switches identical by descent in a simulated African-American population

Wegmann et al. [2011]

Figure and caption heading reproduced from Supplementary Materials (Figure 4).

Wegmann et al. [2011] performed a sophisticated simulations of African-American populations, with different effective population sizes, and a sample with 2,864 individuals. They first simulated African and European reference populations, and then performed a forward-in-time simulation with recombination of a diploid admixed population. The admixture was specified to have a single pulse of admixture 7 generations ago, with 80% African and 20% European ancestry. Further, they keep track of the chromosomal segments inherited by a sample at each stage of the simulation. They subsequently use an algorithm similar to HAPMIX (Chapter 2) to identify ancestry blocks on the simulated samples. Finally, they use this simulation framework to identify which ancestry switches were inherited identical-by-descent (IBD) by more than one sample.

Figure B.1 shows the fraction of crossovers inherited IBD by multiple samples, as a function of the effective population size.

Fraction of ancestry switches identical by descent in a simulated African-American population

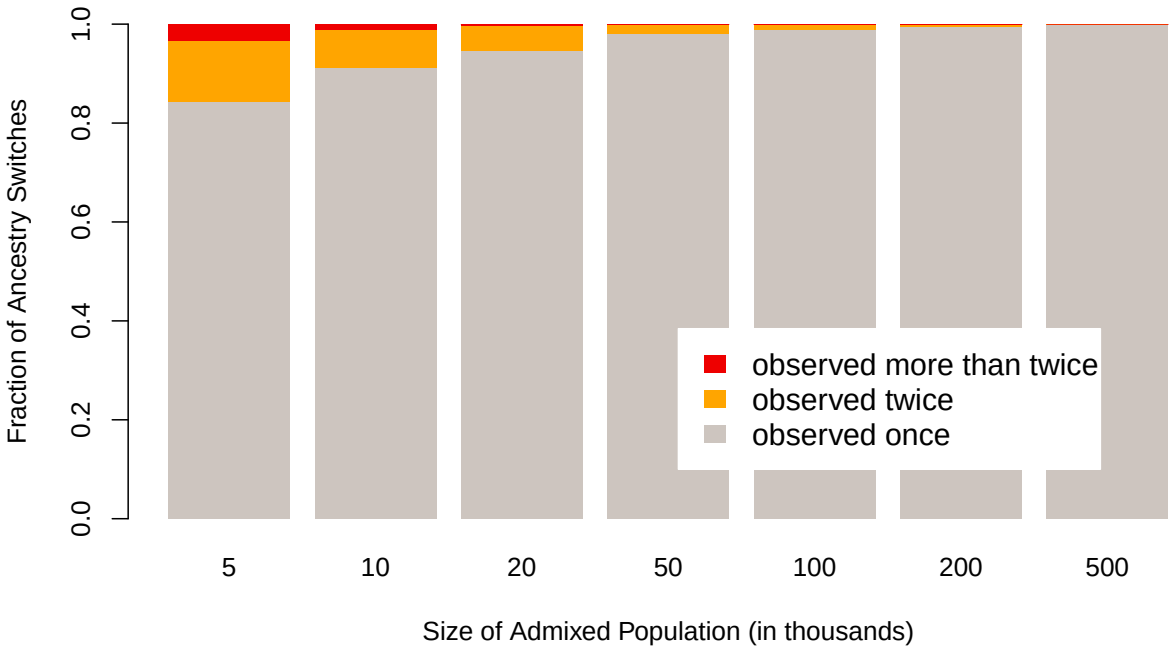


Figure B.1: **Number of times a switch point is expected to be observed in a sample of 2864 African Americans**. Wegmann et al. [2011] find that almost 89% of switches are unique in a population of only 10,000 individuals. This number increases to about 94% in a population of size 20,000 and over 99.8% in the more realistic scenario of 200,000 individuals.

Bibliography

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct. 2010.

1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature cell biology*, 491(7422):56–65, Oct. 2012.

G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30 (1):97–101, Jan. 2002.

G. R. Abecasis, D. Ghosh, and T. E. Nichols. Linkage disequilibrium: ancient history drives the new genetics. *Human Heredity*, 59(2):118–124, 2005.

S. Agarwal and G. S. Roeder. Zip3 provides a link between recombination enzymes and synaptonemal complex proteins. *Cell*, 102(2):245–255, July 2000.

E. Alani, R. A. Reenan, and R. D. Kolodner. Interaction between mismatch repair and genetic recombination in *Saccharomyces cerevisiae*. *Genetics*, 137(1):19–39, May 1994.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 5 edition, Nov. 2007.

T. Allers and M. Lichten. Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell*, 106(1):47–57, July 2001.

S. L. Andersen and J. Sekelsky. Meiotic versus mitotic recombination: Two different routes for double-strand break repair. *BioEssays*, 32(12):1058–1066, Oct. 2010.

L. K. Anderson, A. Reeves, L. M. Webb, and T. Ashley. Distribution of crossing over on mouse synaptonemal complexes using immunofluorescent localization of MLH1 protein. *Genetics*, 151(4):1569–1579, Apr. 1999.

S. Anuradha and K. Muniyappa. Molecular Aspects of Meiotic Chromosome Synapsis and Recombination. *Progress in Nucleic Acid Research and Molecular Biology*, 79: 49–132, 2005.

J. L. Argueso, J. Wanat, Z. Gemici, and E. Alani. Competing crossover pathways act during meiosis in Saccharomyces cerevisiae. *Genetics*, 168(4):1805–1816, Dec. 2004.

A. Auton. *The estimation of recombination rates from population genetic data.* PhD thesis, Oxford University Research Archive, Jan. 2007.

A. Auton, A. Fledel-Alon, S. Pfeifer, O. Venn, L. Ségurel, T. Street, E. M. Leffler, R. Bowden, I. Aneas, J. Broxholme, P. Humburg, Z. Iqbal, G. Lunter, J. Maller, R. D. Hernandez, C. Melton, A. Venkat, M. A. Nobrega, R. Bontrop, S. Myers, P. Donnelly, M. Przeworski, and G. McVean. A fine-scale chimpanzee genetic map from population sequencing. *Science (New York, NY)*, 336(6078):193–198, Apr. 2012.

E. Axelsson, M. T. Webster, A. Ratnakumar, The LUPA Consortium, C. P. Ponting, and K. Lindblad-Toh. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome research*, 22(1):51–63, Jan. 2012.

Y. Baran, B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux, C. Eng, W. Rodriguez-Cintron, R. Chapela, J. G. Ford, P. C. Avila, J. Rodriguez-Santana,

E. G. Burchard, and E. Halperin. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics (Oxford, England)*, 28(10):1359–1367, May 2012.

F. Baudat and B. de Massy. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Research*, 15 (5):565–577, 2007.

F. Baudat, K. Manova, J. P. Yuen, M. Jasin, and S. Keeney. Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Molecular cell*, 6(5):989–998, Nov. 2000.

F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science (New York, NY)*, 327(5967):836–840, Feb. 2010.

I. L. Berg, R. Neumann, K.-W. G. Lam, S. Sarbajna, L. Odenthal-Hesse, C. A. May, and A. J. Jeffreys. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics*, 42(10):859–863, Oct. 2010.

I. L. Berg, R. Neumann, S. Sarbajna, L. Odenthal-Hesse, N. J. Butler, and A. J. Jeffreys. Variants of the protein *PRDM9* differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(30): 12378–12383, July 2011.

A. Bergerat, B. de Massy, D. Gadelle, P. C. Varoutas, A. Nicolas, and P. Forterre. An atypical topoisomerase II from archaea with implications for meiotic recombination. *Nature*, 386(6623):414–417, Mar. 1997.

D. K. Bishop, D. Park, L. Xu, and N. Kleckner. *DMC1*: a meiosis-specific yeast homolog of E. coli *recA* required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell*, 69(3):439–456, May 1992.

R. J. Blaschke and G. Rappold. The pseudoautosomal regions, SHOX and disease. *Current opinion in genetics & development*, 16(3):233–239, June 2006.

Y. Blat, R. U. Protacio, N. Hunter, and N. Kleckner. Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell*, 111(6):791–802, Dec. 2002.

L. M. Boettger, R. E. Handsaker, M. C. Zody, and S. A. McCarroll. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nature Genetics*, 44(8):881–885, July 2012.

V. Borde, N. Robine, W. Lin, S. Bonfils, V. Géli, and A. Nicolas. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *The EMBO journal*, 28(2):99–111, Jan. 2009.

C. Borel, F. Cheung, H. Stewart, D. A. Koolen, C. Phillips, N. S. Thomas, P. A. Jacobs, S. Eliez, and A. J. Sharp. Evaluation of PRDM9 variation as a risk factor for recurrent genomic disorders and chromosomal non-disjunction. *Human genetics*, 131(9):1519–1524, May 2012.

G. V. Börner, N. Kleckner, and N. Hunter. Crossover/Noncrossover Differentiation, Synaptonemal Complex Formation, and Regulatory Surveillance at the Leptotene/Zygotene Transition of Meiosis. *Cell*, 117(1):29–45, Apr. 2004.

Y. Brandvain and G. Coop. Scrambling Eggs: Meiotic Drive and the Evolution of Female Recombination Rates. *Genetics*, 190(2):709–723, Feb. 2012.

K. Brick, F. Smagulova, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova. Genetic recombination is directed away from functional genomic elements in mice. *Nature*, 485(7400):642–645, May 2012.

K. W. Broman and J. L. Weber. Characterization of human crossover interference. *American journal of human genetics*, 66(6):1911–1926, June 2000.

K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American journal of human genetics*, 63(3):861–869, Sept. 1998.

K. W. Broman, L. B. Rowe, G. A. Churchill, and K. Paigen. Crossover interference in the mouse. *Genetics*, 160(3):1123–1131, Mar. 2002.

K. Bryc, A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser, S. Williams, A. Froment, J. M. Bodo, C. Wambebe, and S. A. Tishkoff. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences of the United States of America*, 107 (2):786–791, 2010.

J. Buard, P. Barthès, C. Grey, and B. de Massy. Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. *The EMBO journal*, 28(17):2616–2624, Sept. 2009.

L. Cao, E. Alani, and N. Kleckner. A pathway for generation and processing of double-strand breaks during meiotic recombination in *S. cerevisiae*. *Cell*, 61(6): 1089–1101, June 1990.

A. Chovnick. Gene conversion and transfer of genetic information within the inverted region of inversion heterozygotes. *Genetics*, 75(1):123–131, Sept. 1973.

R. Chowdhury, P. R. J. Bois, E. Feingold, S. L. Sherman, and V. G. Cheung. Genetic analysis of variation in human meiotic recombination. *PLoS genetics*, 5(9): e1000648, Sept. 2009.

C. Churchhouse and J. Marchini. Multiway Admixture Deconvolution Using Phased or Unphased Ancestral Panels. *Genetic epidemiology*, 37(1):1–12, 2013.

F. Cole, S. Keeney, and M. Jasin. Comprehensive, fine-scale dissection of homologous recombination outcomes at a hot spot in mouse meiosis. *Molecular cell*, 39(5): 700–710, Sept. 2010.

G. Coop and S. R. Myers. Live hot, die young: transmission distortion in recombination hotspots. *PLoS genetics*, 3(3):e35, Mar. 2007.

G. Coop and M. Przeworski. An evolutionary view of human recombination. *Nature reviews Genetics*, 8(1):23–34, Jan. 2007.

G. Coop, X. Wen, C. Ober, J. K. Pritchard, and M. Przeworski. High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. *Science (New York, NY)*, 319(5868):1395–1398, Mar. 2008.

F. Couteau, K. Nabeshima, A. Villeneuve, and M. Zetka. A Component of C. elegans Meiotic Chromosome Axes at the Interface of Homolog Alignment, Synapsis, Nuclear Reorganization, and Recombination. *Current Biology*, 14(7):585–592, Apr. 2004.

R. M. Dawley, J. P. Bogart, and N. Y. S. Museum. *Evolution and ecology of unisexual vertebrates*. New York State Museum, Oct. 1989.

E. de Boer and C. Heyting. The diverse roles of transverse filaments of synaptonemal complexes in meiosis. *Chromosoma*, 115(3):220–234, June 2006.

A. F. Dernburg, K. McDonald, G. Moulder, R. Barstead, M. Dresser, and A. M. Villeneuve. Meiotic recombination in C. elegans initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis. *Cell*, 94(3):387–398, Aug. 1998.

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4): 997–1004, Dec. 1999.

R. L. Diaz, A. D. Alcid, J. M. Berger, and S. Keeney. Identification of residues in yeast Spo11p critical for meiotic DNA double-strand break formation. *Molecular and cellular biology*, 22(4):1106–1115, Feb. 2002.

S. C. Dillon, X. Zhang, R. C. Trievel, and X. Cheng. The SET-domain protein superfamily: protein lysine methyltransferases. *Genome biology*, 6(8):227, 2005.

K. D. Dorfman, R. Fulconis, M. Dutreix, and J.-L. Viovy. Model of RecA-mediated homologous recognition. *Physical review letters*, 93(26 Pt 1):268102, Dec. 2004.

M. E. Dresser, D. J. Ewing, S. N. Harwell, D. Coody, and M. N. Conrad. Nonhomologous synapsis and reduced crossing over in a heterozygous paracentric inversion in Saccharomyces cerevisiae. *Genetics*, 138(3):633–647, Nov. 1994.

D. Dumas and J. Britton-Davidian. Chromosomal rearrangements and evolution of recombination: comparison of chiasma distribution patterns in standard and Robertsonian populations of the house mouse. *Genetics*, 162(3):1355–1366, Nov. 2002.

B. L. Dumont and B. A. Payseur. Genetic analysis of genome-scale recombination rate evolution in house mice. *PLoS genetics*, 7(6):e1002116, June 2011.

I. Ebersberger, D. Metzler, C. Schwarz, and S. Pääbo. Genomewide comparison of

DNA sequences between humans and chimpanzees. *American journal of human genetics*, 70(6):1490–1497, June 2002.

R. Egel. The synaptonemal complex and the distribution of meiotic recombination events. *Trends in genetics : TIG*, 11(6):206–208, June 1995.

R. C. Elston and J. Stewart. A General Model for the Genetic Analysis of Pedigree Data. *Human Heredity*, 21(6):523–542, 1971.

D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, Aug. 2003.

P. Fearnhead, R. M. Harding, J. A. Schneider, S. Myers, and P. Donnelly. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, 167(4):2067–2081, Aug. 2004.

D. O. Ferguson and W. K. Holloman. Recombinational repair of gaps in DNA is asymmetric in Ustilago maydis and can be explained by a migrating D-loop model. *Proceedings of the National Academy of Sciences of the United States of America*, 93(11):5419–5424, May 1996.

J. Feunteun. Breast cancer and genetic instability: the molecules behind the scenes. *Molecular medicine today*, 4(6):263–267, June 1998.

A. Flaquer, C. Fischer, and T. F. Wienker. A new sex-specific genetic map of the human pseudoautosomal regions (PAR1 and PAR2). *Human Heredity*, 68(3):192–200, 2009.

A. Fledel-Alon, D. J. Wilson, K. Broman, X. Wen, C. Ober, G. Coop, and M. Przeworski. Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS genetics*, 5(9):e1000658, Sept. 2009.

A. Fledel-Alon, E. M. Leffler, Y. Guan, M. Stephens, G. Coop, and M. Przeworski. Variation in human recombination rates and its genetic determinants. *PloS one*, 6 (6):e20321, 2011.

E. Folta-Stogniew, S. O'Malley, R. Gupta, K. S. Anderson, and C. M. Radding. Exchange of DNA base pairs that coincides with recognition of homology promoted by E. coli RecA protein. *Molecular cell*, 15(6):965–975, Sept. 2004.

J. Forejt. Hybrid sterility in the mouse. *Trends in genetics : TIG*, 12(10):412–417, Oct. 1996.

S. M. Fullerton, A. Bernardo Carvalho, and A. G. Clark. Local rates of recombination are positively correlated with GC content in the human genome. *Molecular biology and evolution*, 18(6):1139–1142, June 2001.

T. J. Getz, S. A. Banse, L. S. Young, A. V. Banse, J. Swanson, G. M. Wang, B. L. Browne, H. M. Foss, and F. W. Stahl. Reduced Mismatch Repair of Heteroduplexes Reveals "Non"-interfering Crossing Over in Wild-Type Saccharomyces cerevisiae. *Genetics*, 178(3):1251–1269, Feb. 2008.

S. Giglio, K. W. Broman, N. Matsumoto, V. Calvari, G. Gimelli, T. Neumann, H. Ohashi, L. Voullaire, D. Larizza, R. Giorda, J. L. Weber, D. H. Ledbetter, and O. Zuffardi. Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *American journal of human genetics*, 68(4):874–883, Apr. 2001.

G. B. Gloor, N. A. Nassif, D. M. Johnson-Schlitz, C. R. Preston, and W. R. Engels. Targeted gene replacement in Drosophila via P element-induced gap repair. *Science (New York, NY)*, 253(5024):1110–1117, Sept. 1991.

W. Goedecke, M. Eijpe, H. H. Offenberg, M. van Aalderen, and C. Heyting. Mre11

and Ku70 interact in somatic cells, but are differentially expressed in early meiosis. *Nature Genetics*, 23(2):194–198, Oct. 1999.

T. Goldfarb and M. Lichten. Frequent and efficient use of the sister chromatid for DNA double-strand break repair during budding yeast meiosis. *PLoS biology*, 8 (10):e1000520, 2010.

I. P. Gorlov and P. M. Borodin. Recombination in single and double heterozygotes for two partially overlapping inversions in chromosome 1 of the house mouse. *Heredity*, 75 (Pt 2):113–125, Aug. 1995.

J. Govin, C. Caron, C. Lestrat, S. Rousseaux, and S. Khochbin. The role of histones in chromatin remodelling during mammalian spermiogenesis. *European Journal of Biochemistry*, 271(17):3459–3469, Aug. 2004.

J. Graves. The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Human molecular genetics*, 7(13):1991–1996, Dec. 1998.

C. Grey, F. Baudat, and B. de Massy. Genome-wide control of the distribution of meiotic recombination. *PLoS biology*, 7(2):e35, Feb. 2009.

C. Grey, P. Barthès, G. Chauveau-Le Friec, F. Langa, F. Baudat, and B. de Massy. Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS biology*, 9(10): e1001176, Oct. 2011.

R. C. Griffiths and S. Tavaré. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Mathematical biosciences*, 127(1):77–98, May 1995.

L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel. Domain organization of

human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, May 2008.

H. Guillon, F. Baudat, C. Grey, R. M. Liskay, and B. de Massy. Crossover and noncrossover pathways in mouse meiosis. *Molecular cell*, 20(4):563–573, Nov. 2005.

M. A. Handel and J. C. Schimenti. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nature reviews Genetics*, Jan. 2010.

T. Hassold and P. Hunt. To err (meiotically) is human: the genesis of human aneuploidy. *Nature reviews Genetics*, 2(4):280–291, Apr. 2001.

T. Hassold, H. Hall, and P. Hunt. The origin of human aneuploidy: where we have been, where we are going. *Human molecular genetics*, 16 Spec No. 2:R203–8, Oct. 2007.

R. S. Hawley and W. D. Gilliland. Homologue pairing: getting it right. *Nature cell biology*, 11(8):917–918, Aug. 2009.

K. Hayashi, K. Yoshida, and Y. Matsui. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature*, 438(7066):374–378, Nov. 2005.

B. Heine and D. Nurse. *African Languages: An Introduction*. Cambridge University Press, 2000.

I. Hellmann, I. Ebersberger, S. E. Ptak, S. Pääbo, and M. Przeworski. A neutral explanation for the correlation of diversity with recombination rates in humans. *American journal of human genetics*, 72(6):1527–1535, June 2003.

A. Henke, C. Fischer, and G. A. Rappold. Genetic map of the human pseudoautosomal region reveals a high rate of recombination in female meiosis at the Xp telomere. *Genomics*, 18(3):478–485, Dec. 1993.

D. R. Higgs, M. A. Vickers, A. O. Wilkie, I. M. Pretorius, A. P. Jarman, and D. J. Weatherall. A review of the molecular genetics of the human alpha-globin gene cluster. *Blood*, 73(5):1081–1104, Apr. 1989.

W. G. Hill. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical population biology*, 8(2):117–126, Oct. 1975.

A. G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, C. D. Palmer, G. K. Chen, K. Wang, S. G. Buxbaum, E. L. Akylbekova, M. C. Aldrich, C. B. Ambrosone, C. Amos, E. V. Bandera, S. I. Berndt, L. Bernstein, W. J. Blot, C. H. Bock, E. Boerwinkle, Q. Cai, N. Caporaso, G. Casey, L. A. Cupples, S. L. Deming, W. R. Diver, J. Divers, M. Fornage, E. M. Gillanders, J. Glessner, C. C. Harris, J. J. Hu, S. A. Ingles, W. Isaacs, E. M. John, W. H. L. Kao, B. Keating, R. A. Kittles, L. N. Kolonel, E. Larkin, L. Le Marchand, L. H. McNeill, R. C. Millikan, A. Murphy, S. Musani, C. Neslund-Dudas, S. Nyante, G. J. Papanicolaou, M. F. Press, B. M. Psaty, A. P. Reiner, S. S. Rich, J. L. Rodriguez-Gil, J. I. Rotter, B. A. Rybicki, A. G. Schwartz, L. B. Signorello, M. Spitz, S. S. Strom, M. J. Thun, M. A. Tucker, Z. Wang, J. K. Wiencke, J. S. Witte, M. Wrensch, X. Wu, Y. Yamamura, K. A. Zanetti, W. Zheng, R. G. Ziegler, X. Zhu, S. Redline, J. N. Hirschhorn, B. E. Henderson, H. A. Taylor, A. L. Price, H. The landscape of recombination in African Americans. *Nature*, 476(7359):170–175, Aug. 2011.

A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(Database issue):D590–8, Jan. 2006.

E. R. Hoffmann. MLH1 and MSH2 Promote the Symmetry of Double-Strand Break Repair Events at the HIS4 Hotspot in Saccharomyces cerevisiae. *Genetics*, 169(3): 1291–1303, Nov. 2004.

S. Huang. The PR Domain of the Rb-binding Zinc Finger Protein RIZ1 Is a Protein Binding Interface and Is Related to the SET Domain Functioning in Chromatin-mediated Gene Expression. *Journal of Biological Chemistry*, 273(26):15933–15939, June 1998.

R. R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159 (4):1805–1817, Dec. 2001.

P. A. Hunt and T. J. Hassold. Human female meiosis: what makes a good egg go bad? *Trends in genetics : TIG*, 24(2):86–93, Feb. 2008.

N. Hunter and N. Kleckner. The single-end invasion: an asymmetric intermediate at the double-strand break to double-Holliday junction transition of meiotic recombination. *Cell*, 106(1):59–70, July 2001.

J. Hussin, M.-H. Roy-Gagnon, R. Gendron, G. Andelfinger, and P. Awadalla. Age-dependent recombination rates in human pedigrees. *PLoS genetics*, 7(9):e1002251, Sept. 2011.

R. W. Hyppa and G. R. Smith. Crossover Invariance Determined by Partner Choice for Meiotic DNA Break Repair. *Cell*, 142(2):243–255, July 2010.

N. Ihara, A. Takasuga, K. Mizoshita, H. Takeda, M. Sugimoto, Y. Mizoguchi, T. Hirano, T. Itoh, T. Watanabe, K. M. Reed, W. M. Snelling, S. M. Kappes, C. W. Beattie, G. L. Bennett, and Y. Sugimoto. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome research*, 14(10A):1987–1998, Oct. 2004.

J. W. IJdo, A. Baldini, D. C. Ward, S. T. Reeders, and R. A. Wells. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences of the United States of America*, 88(20):9051–9055, Oct. 1991.

International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, Sept. 2010.

International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Oct. 2005.

International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, Oct. 2007.

A. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29 (2):217–222, 2001.

A. J. Jeffreys and R. Neumann. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature Genetics*, 31(3):267–271, July 2002.

A. J. Jeffreys and R. Neumann. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Human molecular genetics*, 14 (15):2277–2287, Aug. 2005.

A. J. Jeffreys, J. Murray, and R. Neumann. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Molecular cell*, 2(2):267–273, Aug. 1998.

A. J. Jeffreys, J. K. Holloway, L. Kauppi, C. A. May, R. Neumann, M. T. Slingsby, and A. J. Webb. Meiotic recombination hot spots and human DNA diversity. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, 359(1441):141–152, Jan. 2004.

A. J. Jeffreys, R. Neumann, M. Panayi, S. Myers, and P. Donnelly. Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics*, 37(6):601–606, June 2005.

L. Jessop, B. Rockmill, G. S. Roeder, and M. Lichten. Meiotic chromosome synapsis-promoting proteins antagonize the anti-crossover activity of Sgs1. *PLoS genetics*, 2(9):e155, Sept. 2006.

L. B. Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome research*, 10(10):1435–1444, Oct. 2000.

E. Jorgenson, H. Tang, M. Gadde, M. Province, M. Leppert, S. Kardia, N. Schork, R. Cooper, D. C. Rao, E. Boerwinkle, and N. Risch. Ethnicity and human genetic linkage maps. *American journal of human genetics*, 76(2):276–290, Feb. 2005.

Y. S. Ju, H. Park, M.-K. Lee, J.-I. Kim, J. Sung, S.-I. Cho, and J.-S. Seo. A genome-wide Asian genetic map and ethnic comparison: the GENDISCAN study. *BMC Genomics*, 9:554, 2008.

L. Kauppi, M. Barchi, F. Baudat, P. J. Romanienko, S. Keeney, and M. Jasin. Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science (New York, NY)*, 331(6019):916–920, Feb. 2011.

S. Keeney. Mechanism and control of meiotic recombination initiation. *Current topics in developmental biology*, 52:1–53, 2001.

S. Keeney. Spo11 and the Formation of DNA Double-Strand Breaks in Meiosis. *Genome dynamics and stability*, 2:81–123, Jan. 2008.

S. Keeney, C. N. Giroux, and N. Kleckner. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*, 88 (3):375–384, Feb. 1997.

K. P. Kim, B. M. Weiner, L. Zhang, A. Jordan, J. Dekker, and N. Kleckner. Sister Cohesion and Structural Axis Components Mediate Homolog Bias of Meiotic Recombination. *Cell*, 143(6):924–937, Dec. 2010.

J. F. C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3): 235–248, 1982a.

J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982b.

Y. Kitani, L. S. Olive, and A. S. El-Ani. Genetics of *Sordaria fimicola*. V. Aberrant Segregation at the G Locus. *American Journal of Botany*, 49(7):697–706, 1962.

N. Kleckner, A. Storlazzi, and D. Zickler. Coordinate variation in meiotic pachytene SC length and total crossover/chiasma frequency under conditions of constant DNA length. *Trends in genetics : TIG*, 19(11):623–628, Nov. 2003.

F. Klein, P. Mahr, M. Galova, S. B. Buonomo, C. Michaelis, K. Nairz, and K. Nasmyth. A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. *Cell*, 98(1):91–103, July 1999.

K. E. Koehler and T. J. Hassold. Human aneuploidy: lessons from achiasmate segregation in Drosophila melanogaster. *Annals of human genetics*, 62(Pt 6):467–479, Nov. 1998.

K. E. Koehler, J. P. Cherry, A. Lynn, P. A. Hunt, and T. J. Hassold. Genetic control of mammalian meiotic recombination. I. Variation in exchange frequencies among males from inbred mouse strains. *Genetics*, 162(1):297–306, Sept. 2002a.

K. E. Koehler, E. A. Millie, J. P. Cherry, P. S. Burgoyne, E. P. Evans, P. A. Hunt, and T. J. Hassold. Sex-specific differences in meiotic chromosome segregation revealed by dicentric bridge resolution in mice. *Genetics*, 162(3):1367–1379, Nov. 2002b.

K. E. Koehler, E. A. Millie, J. P. Cherry, S. E. Schrump, and T. J. Hassold. Meiotic exchange and segregation in female mice heterozygous for paracentric inversions. *Genetics*, 166(3):1199–1214, Mar. 2004.

A. Kong, D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson. A high-resolution recombination map of the human genome. *Nature Genetics*, 31 (3):241–247, July 2002.

A. Kong, J. Barnard, D. F. Gudbjartsson, G. Thorleifsson, G. Jonsdottir, S. Sigurdardottir, B. Richardsson, J. Jonsdottir, T. Thorgeirsson, M. L. Frigge, N. E. Lamb, S. Sherman, J. R. Gulcher, and K. Stefansson. Recombination rate and reproductive success in humans. *Nature Genetics*, 36(11):1203–1206, Oct. 2004a.

A. Kong, G. Thorleifsson, H. Stefansson, G. Masson, A. Helgason, D. F. Gudbjartsson, G. M. Jonsdottir, S. A. Gudjonsson, S. Sverrisson, T. Thorlacius, A. Jonasdottir, G. A. Hardarson, S. T. Palsson, M. L. Frigge, J. R. Gulcher, U. Thorsteinsdottir, and K. Stefansson. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science (New York, NY)*, 319(5868):1398–1401, Mar. 2008.

A. Kong, G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, K. T. Kristinsson, S. A. Gudjonsson, M. L. Frigge, A. Helgason, U. Thorsteinsdottir, and K. Stefansson. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, Oct. 2010.

X. Kong, K. Murphy, T. Raj, C. He, P. S. White, and T. C. Matise. A combined

linkage-physical map of the human genome. *American journal of human genetics*, 75(6):1143–1148, Dec. 2004b.

L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American journal of human genetics*, 58(6):1347–1363, June 1996.

K. Kugou, T. Fukuda, S. Yamada, M. Ito, H. Sasanuma, S. Mori, Y. Katou, T. Itoh, K. Matsumoto, T. Shibata, K. Shirahige, and K. Ohta. Rec8 guides canonical Spo11 distribution along yeast meiotic chromosomes. *Molecular Biology of the Cell*, 20 (13):3064–3076, July 2009.

K. Kvaløy, F. Galvagni, and W. R. Brown. The sequence organization of the long arm pseudoautosomal region of the human sex chromosomes. *Human molecular genetics*, 3(5):771–778, May 1994.

E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 84(8):2363–2367, Apr. 1987.

G. M. Lathrop, J. M. Lalouel, C. Julier, and J. Ott. Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 81(11):3443–3446, June 1984.

N. M. Lawrie, C. Tease, and M. A. Hultén. Chiasma frequency, distribution and interference maps of mouse autosomes. *Chromosoma*, 104(4):308–314, Dec. 1995.

M. L. Lenzi, J. Smith, T. Snowden, M. Kim, R. Fishel, B. K. Poulos, and P. E. Cohen. Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis I in human oocytes. *American journal of human genetics*, 76(1):112–127, Jan. 2005.

R. P. Levine. Crossing over and inversions in coadapted systems. *American Naturalist*, pages 41–45, 1956.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, Dec. 2003.

Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, Dec. 2010.

S. Lien, J. Szyda, B. Schechinger, G. Rappold, and N. Arnheim. Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *American journal of human genetics*, 66(2):557–566, Feb. 2000.

S. J. Lindsay, M. Khajavi, J. R. Lupski, and M. E. Hurles. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *American journal of human genetics*, 79 (5):890–902, Nov. 2006.

A. M. Livernois, J. A. M. Graves, and P. D. Waters. The origin and evolution of vertebrate sex chromosomes and dosage compensation. *Heredity*, 108(1):50–58, Nov. 2011.

K. E. Lohmueller, C. D. Bustamante, and A. G. Clark. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics*, 182(1):217–231, May 2009.

P. E. Lovejoy. *Transformations in Slavery: A History of Slavery in Africa*. Cambridge University Press, 3rd edition, 2012.

J. R. Lupski and P. Stankiewicz. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS genetics*, 1(6):e49, Dec. 2005.

A. Lynn, K. E. Koehler, L. Judis, E. R. Chan, J. P. Cherry, S. Schwartz, A. Seftel, P. A. Hunt, and T. J. Hassold. Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science (New York, NY)*, 296(5576):2222–2225, June 2002.

D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39(Database issue):D52–7, Jan. 2011.

M. M. Mahtani and H. F. Willard. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome research*, 8(2): 100–110, Feb. 1998.

E. Mancera, R. Bourgon, A. Brozzi, W. Huber, and L. M. Steinmetz. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–485, July 2008.

E. A. Manheim and K. S. McKim. The Synaptonemal complex component C(2)M regulates meiotic crossing over in Drosophila. *Current biology : CB*, 13(4):276–285, Feb. 2003.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913, July 2007.

T. C. Matise, F. Chen, W. Chen, F. M. De La Vega, M. Hansen, C. He, F. C. L. Hyland, G. C. Kennedy, X. Kong, S. S. Murray, J. S. Ziegle, W. C. L. Stewart, and S. Buyske. A second-generation combined linkage physical map of the human genome. *Genome research*, 17(12):1783–1786, Dec. 2007.

C. May, M. Slingsby, and A. Jeffreys. Human Recombination Hotspots: Before and After the HapMap Project. *Genome Dyn Stab: Recombination and Meiosis*, 2: 195–244, 2007.

C. A. May, A. C. Shone, L. Kalaydjieva, A. Sajantila, and A. J. Jeffreys. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nature Genetics*, 31(3):272–275, June 2002.

J. P. McPherson, B. Lemmers, R. Chahwan, A. Pamidi, E. Migon, E. Matysiak-Zablocki, M. E. Moynahan, J. Essers, K. Hanada, A. Poonepalli, O. Sanchez-Sweatman, R. Khokha, R. Kanaar, M. Jasin, M. P. Hande, and R. Hakem. Involvement of mammalian Mus81 in genome integrity and tumor suppression. *Science (New York, NY)*, 304(5678):1822–1826, June 2004.

G. McVean. What drives recombination hotspots to repeat DNA in humans? *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, 365(1544):1213–1218, Apr. 2010.

G. McVean, P. Awadalla, and P. Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–1241, Mar. 2002.

G. A. T. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, 360(1459):1387–1393, July 2005.

G. A. T. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science (New York, NY)*, 304(5670):581–584, Apr. 2004.

G. McVicker and P. Green. Genomic signatures of germline gene expression. *Genome research*, 20(11):1503–1511, Nov. 2010.

D. G. Mets and B. J. Meyer. Condensins Regulate Meiotic DNA Break Distribution, thus Crossover Frequency, by Controlling Chromosome Structure. *Cell*, 139(1): 73–86, Oct. 2009.

J. Meunier and L. Duret. Recombination drives the evolution of GC-content in the human genome. *Molecular biology and evolution*, 21(6):984–990, June 2004.

O. Mihola, Z. Trachtulec, C. Vlcek, J. C. Schimenti, and J. Forejt. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science (New York, NY)*, 323 (5912):373–375, Jan. 2009.

M. B. Mitchell. Aberrant Recombination of Pyridoxine Mutants of *Neurospora*. *Proceedings of the National Academy of Sciences of the United States of America*, 41 (4):215–220, Apr. 1955.

S. Mlynarczyk-Evans and A. M. Villeneuve. Homologous Chromosome Pairing and Synapsis during Oogenesis. *Oogenesis*, page e1002231, Aug. 2010.

T. K. Mohandas, R. M. Speed, M. B. Passage, P. H. Yen, A. C. Chandley, and L. J. Shapiro. Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp. *American journal of human genetics*, 51(3):526–533, Sept. 1992.

M. A. Morelli and P. E. Cohen. Not all germ cells are created equal: aspects of sexual dimorphism in mammalian meiosis. *Reproduction*, 130(6):761–781, Dec. 2005.

S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, NY)*, 310(5746):321–324, Oct. 2005.

S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean. A common sequence

motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, 40(9):1124–1129, Sept. 2008.

S. Myers, R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, T. S. MacFie, G. McVean, and P. Donnelly. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science (New York, NY)*, 327 (5967):876–879, Feb. 2010.

M. W. Nachman. Single nucleotide polymorphisms and recombination rate in humans. *Trends in genetics : TIG*, 17(9):481–485, Sept. 2001.

M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, Sept. 2000.

M. J. Neale and S. Keeney. Clarifying the mechanics of DNA strand exchange in meiotic recombination. *Nature*, 442(7099):153–158, July 2006.

M. J. Neale, J. Pan, and S. Keeney. Endonucleolytic processing of covalent protein-linked DNA double-strand breaks. *Nature*, 436(7053):1053–1057, Aug. 2005.

R. Neumann and A. J. Jeffreys. Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. *Human molecular genetics*, 15(9):1401–1411, May 2006.

H. Niu, L. Wan, B. Baumgartner, D. Schaefer, J. Loidl, and N. M. Hollingsworth. Partner choice during meiosis is regulated by Hop1-promoted dimerization of Mek1. *Molecular Biology of the Cell*, 16(12):5804–5818, Dec. 2005.

M. Nordborg. *Coalescent Theory*. Handbook of Statistical Genetics, 3 edition, Oct. 2007.

T. Ohta and M. Kimura. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, 63(1):229–238, Sept. 1969.

T. Ohta and M. Kimura. Linkage Disequilibrium between Two Segregating Nucleotide Sites under the Steady Flux of Mutations in a Finite Population. *Genetics*, 68(4): 571, Aug. 1971.

P. L. Oliver, L. Goodstadt, J. J. Bayes, Z. Birtle, K. C. Roach, N. Phadnis, S. A. Beatson, G. Lunter, H. S. Malik, and C. P. Ponting. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS genetics*, 5(12): e1000753, Dec. 2009.

T. R. Oliver, E. Feingold, K. Yu, V. Cheung, S. Tinker, M. Yadav-Shah, N. Masse, and S. L. Sherman. New Insights into Human Nondisjunction of Chromosome 21 in Oocytes. *PLoS genetics*, 4(3):e1000033, Mar. 2008.

T. L. Orr-Weaver. Meiosis in Drosophila: seeing is believing. *Proceedings of the National Academy of Sciences of the United States of America*, 92(23):10443–10449, Nov. 1995.

F. Osman, J. Dixon, C. L. Doe, and M. C. Whitby. Generating Crossovers by Resolution of Nicked Holliday Junctions. *Molecular cell*, 12(3):761–774, Sept. 2003.

S. P. Otto and T. Lenormand. Resolving the paradox of sex and recombination. *Nature reviews Genetics*, 3(4):252–261, Apr. 2002.

S. Pääbo. The mosaic that is our genome. *Nature*, 421(6921):409–412, Jan. 2003.

D. C. Page, K. Bieker, L. G. Brown, S. Hinton, M. Leppert, J. M. Lalouel, M. Lathrop, M. Nystrom-Lahti, A. de la Chapelle, and R. White. Linkage, physical mapping, and DNA sequence analysis of pseudoautosomal loci on the human X and Y chromosomes. *Genomics*, 1(3):243–256, Nov. 1987.

S. L. Page and R. S. Hawley. The Genetics and Molecular Biology of the Synaptonemal

Complex. *Annual Review of Cell and Developmental Biology*, 20(1):525–558, Nov. 2004.

K. Paigen and P. Petkov. Mammalian recombination hot spots: properties, control and evolution. *Nature reviews Genetics*, 11(3):221–233, Mar. 2010.

K. Paigen, J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov, S. H. S. Ng, J. H. Graber, K. W. Broman, and P. M. Petkov. The Recombinational Anatomy of a Mouse Chromosome. *PLoS genetics*, 4(7):e1000119, July 2008.

J. Pan, M. Sasaki, R. Kniewel, H. Murakami, H. G. Blitzblau, S. E. Tischfield, X. Zhu, M. J. Neale, M. Jasin, N. D. Socci, A. Hochwagen, and S. Keeney. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, 144(5):719–731, Mar. 2011.

F. Pâques, W. Y. Leung, and J. E. Haber. Expansions and contractions in a tandem repeat induced by double-strand break repair. *Molecular and cellular biology*, 18 (4):2045–2054, Apr. 1998.

E. D. Parvanov, S. H. S. Ng, P. M. Petkov, and K. Paigen. Trans-regulation of mouse meiotic recombination hotspots by *Rcr1*. *PLoS biology*, 7(2):e36, Feb. 2009.

E. D. Parvanov, P. M. Petkov, and K. Paigen. *Prdm9* controls activation of mammalian recombination hotspots. *Science (New York, NY)*, 327(5967):835, Feb. 2010.

N. Patterson, N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, S. L. Hauser, M. W. Smith, S. J. O'Brien, D. Altshuler, M. J. Daly, and D. Reich. Methods for high-density admixture mapping of disease genes. *American journal of human genetics*, 74(5):979–1000, May 2004.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, Dec. 2006.

L. Pentao, C. A. Wise, A. C. Chinault, P. I. Patel, and J. R. Lupski. Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nature Genetics*, 2(4):292–300, Dec. 1992.

J. Perry, S. Palmer, A. Gabriel, and A. Ashworth. A short pseudoautosomal region in laboratory mice. *Genome research*, 11(11):1826–1832, Nov. 2001.

A. V. Persikov, R. Osada, and M. Singh. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics (Oxford, England)*, 25(1):22–29, Jan. 2009.

M. Petronczki, M. F. Siomos, and K. Nasmyth. Un ménage à quatre: the molecular biology of chromosome segregation in meiosis. *Cell*, 112(4):423–440, Feb. 2003.

C. P. Ponting. What are the genomic drivers of the rapid evolution of PRDM9? *Trends in genetics : TIG*, 27(5):165–171, May 2011.

M. Pradillo and J. L. Santos. The template choice decision in meiosis: is the sister important? *Plant Cell Reports*, 120(5):447–454, Aug. 2011.

A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, 5(6):e1000519, June 2009.

T. Pringle. PRDM9: meiosis and recombination, Apr. 2012. URL `http://genomewiki.ucsc.edu/index.php/PRDM9:_meiosis_and_recombination`.

S. E. Ptak, D. A. Hinds, K. Koehler, B. Nickel, N. Patil, D. G. Ballinger, M. Przeworski, K. A. Frazer, and S. Pääbo. Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics*, 37(4):429–434, Apr. 2005.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool

set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–575, Sept. 2007.

H. Qiao, J. K. Chen, A. Reynolds, C. Höög, M. Paddy, and N. Hunter. Interplay between synaptonemal complex, homologous recombination, and centromeres during mammalian meiosis. *PLoS genetics*, 8(6):e1002790, June 2012.

L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

T. D. Raedt, M. Stephens, I. Heyns, H. Brems, D. Thijs, L. Messiaen, K. Stephens, C. Lazaro, K. Wimmer, H. Kehrer-Sawatzki, D. Vidaud, L. Kluwe, P. Marynen, and E. Legius. Conservation of hotspots for recombination in low-copy repeats associated with the NF1 microdeletion. *Nature Genetics*, 38(12):1419–1423, Dec. 2006.

C. L. Ramirez, J. E. Foley, D. A. Wright, F. Müller-Lerch, S. H. Rahman, T. I. Cornu, R. J. Winfrey, J. D. Sander, F. Fu, J. A. Townsend, T. Cathomen, D. F. Voytas, and J. K. Joung. Unexpected failure rates for modular assembly of engineered zinc fingers. *Nature methods*, 5(5):374–375, May 2008.

J. A. Rawley and S. D. Behrendt. *The Transatlantic Slave Trade: A History, Revised Edition*. University of Nebraska Press, 2005.

B. Rockmill, J. C. Fung, S. S. Branda, and G. S. Roeder. The Sgs1 helicase regulates chromosome synapsis and meiotic crossing over. *Current biology : CB*, 13(22): 1954–1962, Nov. 2003.

P. J. Romanienko and R. D. Camerini-Otero. The mouse Spo11 gene is required for meiotic chromosome synapsis. *Molecular cell*, 6(5):975–987, Nov. 2000.

M. T. Ross, D. V. Grafham, A. J. Coffey, S. Scherer, K. McLay, D. Muzny, M. Platzer, G. R. Howell, C. Burrows, C. P. Bird, A. Frankish, F. L. Lovell, K. L. Howe, J. L. Ashurst, R. S. Fulton, R. Sudbrak, G. Wen, M. C. Jones, M. E. Hurles, T. D. Andrews, C. E. Scott, S. Searle, J. Ramser, A. Whittaker, R. Deadman, N. P. Carter, S. E. Hunt, R. Chen, A. Cree, P. Gunaratne, P. Havlak, A. Hodgson, M. L. Metzker, S. Richards, G. Scott, D. Steffen, E. Sodergren, D. A. Wheeler, K. C. Worley, R. Ainscough, K. D. Ambrose, M. A. Ansari-Lari, S. Aradhya, R. I. S. Ashwell, A. K. Babbage, C. L. Bagguley, A. Ballabio, R. Banerjee, G. E. Barker, K. F. Barlow, I. P. Barrett, K. N. Bates, D. M. Beare, H. Beasley, O. Beasley, A. Beck, G. Bethel, K. Blechschmidt, N. Brady, S. Bray-Allen, A. M. Bridgeman, A. J. Brown, M. J. Brown, D. Bonnin, E. A. Bruford, C. Buhay, P. Burch, D. Burford, J. Burgess, W. Burrill, J. Burton, J. M. Bye, C. Carder, L. Carrel, J. Chako, J. C. Chapman, D. Chavez, E. Chen, G. Chen,. The DNA sequence of the human X chromosome. *Nature*, 434(7031):325–337, Mar. 2005.

S. Rosu, D. E. Libuda, and A. M. Villeneuve. Robust crossover assurance and regulated interhomolog access maintain meiotic crossover number. *Science (New York, NY)*, 334(6060):1286–1289, Dec. 2011.

C. Sandor, W. Li, W. Coppieters, T. Druet, C. Charlier, and M. Georges. Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination in Cattle. *PLoS genetics*, 8(7):e1002854, July 2012.

S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating local ancestry in admixed populations. *American journal of human genetics*, 82(2):290–303, Feb. 2008.

S. Sarbajna. Analysis of Meiotic Recombination in the Human Pseudoautosomal Regions. *PhD Thesis*, June 2012.

S. Sarbajna, M. Denniff, A. J. Jeffreys, R. Neumann, M. Soler Artigas, A. Veselis, and C. A. May. A major recombination hotspot in the XqYq pseudoautosomal region gives new insight into processing of human gene conversion events. *Human molecular genetics*, Jan. 2012.

S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, 15 (11):1576–1583, Nov. 2005.

H. Scherthan. Telomere attachment and clustering during meiosis. *Cellular and molecular life sciences : CMLS*, 64(2):117–124, Jan. 2007.

M. Schierup, T. Mailund, and C. Wiuf. The effect of a single recombination event on estimates of the scaled recombination rate and the decay of linkage disequilibrium. *Unpublished*.

K. Schmitt, L. C. Lazzeroni, S. Foote, D. Vollrath, E. M. Fisher, T. M. Goradia, K. Lange, D. C. Page, and N. Arnheim. Multipoint linkage map of the human pseudoautosomal region, based on single-sperm typing: do double crossovers occur during male meiosis? *American journal of human genetics*, 55(3):423–430, Sept. 1994.

L. Ségurel, E. M. Leffler, and M. Przeworski. The Case of the Fickle Fingers: How the PRDM9 Zinc Finger Protein Specifies Meiotic Recombination Hotspots in Humans. *PLoS biology*, 9(12):e1001211, Dec. 2011.

A. Shinohara and M. Shinohara. Roles of RecA homologues Rad51 and Dmc1 during meiotic recombination. *Cytogenetic and genome research*, 107(3-4):201–207, 2004.

F. Smagulova, I. V. Gregoretti, K. Brick, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, 472(7343):375–378, Apr. 2011.

M. W. Smith, N. Patterson, J. A. Lautenberger, A. L. Truelove, G. J. McDonald, A. Waliszewska, B. D. Kessing, M. J. Malasky, C. Scafe, E. Le, P. L. De Jager, A. A. Mignault, Z. Yi, G. De The, M. Essex, J.-L. Sankale, J. H. Moore, K. Poku, J. P. Phair, J. J. Goedert, D. Vlahov, S. M. Williams, S. A. Tishkoff, C. A. Winkler, F. M. De La Vega, T. Woodage, J. J. Sninsky, D. A. Hafler, D. Altshuler, D. A. Gilbert, S. J. O'Brien, and D. Reich. A high-density admixture map for disease gene discovery in african americans. *American journal of human genetics*, 74(5): 1001–1013, May 2004.

T. Snowden, S. Acharya, C. Butz, M. Berardini, and R. Fishel. hMSH4-hMSH5 recognizes Holliday Junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Molecular cell*, 15(3):437–451, Aug. 2004.

C. C. A. Spencer, P. Deloukas, S. Hunt, J. Mullikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, and G. McVean. The Influence of Recombination on Human Genetic Diversity. *PLoS genetics*, 2(9):e148, Sept. 2006.

F. W. Stahl and H. M. Foss. A Two-Pathway Analysis of Meiotic Crossing Over and Gene Conversion in Saccharomyces cerevisiae. *Genetics*, 186(2):515–536, Oct. 2010.

H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir, N. Desnica, A. Hicks, A. Gylfason, D. F. Gudbjartsson, G. M. Jonsdottir, J. Sainz, K. Agnarsson, B. Birgisdottir, S. Ghosh, A. Olafsdottir, J.-B. Cazier, K. Kristjansson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, A. Kong, and K. Stefansson. A common inversion under selection in Europeans. *Nature Genetics*, 37(2):129–137, Feb. 2005.

W. W. Steiner and G. R. Smith. Natural meiotic recombination hot spots in the

Schizosaccharomyces pombe genome successfully predicted from the simple sequence motif M26. *Molecular and cellular biology*, 25(20):9054–9062, Oct. 2005.

L. S. Stevison, K. B. Hoehn, and M. A. F. Noor. Effects of inversions on within- and between-species recombination and divergence. *Genome biology and evolution*, 3: 830–841, 2011.

A. Storlazzi, S. Tessé, S. Gargano, F. James, N. Kleckner, and D. Zickler. Meiotic double-strand breaks at the interface of chromosome movement, chromosome remodeling, and reductional division. *Genes & development*, 17(21):2675–2687, Nov. 2003.

A. H. Sturtevant and G. W. Beadle. The Relations of Inversions in the X Chromosome of Drosophila Melanogaster to Crossing over and Disjunction. *Genetics*, 21(5):554–604, Sept. 1936.

H. Sun, D. Treco, N. P. Schultes, and J. W. Szostak. Double-strand breaks at an initiation site for meiotic gene conversion. *Nature*, 338(6210):87–90, Mar. 1989.

H. Sun, D. Treco, and J. W. Szostak. Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the ARG4 recombination initiation site. *Cell*, 64(6):1155–1161, Mar. 1991.

A. Sundquist, E. Fratkin, C. B. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome research*, 18(4):676–682, Apr. 2008.

J. W. Szostak, T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl. The double-strand-break repair model for recombination. *Cell*, 33(1):25–35, May 1983.

H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch. Reconstructing genetic ancestry

blocks in admixed individuals. *American journal of human genetics*, 79(1):1–12, July 2006.

C. Tease and M. A. Hultén. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenetic and genome research*, 107(3-4):208–215, 2004.

S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, and S. M. Williams. The Genetic Structure and History of Africans and African Americans. *Science (New York, NY)*, 324(5930):1035–1044, May 2009.

J. Tost. *Epigenetics*. Horizon Scientific Press, 2008.

J. E. van Veen and R. S. Hawley. Meiosis: when even two is a crowd. *Current biology : CB*, 13(21):R831–3, Oct. 2003.

J. D. Wall, L. A. Frisse, R. R. Hudson, and A. Di Rienzo. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. *American journal of human genetics*, 73(6):1330–1340, Dec. 2003.

J. Wang, H. C. Fan, B. Behr, and S. R. Quake. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell*, 150 (2):402–412, July 2012.

T.-F. Wang and W.-M. Kung. Supercomplex formation between Mlh1-Mlh3 and Sgs1-Top3 heterocomplexes in meiotic yeast cells. *Biochemical and biophysical research communications*, 296(4):949–953, Aug. 2002.

H. A. Webber, L. Howard, and S. E. Bickel. The cohesion protein ORD is required for homologue bias during meiotic recombination. *The Journal of cell biology*, 164 (6):819–829, Mar. 2004.

M. T. Webster and N. G. C. Smith. Fixation biases affecting human SNPs. *Trends in Genetics*, 20(3):122–126, Mar. 2004.

D. Wegmann, D. E. Kessner, K. R. Veeramah, R. A. Mathias, D. L. Nicolae, L. R. Yanek, Y. V. Sun, D. G. Torgerson, N. Rafaels, T. Mosley, L. C. Becker, I. Ruczinski, T. H. Beaty, S. L. R. Kardia, D. A. Meyers, K. C. Barnes, D. M. Becker, N. B. Freimer, and J. Novembre. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics*, 43(9):847–853, Aug. 2011.

S. C. West. Enzymes and Molecular Mechanisms of Genetic Recombination. *Annual Review of Biochemistry*, 61(1):603–640, June 1992.

M. A. White, M. Stubbings, B. L. Dumont, and B. A. Payseur. Genetics and evolution of hybrid male sterility in house mice. *Genetics*, 191(3):917–934, July 2012.

E. O. Wilson. *Biodiversity*. National Academy of Sciences & the Smithsonian Institution. 1988.

W. Winckler, S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald, R. E. Bontrop, G. A. T. McVean, S. B. Gabriel, D. Reich, P. Donnelly, and D. Altshuler. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science (New York, NY)*, 308(5718):107–111, Apr. 2005.

C. Wiuf. Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology*, 53(5):821–841, Sept. 2006.

H.-G. Yu and D. E. Koshland. Meiotic condensin is required for proper chromosome

compaction, SC assembly, and resolution of recombination-dependent chromosome linkages. *The Journal of cell biology*, 163(5):937–947, Dec. 2003.

F. Zakharia, A. Basu, D. Absher, T. L. Assimes, A. S. Go, M. A. Hlatky, C. Iribarren, J. W. Knowles, J. Li, B. Narasimhan, S. Sidney, A. Southwick, R. M. Myers, T. Quertermous, N. Risch, and H. Tang. Characterizing the admixed African ancestry of African Americans. *Genome biology*, 10(12):R141, 2009.

K. Zeng and B. Charlesworth. The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics*, 189(1):251–266, Sept. 2011.

D. Zickler and N. Kleckner. Meiotic chromosomes: integrating structure and function. *Annual review of genetics*, 33:603–754, 1999.