Patterns of crossing over and gene conversion in meiotic recombination

Christopher L Campbell

Candidate:

_____
Signature

Thesis Advisor:

_____
Signature

Adam Auton, D.Phil.
Assistant Professor, Department of Genetics
Assistant Professor, Department of Epidemiology & Population Health

Co-advisor:

_____
Signature

Bernice E. Morrow, Ph.D.
Professor, Department of Genetics
Professor, Department of Obstetrics & Gynecology and Women's Health
Professor, Department of Pediatrics (Cardiology)
Sidney L. and Miriam K. Olson Chair in Cardiology
Director, Division of Translational Genetics, Department of Genetics

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Graduate Division of Medical Sciences.

Albert Einstein College of Medicine
Yeshiva University
New York
May 31, 2016

ABSTRACT

Patterns of crossing over and gene conversion in meiotic recombination
Christopher L Campbell

Recombination during meiosis is an essential process to the generation of gametes. It acts to shuffle genetic variation within the genome, generating new combination of alleles that can be subject to natural selection. Failures of the chromosomes to synapse and recombine can result in aneuploidy, while improper repair of double strand breaks can cause chromosome rearrangements, both of which are highly deleterious.

Recombination has two outcomes: crossover and gene conversion. In crossover, there is an equal exchange of genetic material between homologous chromosomes, while gene conversion is the non-reciprocal transfer of genetic information between homologues that is limited to smaller intervals. Most research has focused on crossover, which varies widely in frequency and placement within the genome, and between individuals, sexes, populations, and species. This variation is controlled by multiple factors, including the protein PRDM9, which drives crossovers to concentrated areas of recombination known as hotspots. In addition, crossover interference acts to inhibit crossovers in close proximity. Crossover properties also vary greatly by sex, and it has been suggested that these change with age, a possible cause of chromosomal aneuploidy.

In order to better understand recombination in humans I have employed statistical methods to identify sex specific differences in recombination in detail. In Chapter 2, I performed a pedigree analysis using data derived from over 18,000 meioses. I find that males and females had different hotspot usage, with males having a higher hotspot overlap proportion than females, by 4.6%. I measure crossover interference, which affects the spatial positioning of crossovers, finding that older mothers had a steep increase in crossovers that escape regulation by interference. These crossovers appear closely spaced, pointing to a possible deregulation of recombination that increases with maternal age.

In Chapter 3, I present an extension to Chapter 2, which confirms the age effect using samples from older individuals. I include a re-analysis of public data from single cell sperm and oocytes, showing that interference varies widely on an individual basis.

In Chapter 4, I focused on crossover patterns using a complex pedigree of inbred domestic dogs, consisting of 408 meioses. Dogs are unique in that their PRDM9 ortholog has accumulated muta-

tions, rendering it inactive, and raising question as to how crossovers are placed without this regulatory mechanism. I found that dog recombination is broadly similar to humans: females have a higher recombination rate than males, and male recombination strongly prefers telomeric regions. Despite the absence of PRDM9, dog recombination appears to be more concentrated to a smaller proportion of sequence when compared to humans, suggesting the existence of hotspots. Finally, I show evidence for positive crossover interference acting in the dog genome with a similar mechanism to that observed in humans.

In Chapter 5, I developed a novel computational approach using a hidden Markov model to detect gene conversion events in admixed population genetic data. The model uses two divergent ancestral reference populations in order to predict the location of gene conversion events on an admixed haplotype. Simulation results show that the model can plausibly be used to detect gene conversion.

Overall, I have performed three separate but interrelated studies of recombination that expand on existing studies. First, my work expands on the current knowledge of recombination variation and provides evidence for sexual dimorphism in crossover modifying properties, including placement and frequency. Second, my work has further investigated the effects of the loss of PRDM9 on the recombination landscape in dogs, and through comparison to humans, provided insight into recombination in both species. Finally, I developed a novel computational approach for investigating gene conversion. These studies come together to enable further understanding of the properties of genetic recombination as a whole.

Contents

List of Figures

List of Tables

List of Abbreviations

B-ALL	B-cell precursor acute lymphoblastic leukemia.
BGC	Biased gene conversion.
BGC	GC biased gene conversion.
BIC	Bayesian Information Criterion.
bp	Base pairs.
CDF	Cumulative distribution function.
CEPH	Center for the Study of Human Polymorphisms. French: Centre d'Etude du Polymorphisme Humain.
CEU	Utah Residents (CEPH) with Northern and Western Ancestry.
cM	centi-Morgans.
CMT1A	Charcot-Marie-Tooth disease type 1A.
CO	Crossover.
DNA	Deoxyribonucleic acid.
DSB	Double strand breaks.
EM	Expectation maximization.
FDR	False discovery rate.
FPN	Female pronucleus.
GC	Gene conversion, also known as NCO, non-crossover.
GWAS	Genome wide association studies.
H3K4	Histone 3, lysine (K) 4.
HMM	Hidden Markov model.
HNPP	Hereditary neuropathy with liability to pressure palsies.
IVF	In vitro fertilization.
kb	Kilo-base pairs, 1,000 bases.

LCR	Low copy repeats.
LD	Linkage disequilibrium.
LS	The Li and Stephens model.
Mb	Mega-base pairs, 1,000,000 bases.
MHC	Major histocompatability complex.
MLE	Maximimum likelihood estimation.
NAHR	Nonallelic homologous recombination.
PAC	Product of approximate conditionals.
PAR	Pseudoautosomal region.
PB1	Polar body 1.
PB2	Polar body 2.
PCR	Polymerase chain reaction.
RFLP	Restriction fragment length polymorphism.
SC	Synaptonemal complex.
SNP	Single nucleotide polymorphism.
STRP	Short tandem repeat polymorphism.
TPR	True positive rate.
TSS	Transcription start site.
YRI	Yoruba in Ibadan, Nigeria (abbreviation: YRI).
ZF	Zinc finger.

Chapter 1

Introduction

Meiosis occurs in all sexually reproducing organisms, and is essential to gametogenesis. Recombination plays a key role in this process, facilitating the pairing and alignment of chromosomes, while the exchange of genetic material has important implications in inheritance, natural selection, and evolution. Double strand breaks begin the recombination process, and resolve to one of two outcomes, crossover or gene conversion (outlined in Figure 1.1). Most research has focused on genetic crossover, which is the reciprocal exchange of genetic material between homologous chromosomes during meiosis.

There are a number of factors that influence the placement of crossovers within the genome, and there is tremendous variability between individuals, sexes, and species. Recombination properties differ greatly between males and females, both in frequency and in placement within the genome. In this introductory chapter, I will review historical literature that has advanced our understanding of recombination, building up to the current knowledge of how recombination is affected by these recombination modifying factors. Much of existing research focuses on human subjects, and I have focused the majority of work within this thesis on the research of recombination in humans, with the specific goal of learning more about human recombination. I have focused on research in humans, but consider recombination across a number of other species as well.

First, I will review what is known about the biology of meiotic recombination. This includes the cellular stages that make up meiosis I and II, and how double strand breaks are initiated and resolved. Then I will outline a variety of methods that have been used to study recombination. This includes molecular methods that observe recombination in a direct fashion, as well as indirect approaches that rely on inferring crossover from genetic data from families and unrelated individuals.

I will then describe what is currently known about the various factors that act together to modify properites of recombination freqeuncy and placement. This includes the hotspot specifying protein PRDM9, which was recently identified, and influences recombination placement in humans and other mammals. Another important modifier of recombination is sex, and males and females have been shown to vary widely in their recombination properties on both broad and fine scales. In addition, a phenomenon called interference affects the spacing of bewteen crossover events on a chromosome.

Gene conversion, the non-reciprocal transfer of genetic material between chromosomes is understudied due to the difficulty of its detection. Most research so far has relied on sperm typing, which does not provide any information on gene conversion in females. I will review these, as well as

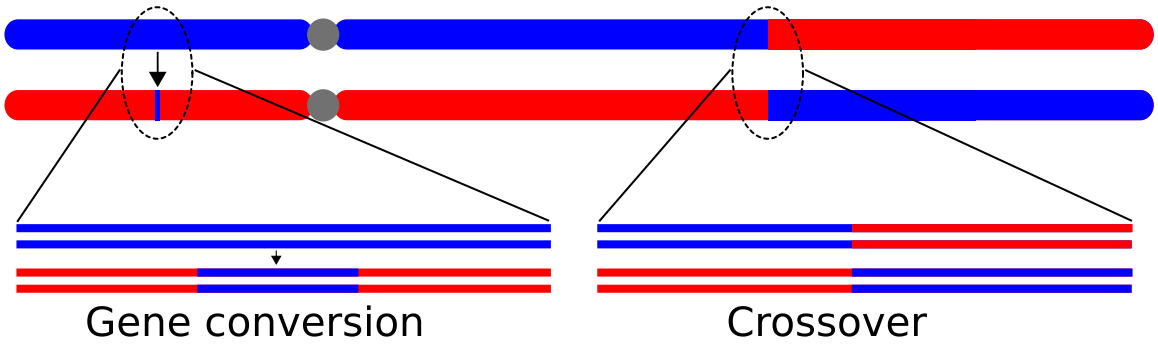

Figure 1.1: **Recombination produces crossover and gene conversion.** Recombination is initiated by DNA double strand breaks that resolve into one of two possible outcomes. Crossover (shown at right) is the large-scale, reciprocal exchange of genetic material between two chromosomes around the break point. Gene conversion (or non-crossover, shown at left) is the one-way transfer of small amounts of DNA from one chromosome to the other during the repair of the break.

genome wide methods that focus on statistical approaches to infer gene conversion events in genetic data.

In this thesis, I used a number of biological and statistical methods to gain insight into factors affecting crossover placement, and into the recombination process as a whole. I will conclude this introduction with an overview of these methods used to study both crossover and gene conversion. Specifically, this focuses on obtaining genotypes from family pedigrees followed by a hidden Markov model (HMM) approach to call crossovers. For gene conversions, I will review several previous HMMs that were used to call recombination in population genetic data.

Overall, the study of recombination has advanced greatly over the past 100 years, with substantial progress made in humans following the completion of the Human Genome Project. Although a number of advances have been made in recent years, our current understanding of the recombination process is incomplete. The research presented within this thesis provides an expanded picture of the recombination landscape as a whole.

1.1 An overview of meiotic recombination

1.1.1 The biology of meiotic recombination

Prior to meiosis, a diploid cell contains pairs of homologous chromosomes, with one of the pair inherited from the father, and the other from the mother. This diploid DNA is replicated just prior to the cell entering the meiotic cycle, in premeiotic S-phase[1] to generate exact copies of each pair of chromosomes, referred to as sister chromatids. Meiosis consists of two stages; recombination occurs in the first stage, referred to as meiosis I, while sister chromatids separate into their respective daughter cells in meiosis II. Meiosis I is the most complex and lengthy stage, with chromosome pairing, synapsis, and recombination all occurring in succession within prophase I, which is divided into several sub-stages. Meiosis II is similar to mitotic divisions and results in separation of chromosomes to haploid daughter cells.

Prophase of meiosis I has several stages which have been given individual names, shown in Figure 1.2. The stages are leptotene, zygotene, pachytene, diplotene, and diakinesis. In the first step, leptotene (derived from Greek meaning “thin threads”), changes in chromatin cause the newly replicated chromosomes to form thin individual strands. Here, the synaptonemal complex (SC), a

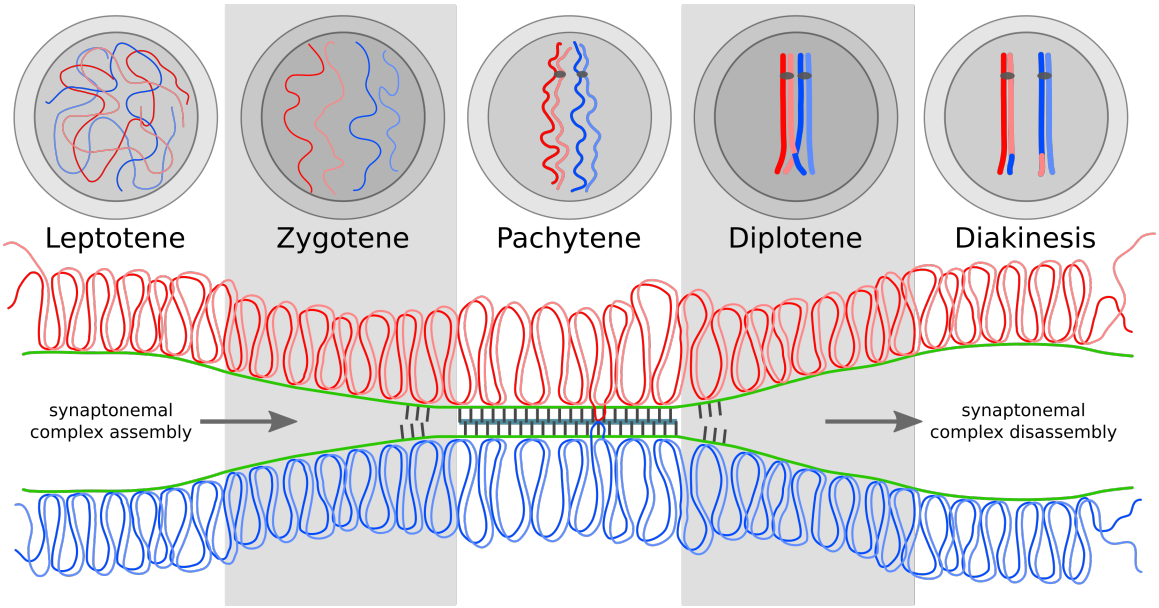

Figure 1.2: **The stages of prophase I in meiosis.** The top panel shows an overview of the five stages of prophase I, progressing from left to right. One pair of sister chromatids is shown in shades of red, and represents genetic material inherited from one parent. A second pair of sister chromatids, inherited from the other parent, is shown in shades of blue. The pairs of sister chromatids are homologous to each other and this homology aids in pairing and synapsis. The bottom panel shows the progression of the assembly and disassembly of the synaptonemal complex (SC), corresponding with the stages in the top panel. A recombination event is shown in the pachytene stage, mediated within the central region of the SC. The axial element forms the backbone of the SC (shown in green) and binds to DNA that is tightly packaged into large chromatin loops. Modified from Yang and Wang [2] and de Boer and Heyting [3].

protein structure that will bind the sister chromatids and their homologues into a tetrad begins to assemble. The SC consists of a scaffold of proteins that first forms with axial elements that associate with the already paired sister chromatids, solidifying their association.

In the zygotene phase (“paired threads”), pairing between the unraveled DNA begins to occur at regions of homology in a process known as synapsis. Homologous chromosomes connect to the SC by transverse filaments, drawing them into the SC structure in a progressive zipper-like mechanism, completing synapsis[2]. Evidence from cytological studies suggests that DNA DSBs occur at this stage[4,5].

By the pachytene stage (“thick threads”), synapsis, and the SC assembly are fully complete, and the pairs of homologous chromosomes bound within the SC are referred to as a tetrad or bivalent. Strand exchange and DSB repair occurs at this stage, mediated by the structure of the SC. Several models for strand exchange and resolution have been proposed, but one widely accepted possibility is the Szostak model. Here, the single stranded DNA from the severed chromatid invades and pairs with the double stranded homologue, which forms a double Holliday junction that is resolved in two possible ways[6]. A subset of the DSBs are processed as crossovers, at locations called chiasmata. The remaining DSBs are repaired through a different pathway, as non-crossovers, also known as gene conversion. It is proposed that two distinct enzymatic reactions occur that sever the Holliday junctions in two different ways, producing either gene conversion or crossover[7].

In the diplotene stage (“two threads”), the SC is disassembled, allowing the tetrad to relax slightly. The homologous chromosomes are still held together at chiasmata locations. In the final substage of prophase I, diakinesis (“moving through”), the chromosomes condense into visible threads.

The cellular machinery begins to prepare for cell division, which occurs in the remaining step of meiosis I, metaphase I, and anaphase I. Chiasmata, holding the chromosomes together as crossover points are digested enzymatically, allowing the homologous chromosomes to segregate to their respective cellular poles. Following this, the cell proceeds through meiosis II, which is procedurally similar to mitosis. Here, the separation of sister chromatids occurs, and four haploid gametes are produced.

1.1.2 Double strand breaks and the synaptonemal complex assembly.

Specific proteins are critically important for individual steps of meiosis I. The recombination process in meiosis I begins with programmed double strand breaks (DSBs) in the DNA, catalyzed by the protein SPO11, which has a similar function to DNA topoisomerases[8]. The protein RAD51 localizes to broken DNA and assists in strand invasion and homologous pairing of the cut strand. RAD51 has been detected in autosomes as early as late leptotene or early zygotene[4], suggesting that DSBs form early in meiosis I.

Roughly concurrent with DSB formation in late leptotene, the axial elements of the SC assemble[2]. The axial elements form the backbone of the SC, and each associates with one pair of sister chromatids. The axial elements are attached to chromatin, containing the compacted DNA of the sister chromatids in a series of loops that radiate outwards from the core axis of the SC.

Two DNA binding proteins, MSH4 and MLH1 are critically important in the next step of meiosis I. In the zygotene stage, peak levels of MSH4 foci are found. The MSH4 protein marks most DSBs and is thought to promote synapsis[4]. MLH1 foci, thought to specifically influence DSBs to be repaired as crossovers and not non-crossover[9], begin to appear in late zygotene[4]. By the end of zygotene, the paired sister chromatids in each axial element complete synapsis. The axial elements progressively join together at homologous regions, assisted by transverse elements that “zip” the structure together, and are bound in the central core by central element proteins[2].

When complete by the end of zygotene, the SC is composed of two axial elements, a central element, and a number of transverse elements[2]. The DNA is bound within this complex, with homologous pairing within the SC core, and compacted DNA within chromatin located outside the central core in large loops. The preceding evidence suggests a sequence of events in which DSBs form early in prophase I, just prior to, or concurrent with, synapsis and full assembly of the SC. The decision to repair a DSB as a crossover, as seen by the association of MLH1 and MSH4 proteins, appears to happen in conjunction with synapsis and prior to DSB resolution in pachytene, however the exact timing remains an open quesion[10].

Additional evidence from mouse studies suggests that sex chromosomes have a different timing of these events. There are two isoforms of *Spo11* in humans, and mice, and a recent study in mice suggests that they may have differing functions, with $Spo11\beta$ being expressed earlier in meiosis,

coinciding with most DSBs occurring on the autosomes. Male mice with only *Spo11*β had meiotic defects, with the majority of spermatocytes failing to recombine in the pseudoautosomal region (PAR)[11]. Following this, *Spo11*α was found to be expressed later in meiosis, and coincided with DSBs located within the sex chromosomes, including the PAR[8,11]. This evidence indicates that the initiation of DSBs is a complex, multi-stage process, with autosomal DNA processed earlier than DNA from the sex chromosomes.

1.1.3 Timing of meiotic events

The fundamental steps of meiosis are the same in males and females, but the timing of these events, both prior and during, differs significantly between the sexes[12], and even between species. In humans, male meiosis begins at puberty and continues in a cycle that lasts throughout the lifespan. The precursor cells undergo a minimum of 30 mitotic divisions prior to entering meiosis, and this number continues to rise with age, since male meiosis is continually occurring. For example, a 15 year old male is estimated to have 35 germ-cell divisions, with this number rising to 380 at age 30, and 840 by age 50[13]. As the number of mitotic proliferations increases in males, so does the number of mutations accumulated through DNA replication errors. A recent study using a chimpanzee pedigree estimated that the number of mutations rises linearly with the father's age, with approximately three additional mutations accumulating per year[14]. This contributes to a paternal age effect, with mutations accumulating on the male germ line with increasing age.

Females have 22 cell divisions prior to meiotic entry, and one during, for a total of 23 divisions[13], and this number is fixed for all oocytes. In females, meiosis begins prenatally, and oocytes progress through the diplotene stage of prophase I before undergoing an arrest period[13,15]. This arrest is called the dictyotene stage, or dictyate arrest, and meiosis is frozen at the point at which the chromosomes have fully synapsed and chiasmata have formed (Figure 1.3). This arrest period ends only upon ovulation, and thus meiosis can be potentially very lengthy, taking one to five decades to complete. Additionally, while each male meiosis produces four haploid sperm products, female meiosis yields one haploid oocyte contain the majority of the cytoplasm. The remaining meiosis I and II division products produce polar bodies, which contain DNA but typically apoptose[16].

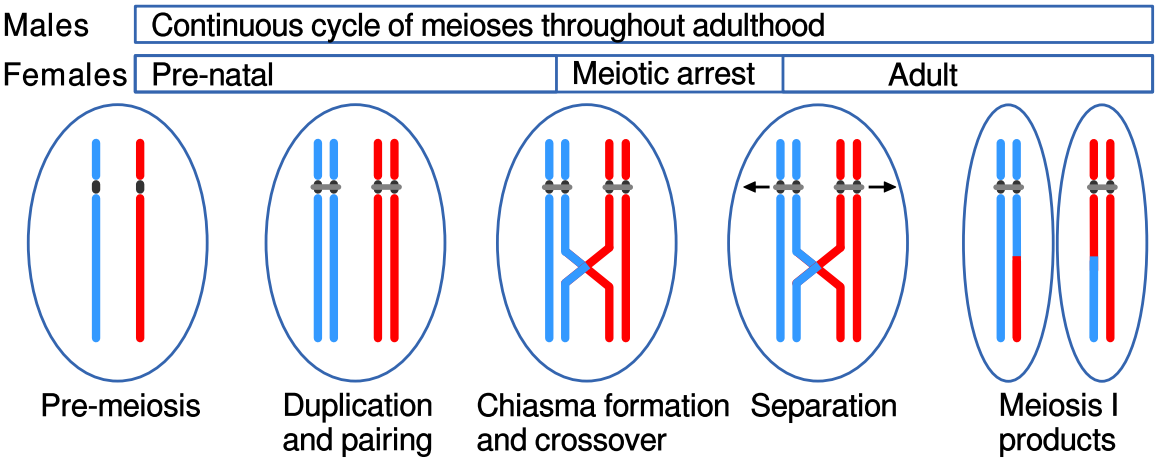

Figure 1.3: **Timing of meiosis I events in humans.** Males exhibit a continuous wave of meioses throughout adulthood. In females, meiosis begins before birth and enters an arrest period that can last decades before completion.

1.2 Historical studies of meiotic recombination

Recombination can be observed in a number of ways, and its discovery came decades prior to the discovery of the structure of DNA. Thomas Hunt Morgan first observed the separation of linked traits while studying Drosophila in 1911[17], and proposed the theory of crossing over between chromosomes. In addition he suggested that the recombination rate could increase with the distance between factors. Morgan's student, Alfred Henry Sturtevant, quantified this change in rate over physical distance into "map distance," using this concept to construct the first genetic map. This map represented the order of, and crossover rates between genes on the X chromosome in Drosophila[18]. In addition, Sturtevant observed that one crossover tended to inhibit the placement of a second nearby, an early description of interference. A later study by Harriet Creighton and Barbara McClintock in corn (*Zea mays*) in 1931 demonstrated that recombination between genes was tied to an exchange of chromosomal segments[19].

Tracing the inheritance of markers from one generation to the next within a family pedigree provided the first genome-wide measurement of recombination across the human genome, prior to the completion of the Human Genome Project. Early studies used restriction fragment length polymorphism (RFLP) probes to identify specific loci within the genome, and determine if they are linked. An early study described the use of RFLPs to generate a linkage map of recombination in the human genome[20]. Further linkage studies increased the marker density across the genome by using microsatellite, short tandem repeat polymorphisms (STRPs), and other approaches to capturing genetic variation[21–23]. The Marshfield map, generated in 1998 by Broman *et al.*[24], was an important step in characterizing recombination on a genome-wide basis.

With the completion of the Human Genome Project and the publication of the draft sequence of the human genome[25,26], human genetic variation has become increasingly well characterized. As a consequence, a number of technologies have sprung up to make genome-wide ascertainment of variation a routine procedure. Currently, microarray technology provides a well-balanced approach for determining genome-wide coverage of genetic variation. These arrays target a pre-selected panel of hundreds of thousands to millions of single-nucleotide polymorphisms (SNPs) across every chromosome for a reasonable cost.

1.3 Methods for studying recombination

Genome-wide methods have the primary goal of creating a genetic map of the frequency of recombination as a function of physical distance across each chromosome. Recombination is represented in units called Morgans, which quantify the amount of recombination between two physical locations, which are given in base pairs. One Morgan corresponds to the physical distance between two markers so that an average of 1 crossover will occur between in any given meiosis. Typically, the centiMorgan (cM) is the preferred unit of measurement, which is equivalent to 0.01 Morgans. This data is used to generate genetic maps, which describe the relationship between genetic distance and physical distance on a chromosome. In contrast to genome- or chromosome-wide methods, several other methods are limited in scope, and reveal information about specific loci. With each method, the detection of recombination is made difficult by the fact that crossovers are rare – only 20-60 are expected to occur in each meiosis.

1.3.1 Pedigree analysis

Tracking the transmission of alleles from one generation to the next within known pedigrees provided the first data on recombination in early linkage studies, and pedigree analyses are still widely in use today. Regardless of the method used for obtaining markers, the principle of detecting recombination in a pedigree remains similar. Crossovers can be identified by tracing the allele transmissions from parent to child. Figure 1.4 provides a simple visual example, showing a family quartet. The father in this pedigree has two informative SNPs producing haplotypes 1-1 and 0-0, while the mother is homozygous at both sites. The male child has a 1-0 haplotype, and therefore must have inherited a recombinant haplotype from his mother. We can identify here a crossover event and localize that event to an interval flanked by two informative genetic variants. This region of uncertainty can vary in size and depends on the spacing and genotypes of polymorphic variants within the genome. It is interesting to note that for a pedigree analysis, while whole-genome sequencing technology allows the discovery of a higher density of markers across the genome, its use is often not worth the higher cost. A higher variant coverage will help to narrow the region of uncertainty surrounding a particular crossover, but it will most likely not assist in the detection of additional crossovers in a single meiosis. Beyond this intuitive example, the problem of determining the parental phase of a recombinant

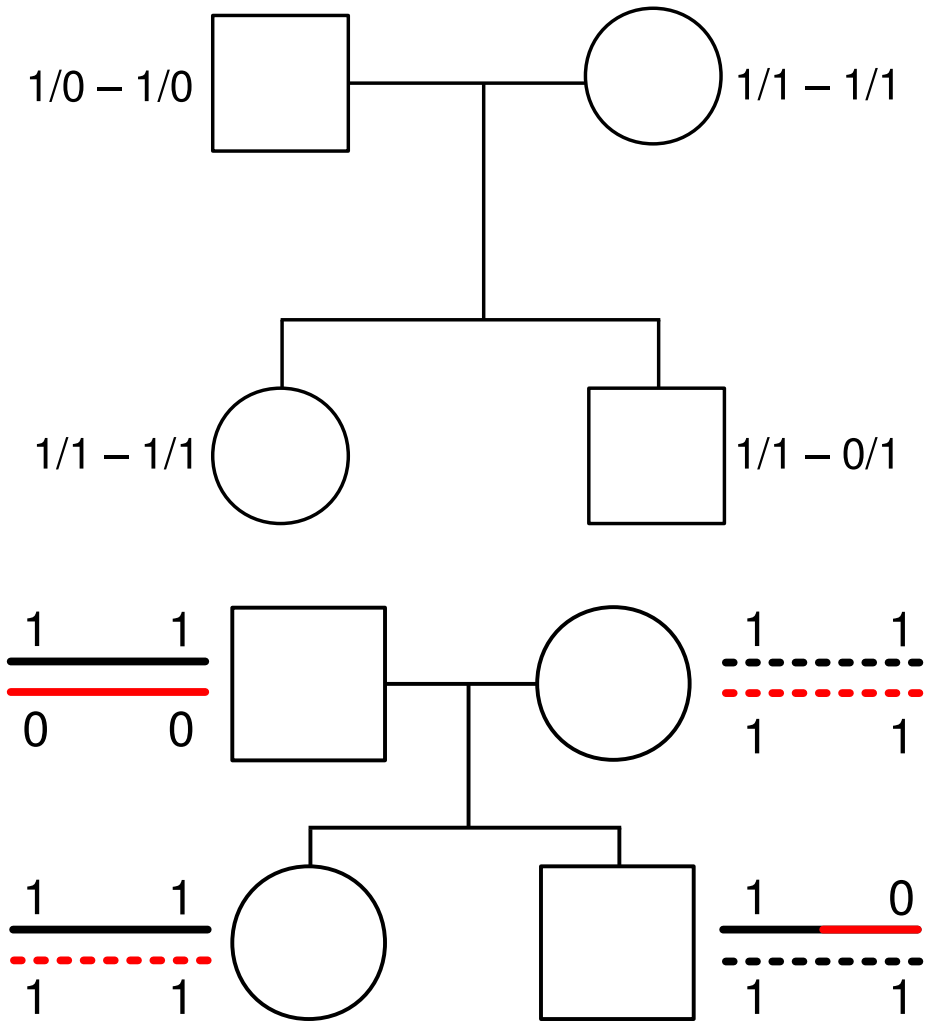

Figure 1.4: **Allelic transmission in a family quartet.** The top panel outlines a simple pedigree in which only genotype information is known. The bottom panel shows a pedigree for which phase-known haplotype information is known. Here, all chromosomes (lines next to each individual) are transmitted in the absence of recombination except for in the male child. Here there was a recombination event in the father to generate the recombinant chromosome (black/red)

chromosome has been addressed in a number of methods.

1.3.2 Linkage disequilibrium approach

Another powerful method uses population genetic data to study recombination. These data can be gathered by genotyping samples from unrelated individuals on a microarray or by whole-genome sequencing.

The inference of recombination in such a dataset relies on the quantification of levels of linkage disequilibrium (LD) within the samples, a measurement of linkage between loci. For example, when alleles at one locus are inherited completely independently of alleles at another locus they are considered to be in linkage equilibrium. However, many alleles exhibit a non-random association, and individuals of the same species tend to share haplotype segments that reflect a shared evolutionary history. When two alleles on an ancestral haplotype are inherited together they are considered linked, and are not independent of each other. These alleles exhibit evidence of linkage disequilibrium, a deviation from the assumption of random assortment of alleles.

LD is measured in a pairwise fashion, considering allele frequencies at each pair of markers in the genome. From measurements of LD within a number of unrelated samples, methods based on coalescent theory have been developed to estimate the recombination rate[27]. In software such as LDhat[28–30], the population-scale recombination rate, ρ, is estimated from the data, and the per-generation recombination rate, r, can be calculated by the relationship $\rho = 4N_e r$, where N_e represents the effective population size. These methods are quite powerful and have produced high-quality estimates of recombination in humans[31], however, they are subject to limitations. First, this method requires the knowledge of the genealogical history of a sample, which is unknown, and thus relies on a simplifying approximation to the true genealogical structure. Second, these maps by their nature generate sex-averaged data only, since recombination events that are inferred have occurred over the course of potentially thousands of generations.

1.3.3 Molecular assays

A number of molecular assays have been developed for studying recombination. Most, however are limited to small regions of the genome, and are more effective in males.

Sperm cell assays

Sperm typing is one method of identifying both gene conversion and crossover events. Sperm typing was first used in 1989 to study crossing over in humans[32], and uses an allele-specific polymerase chain reaction (PCR) assay to identify recombination events at a given locus. In this method, DNA is extracted from multiple haploid sperm cells from a single donor and subject to PCR. A common reverse primer is used in conjunction with two different allele-specific forward primers, which correspond to a polymorphic site in the diploid genome, and are designed to produce different amplicon sizes depending on the matching nucleotide. Analysis of the PCR products from many sperm cells can reveal the phase of the donor individual, and the recombinant status of each sperm cell.

Sperm typing has been used to produce high-quality data from a number of loci throughout the genome. One of the first major findings to come out of sperm typing was the characterization of recombination hotspots in the human major histocompatability complex (MHC), first within one gene, *TAP2*[33], then expanded to cover a wider 216 kb region of the MHC[34]. All six hotspots found within the region of the MHC were found to be tightly correlated to regions in which LD broke down, providing molecular evidence that recombination hotspots have severe effects on LD patterns.

Further work using whole genome, single-cell sequencing approaches has shed light on recombination on an individual basis for the first time. A study by Lu *et al.*[35] sequenced 99 sperm cells from an Asian male, providing valuable information on the level of variation across the entire genome of a single individual. In addition, a further study analyzed genome-wide recombination using more than 100 sperm cells[36]. These studies are promising, and when expanded to include measurements from multiple individuals, will provide much needed data on individual level variation in crossover, and potentially gene conversion as well.

Single oocytes.

Although much harder to study, recent studies have provided much needed insight into recombination in single oocytes through novel methods. Hou *et al.*[37] conducted an analysis in human oocytes, with the motivation to provide a potential new method for applying next generation sequencing methods to genetic screening with in vitro fertilization (IVF) techniques. Here, the researchers extracted and fertilized a total of 70 oocytes from 8 Asian female donors, collecting the first and second polar bodies,

as well as the male and female pronucleus from the fertilized egg. This complete collection of data from female oogenesis and fertilization enabled complete haplotype phasing and crossover calls to be achieved. Most intriguingly, this enabled the generation of personal genetic maps for each donor, allowing the researchers to address the question of how recombination varies on an individual basis. This data allows the observation of all four members of the tetrad bundle, both sister chromatids and their homologues, something usually missed by inferential studies of recombination.

Using a similar approach, recovering genetic material from the first and second polar bodies and oocyte, Ottolini *et al.*[38] were able to observe all four products of female recombination. The researchers generated "MeioMaps," providing valuable information on how meiosis proceeds on an individual level.

Recombination initiation maps

Additional methods have been developed to identify double strand breaks within a individual cells undergoing meiotic recombination. Pratto *et al.*[39] utilized chromatin immunoprecipitation coupled with sequencing to identify DSBs associated with the strand-exchange protein DMC1 to generate DSB maps in four unrelated human males. This method identified the majority of DSBs within the meiotic cell, only a fraction of which would be resolved as crossovers that could be identified via genotyping methods. The remainder would be repaired as non-crossover gene conversions. The researchers found the DSBs cluster into hotspots, of which 51% overlapped with the LD crossover hotspots[31], and 80% of DSB hotspots overlapped regions with elevated recombination rate.

1.4 Genetic maps of recombination.

1.4.1 Marshfield map

The Marshfield map, generated by Broman *et al.*[24] in 1998, was the first genetic map of the human genome with a resolution high enough to make inferences on the recombination properties in humans, using >8,000 short tandem repeat polymorphsims (STRPs) in 188 meioses. Here, estimates of the genome wide map lengths, inferences on individual variation, and sex differences in recombination were highlighted. The ratio of female to male autosomal map length was estimated at 1.56, indicating that the recombination rate in females is substantially higher than males. This ratio has proved stable

over a number of subsequent studies (summarized in Table 1.2). Analysis of this ratio as a function of chromosome position revealed that male recombination tends to be highest in the telomeres, while female recombination tends to be higher towards the centromeres. This study provided valuable insight into broad-scale sex dimorphism in recombination, and raised questions as to the extent of fine-scale differences between males and females.

1.4.2 deCODE maps

Another major stride in pedigree-based genetic maps came from deCODE genetics, an Icelandic pharmaceutical company that used their database of genealogical and genetic data on many Icelandic families to infer recombination, producing several high quality studies. The first was in 2002, in which 146 families, comprising 1,257 meioses, were genotyped using 5,136 microsatellite markers[40]. This data, in conjunction with the draft sequence of the human genome[25,26] was used to improve the marker order and their placement within the reference sequence. The genetic map generated from this study confirmed much found by Broman *et al.*[24] in terms of sex dimorphism in recombination, and further characterized fine scale variation. One particular finding was the discovery of recombination "jungles," which are regions of high crossover rate that are clustered towards telomeres. In addition, recombination rate was found to correlate with GC content, CpG motif occurrence, and tracts of poly(A)/poly(T), together explaining ~37% of variation.

In 2010, using genome wide SNP data, deCODE genetics published an updated sex-specific recombination map consisting of 15,257 meioses[41]. Instead of a typical pedigree-based recombination study, Kong *et al.*[41] leveraged the high degree of relatedness within the Icelandic population to infer phase and parent-of-origin in 20,217 individuals typed on microarrays assaying >289,000 autosomal SNPs. Here, phase was determined for parent-child pairs by taking into account haplotype blocks that are shared with other individuals within the population to determine the parent-of-origin for a particular block. Crossovers were called when a segment of a child's haplotype was inferred to move between maternal and paternal origin. This enables recombination be called in 15,257 parent-offspring pairs. One effect of this parent-child phasing approach is that inference of recombination events near the telomeres was difficult, and Kong *et al.*[41] omitted the most distal 5 Mb for each chromosome. The omission of the telomeric regions, where male recombination is higher, contributed to the inflation of the female:male map length ratio, at 1.78 in this study. The true value is more likely to be closer to

the consensus ratio of around 1.6 (Table 1.2). Since its release in 2010 the deCODE genetic map has proven quite valuable as a high quality sex-specific map of recombination in the human genome.

Other pedigree studies. A study in 2008 performed a pedigree analysis of individuals from Hutterites, an isolated population of European descent that lives in the US[42]. This was the first study to use genome-wide high-density SNP arrays (in this case the Affymetrix GeneChip Mapping 500k array) to study recombination. A total of 728 meioses were analyzed, yielding valuable data regarding the overlap of crossovers and hotspots on an individual basis. Other pedigree studies include a map generated from 980 meioses using individuals of Korean and Mongolian descent[43].

1.4.3 Linkage disequilibrium based maps

LD has been used to great effect in the identification of recombination events, both in the human genome and in other species. Perhaps the most high-quality LD map comes from data generated from the International Hapmap Project, Phase II[31], which includes data from 270 individuals genotyped at over 3.1 million SNPs. This map was generated using samples from both European (CEU) and African (YRI) origin, greatly improving the resolution over the Phase I map. The resolution of this map refined the collection of hotspots first identified by Myers *et al.* [44], and expands this set to include 32,996 within the human genome. This map remains in common use today, still providing the highest resolution of sex-averaged recombination rate across the human genome.

1.5 Fine-scale recombination placement in the genome: recombination hotspots

1.5.1 Initial discovery of recombination hotspots

Sperm typing produced the first set of well-characterized hotspots in humans, initially focusing within the MHC[33,34]. Extensive analysis by sperm typing has indicated that hotspots can have recombination rates thousands of times that of the surrounding region and are generally just 1-2 kb in width[45,46]. The correlation of hotspot locations identified by sperm typing in this regions with the breakdown of LD measurements provided support for the use of LD methods to find hotspots of recombination genome-wide, without the expense and limitations of single-cell assays[34]. Using LD methods[28], hotspots have

been identified on genome-wide scales and in 2005 Myers *et al.*[44] used LD methods to produce the first fine-scale recombination map of the human genome. Using this map, hotspots were found to be a ubiquitous feature of the human genome, with over ~25,000 hotspots identified, occurring roughly every 50 kb. The subsequent HapMap LD map expanded the characterized set of hotspots to more than 30,000 throughout the human genome[31]. These LD maps of recombination also estimated the proportion of recombination occurring within various fractions of the total genome sequence, finding that recombination was intensely concentrated, with 80% of all recombination occupying less than 20% of sequence.

1.5.2 Discovery of PRDM9

Since hotspots were first identified in large numbers, questions regarding the regulatory mechanism have persisted. Myers *et al.*[44] found that DNA sequences within recombination hotspots were enriched for certain DNA repeat elements such as THE1A/B, as well as a short sequence motif (CCTCCCT) located at the centers of many hotspots. A further family of motifs were identified via the use of the HapMap Phase 2 LD map[31], encompassed by the degenerate 13 bp motif (CCNCCNTNNCCNC)[47], estimated to be involved in up to 40% of all hotspots. Hotspots were associated with increased GC content, GC increasing mutations, a likely result of the action of biased gene conversion[48].

In 2010, a series of papers by three separate groups converged on the identification of PRDM9 from different approaches. Myers *et al.*[49] compared the sequence of the 13 bp motif against predicted binding of various zinc finger arrays in the genome, finding PRDM9 as a top candidate. Using a mouse genetic approach, Baudat *et al.*[50] focused on a locus previously determined to be involved in localization of crossover initiation[51,52], and narrowed this genomic interval to a region containing the *Prdm9* gene. A third study used linkage analysis in mice to narrow the region, again identifying PRDM9 as the likely candidate protein[53].

PRDM9 was originally identified in mice, where is was called *Meisetz*, and was shown to be active only in early meiotic prophase[54]. PRDM9 (PR domain containing 9) has multiple domains. It contains a KRAB (Kruppel-associated box) domain, that is potentially involved in protein-protein interactions, a PR/SET domain that serves a methyltransferase function, responsible for the trimethylation of lysine 4 on histone 3 (H3K4)[54]. PRDM9 also contains a high polymorphic C2H2 zinc finger array, responsible

for binding to specific DNA sequences[54].

1.5.3 PRDM9 alleles

The PRDM9 zinc finger array has tremendous variability and there are a number of alleles present, not just in humans but also chimpanzees and other primates[55]. In humans, the major alleles A and B, are both present in a high percentage of European individuals, and differ by only a single amino acid change, while the I allele encodes a longer zinc finger array[50]. The A and B alleles are predicted to recognize the degenerate 13-mer motif, but the I allele is not[50]. This effect is seen at the DSB-level as well, with PRDM9 A and B variants contributing to generate similar DSB hotspots[39].

Furthermore the PRDM9 allele status of an individual has a strong effect on the hotspot overlap. In the Hutterite study, A/A individuals have significantly different overlap compared to A/I, and A/B, with the A/I heterozygous having lower hotspot usage overall[50].

Sperm typing in men of African ancestry revealed that PRDM9 variants similar to the “C” allele (termed C-type), were more common in this population. Furthermore, these C-type alleles specified different hotspots with a motif different from those seen Europeans[56]. These “African-enhanced hotspots” all contained a common motif, CCNCNNTNNNCNTNNC, but were associated with the PRDM9 C-type alleles.

A study using recombination initiation maps of recombination to locate DSBs found that PRDM9 alleles affected DSB locations, which represent a set of both crossovers and gene conversion[39]. The DSB locations were largely tied to the specific PRDM9 allele for each particular individual. $PRDM9_A$ and $PRDM_B$ alleles appear to specify similar DSB hotspots, while $PRDM9_C$ has a separate specificity. PRDM9 heterozygosity was also found to influence hotspot strength.

1.6 Sexual dimorphism in recombination

1.6.1 Broad scale distribution within the genome differs between the sexes

The distribution of recombination within the genome varies considerably between individuals, sexes, and populations. At the broad scale, crossovers are not distributed equally across the genome. The recombination rate is higher towards the telomeres[24,28,31]. This effect seems to be primarily driven by

sex. Crossing over in males occurs more frequently in telomeric regions, while females have higher recombination rates closer to the centromeres[40–42].

Numerous studies in humans have found that recombination tends to be depressed within gene regions overall, with high recombination just upstream and downstream of gene regions[28,31,41,44,48]. When looking further away from gene regions, recombination appears to be elevated up to around around 500 kb away, then falls off after a few hundred kb more. Evidence for sex differences in recombination has also been suggested to occur around gene regions. Kong *et al.*[41] found that recombination was lower within genes overall, and that female recombination was lower within genes, but higher in the surrounding regions. Additionally, while females have a lower rates within genic regions, males have higher rate within exonic regions[41].

In addition, results from LD maps in humans have examined the proportion of recombination occupying various proportions of sequence. From these analyses, it is estimated that the majority of recombination, 80%, occurs in approximately 20% of the sequence[28,31,44] This concentration of recombination could be due to clustering within hotspots. A majority of crossovers in each meiosis were found to cluster into recombination hotspots, driving recombination into limited region of the genome[42], although this overlap varies widely between individuals.

1.6.2 Recombination under genetic control

A number of genetic factors have been identified that alter properties of recombination, a summary of which can be found in Table 1.1. In a sperm typing analysis, Jeffreys and Neumann[57] found a SNP whose minor allele suppressed the ratio of crossover to gene conversion near a particular hotspot. Additionally, this variant was found to be overtransmitted in the offspring. This is an example of meiotic drive, in which forces acting during meiosis alter the expected transmission ratios for a particular locus, causing over- or under-transmission. Another study in the Icelandic population found an association with an inversion at 17q21.31 on recombination rate[58]. The presence of this inversion is associated with an increase in crossover rate in females, but not males.

In addition, RNF212 has been associated with recombination rate in an Icelandic population[59], and this has been replicated in a number of follow-up studies[60–63]. Interestingly, the linkage of two particular SNPs near this gene is associated with an increased recombination rate in males, and a corresponding decreased rate in females. RNF212 is a homologue of ZHP-3, a *C. elegans* gene

involved in crossing over and disjunction, localizing to sites of crossover, and aids in the change in chromatin structure that occurs with the disassembly of the synaptonemal complex[64]. Mouse RNF212 was found to have similar function, facilitating synapsis, and forming crossover-stabilizing structures[62]. In addition RNF212 possibly influences the decision to repair a DSB as a crossover rather than a gene conversion[62].

A study by Chowdhury *et al.*[60] using 2,310 meioses found significant associations with variants associated with four gene regions. Two were associated with female recombination rate (KIAA1462, PDZK1), and the other two with male rate (UGCG, NUB1). These genes are poorly characterized and their function, beyond potential meiotic roles, remains unknown. However, this study reinforces the suggestion that male and female recombination rate are controlled via differing genetic factors.

Another study, again in the Icelandic population with an expanded number of meioses, identified a further 8 genomic regions that have separate associations with male and female recombination rates[63]. In the telomeric region of chromosome 4, near RNF212, one variant was found that was estimated to increase the map length by 386 cM in females, and 124 cM in males. Novel variants were found in an additional six genes (Table 1.1).

The strongest association thus far has been with the PRDM9 protein, which was first associated with hotspot usage and variation in alleles encoding the zinc finger array[50,65]. Berg *et al.*[56] found that DNA sequence variation in *PRDM9* alleles greatly influences hotspot usage in African populations, who appear to have a distinct subset of hotspots[66]. Genome-wide associations with hotspot usage were replicated in the Icelandic population with strong effects[41].

In non-human studies, both RNF212, and REC8, a SC-associated protein, have also been shown to associate with recombination rate in cattle[67].

Heritability of recombination modifiers. Several lines of evidence indicates that modifiers of recombination are heritable. In the 2004 deCODE study, siblings in large families, were analyzed and data suggested that recombination rates are broadly heritable[68]. The pedigree analysis of the Hutterites population by Coop *et al.*[42] was the first to find extensive DNA sequence variation among individuals in terms of their hotspot overlap (the proportion of an individual's crossover that overlap with known hotspots). Furthermore, it was found that the extent that hotspots overlap is also heritable. Further work reinforced the heritability of hotspot usage as a phenotype, and supported the finding

Gene / region	Chr.	Association	Female rate	Male rate	Study	Replication
RNF212	4	rate	+	-	Kong *et al.* [59]	60–63,69
17q21.31 inversion	17	rate	+	none	Stefansson *et al.* [58]	60,61,63
KIAA1462	10	rate	+	none	Chowdhury *et al.* [60]	
PDZK1	1	rate	+	none	Chowdhury *et al.* [60]	
UGCG	9	rate	none	+	Chowdhury *et al.* [60]	
NUB1	7	rate	none	+	Chowdhury *et al.* [60]	
CCNB1IP1	14	rate	+	weakly +	Kong *et al.* [63]	69
C14orf39	14	rate	+	none	Kong *et al.* [63]	
SMEK1	14	rate	+	none	Kong *et al.* [63]	69
RAD21L	20	rate	weakly +	+	Kong *et al.* [63]	
MSH4	1	rate	+	none	Kong *et al.* [63]	
CCDC43	17	rate	+	none	Kong *et al.* [63]	
PRDM9	5	hotspot usage	NA	NA	Baudat *et al.* [50]	61,63,69

Table 1.1: **Genomic regions associated with recombination in humans.**

that both male and female recombination rates are heritable[61]. In addition, selection has been shown to act to retain eggs with higher recombination rates[38], a phenomenon previously suggested to explain the apparent increase in recombination rate in older mothers[68]. This raises the possibility that other aspects of the recombination process may also be heritable, but further work in larger cohorts is will necessary to determine the full extent of these modifiers.

1.6.3 Heterochiasmy

The Marshfield map[24], provided some of the first genome wide evidence of recombination rate variation across the human genome, and between males and females. Particularly interesting was the finding that recombination rates are higher in the telomeres, especially in males, and that females have a 1.56-fold higher rate of recombination in the autosomes. The estimation of this ratio proved to be accurate, despite the low marker coverage and few meioses used, and has been reinforced through numerous follow-up studies[40–43,69–71].

This example in humans provides an illustration of heterochiasmy, the unequal distribution of recombination rates between the sexes of a species. Humans are among a large majority of species with data currently available in which the female recombination rate is higher than that of the male (Table 1.2).

Related to heterochiasmy is the Haldane-Huxley rule[72,73], which states that in organisms in which one sex has an absence of recombination (achiasmate) it is the heterogametic sex (the sex with differing sex chromosomes, e.g. XY males). Because recombination is suppressed between the sex chromosomes in the heterogametic sex, Haldane[72] hypothesized that this suppression carried over to the autosomes as well. But the Haldane-Huxley rule does not apply to species in which neither sex is achiasmate. There are a number of known species in which the heterogametic sex has an equal or higher recombination rate than the homogametic sex (e.g. XX females), and several other possibilities have been proposed.

One interesting explanation is that recombination is lower in the sex that undergoes stronger selection. Trivers[74] suggested that males in general are subject to stronger selective forces than females, who often have more choice in their selection of a mate. Because males with very fit phenotypes reproduce more often that unfit males, these males will tend to keep their high-fitness gene

Common name (species)	Year	Study	Female (cM)	Male (cM)	Ratio	Sex avg (cM)
Human (*Homo sapiens*)	1991	Morton [21]	4782	2809	1.70	3795.50
	1994	Matise *et al.* [22]	4790	2625	1.82	3707.50
	1996	Dib *et al.* [23]	4198.8	2729.7	1.54	3464.25
	1998	Broman *et al.* [24]	4251	2730	1.56	3490.50
	2000	Broman and Weber [70]	4435	2730	1.62	3582.50
	2002	Kong *et al.* [40]	4280.9	2590.5	1.65	3435.70
	2008	Coop *et al.* [42]	4133	2699	1.53	3416.00
	2010	Kong *et al.* [41]	4070.8	2286.6	1.78	3178.70
	2013	Bleazard *et al.* [43]	3933.5	2502.9	1.57	3218.20
	2014	Kong *et al.* [63]	3868	1925	2.01	2896.50
	2015	Campbell *et al.* [69]	4354.91	2707.66	1.61	3531.29
	2016	Bhérer *et al.* [71]	4122.8	2635.2	1.56	3379.00
Mouse (*Mus musculus*)	2009	Cox *et al.* [75]	1495.3	1375.3	1.09	1435.30
Dog (*Canis lupus familiaris*)*	1997	Mellersh *et al.* [76]	1039	766	1.36	902.50
	1999	Neff *et al.* [77]	1820	1290	1.41	1555.00
	2010	Wong *et al.* [78]	2276	1909	1.19	2092.50
	2016	Chapter 4	2162	1816	1.19	1989.00
Sheep (*Ovis aries*)	1995	Crawford *et al.* [79]	1907	2143	0.89	2025.00
Cattle (*Bos taurus*)	1997	Kappes *et al.* [80]	2879	2808	1.03	2843.50
	2015	Ma *et al.* [81]	2320	2550	0.91	
Pig (*Sus scrofa*)	1995	Archibald *et al.* [82]	2150	1650	1.30	1900.00
	1996	Marklund *et al.* [83]	2565	1830	1.40	2197.50
Short-tailed opossum (*Monodelphis domestica*)*	2004	Samollow *et al.* [84]	443.1	884.6	0.50	663.85
European tree frogs (*Hyla arborea*)*	2008	Berset-Brändli *et al.* [85]	296.6	20.7	14.33	158.65
Zebrafish (*Danio rerio*)*	2002	Singer *et al.* [86]	2582.7	942.5	2.74	1762.60
Rainbow trout (*Oncorhynchus mykiss*)*	2000	Sakamoto *et al.* [87]			3.25	
	2005	Danzmann *et al.* [88]			4.31	
Arctic charr (*Salvelinus alpinus*)*	2005	Danzmann *et al.* [88]			1.69	
Atlantic salmon (*Salmo salar*)*	2005	Danzmann *et al.* [88]			16.81	
Pigeon (*Columba livia*)*	1999	Pigozzi and Solari [89]			∼equal	
Chicken (*Gallus gallus domesticus*)*	2009	Groenen *et al.* [90]	3097.7	2913.7	1.06	3005.70
Great reed warbler (*Acrocephalus arundinaceus*)*	2005	Hansson *et al.* [91]	237.2	110.2	2.15	173.70
Fruit fly (*Drosophila melanogaster*)*	1914	Morgan [92]		no rec.	NA	
Arabadopsis (*Arabidopsis thaliana*)*	2013	Basu-Roy *et al.* [93]	332	575	0.58	453.50

Table 1.2: **Autosomal map length estimates in various species.** Total map lengths are given in centimorgans, while the ratio represents the female to male map lengths. Some studies do not provide map lengths but ratios are reported here as they are found in the text. PRDM9-absent species are marked with an asterisk (*).

combinations because disrupting these favorable haplotypes will be selected against. In contrast, females do not have as much variability in their ability to reproduce and selective pressure is lower.

This idea has been criticized however, especially because selection in an adult, diploid organism is unlikely to affect recombination rate at the haploid stage[94]. Lenormand and Dutheil[95] suggested that haploid selection may act differently in males and females and that this, and not heteromorphic sex chromosomes, could account for the heterochiasmy observed in plants. A more recent study[96] supports this idea, suggesting an additional level of parental control over the level of selection experienced by a given gamete.

In mammals, the idea of haploid selection to effect dimorphism is an attractive one to explain heterochiasmy, although the strength of selection is markedly lower. Since oogenesis is arrested at dictyotene and does not complete until fertilization, the haploid stage is essentially absent in females. Therefore, there is a higher potential for selection in male haploids, even if only a handful of genes are actually expressed[97]. In addition, there is evidence for elevated female recombination rates in imprinted regions[98], which further confounds this issue of male or female control over haploid selection.

In most studied species, including humans, females have a higher rate and a longer map length than males, with a ratio ranging from just above 1, to 14 in the European tree frog[85], and nearly 17 in Atlantic salmon[88]. Estimates in humans, perhaps the most well studied organism, have demonstrated little variation since the earliest linkage studies, with a ratio of 1.6.

However, the heterogametic sex does not always have a lower recombination rate. A number of species have roughly equivalent rates of recombination between the sexes, including pigeons[89] and possibly cattle[80]. Indeed, many species have higher recombination rates in males. Sheep[79] (0.89), and a more recent study in cattle[81] (0.91), demonstrate only slightly higher rates in males. However, marsupials provide the most extreme example. In the short-tailed opossum[84], males have twice the amount of recombination as females (a ratio of 0.5).

Cytological studies of meiotic cells have provided convincing evidence that the sex differences in crossover count are driven by, or parallel differing chromatin configurations across meiotic prophase. Using molecular cytological studies, the entirely of the SC can be visualized and measured with markers to SYCP3, which is a transverse element protein. The SC is substantially longer in females

than males, translating to a correspondingly larger DNA loop size[5,99]. This relationship has been proposed to account for the differences in crossovers observed between males and females. Males and females have chromosomes of identical size (among the autosomes), so packaging the same amount of DNA into a smaller total SC length would naturally mean a larger loop size. In addition males have fewer DSBs overall, when examining markers for RAD51, which associate with DSBs, indicating that these differences occur prior to or at the onset of DSB formation[5]. There is also evidence for variation in the SC length for chromosome arms, suggesting that SC length depends on the chromatin configuration for each chromosome arm and not the physical length[100].

1.6.4 Conflicting evidence for a maternal age effect on recombination rate.

Age effects on reproduction have been tied with an increased incidence of aneuploidy in humans[15,101]. Over the past decade, several studies have provided conflicting reports on an association of recombination rate and maternal age. Those that have demonstrated an increase in the number of crossovers with maternal age suggest that this increase could be a protective mechanism against aneuploidy.

With the aim of investigating age effects in recombination, a study in 2004 used the deCODE genetics dataset with 23,066 meioses, with a reduced set of 1000 microsatellite markers[68]. Genotyping was not done for every individual, with some families having only one parent genotyped, and recombination events were therefore imputed. The main finding from this study is that the number of crossovers observed in females appears to increase with age, with an additional 0.082 events per year ($\pm$0.012), corresponding to a 4% increase over 25 years. This effect was seen to increase within families, such that a child born later in the mother's life has a higher number of maternal crossover than their siblings born earlier. Coop *et al.*[42] analyzed recombination in 728 meioses from Hutterites, and found that mothers 35 year or older have an extra 3.1 crossover events compared to mothers under 25 years old. This effect corresponds with an extra 0.19 events per year ($\pm$0.092). No such effect was found in males in either study.

A study by Hussin *et al.*[102] examined in recombination in 195 maternal meioses from French-Canadian pedigrees. Here, the opposite effect was seen, with recombination rate found to decrease with maternal age, with a larger effect size. The crossover count was estimated to decrease by -0.49 to -0.42 events each year. Differences in the direction of the effect here may be due to real differences between populations, when considering the French-Canadian population as a genetic

isolate, or simply due to a lack of power with only 195 meioses. Another study reported a slight decrease in crossover count with maternal age in 338 meioses, finding a decrease of -0.29 events per year[43].

While evidence for an age effect on crossover count in females is conflicting, these studies all agree that there is no age effect present in males. A recent multi-cohort analysis by Martin *et al.*[103] provided much need insight into the age effect issue. Using a combination of nine cohorts comprising $>$6,000 meioses, the authors report a modest but significant increase in the crossover count with age. The authors used a comprehensive and systematic approach that avoids the methodological differences between previous studies. An interesting suggestion from this study was that of possible confounding factors upon the maternal age effect. First, that assisted reproductive technologies, including IVF, may provide an artificial selection for oocytes with a greater number of crossovers. Second was the possibility that oral contraception, which suppresses ovulation, could somehow alter the age-count association. However, neither of these possibilities could be controlled for with the power of this study and remain undetermined.

Several other studies approach this question from another direction. Evidence in mice point to sex differences in cell cycle checkpoints that control for, and terminate cells with excessive DNA damage, or without chiasmata[104]. The first of at least two checkpoints in prophase I is triggered by lack of synapsis, or breaks in the DNA that have been missed by repair machinery. A second checkpoint, in late pachytene, senses chromosomes without chiasma, and female mice appear to frequently bypass this checkpoint, while male mice do not. Thus is appears that female mice may have a less stringent set of checkpoints, and that this may account for the increased incidence of meiotic abnormalities[104].

An analysis of single oocytes provides evidence that maternal recombination rate is highly variable within a single individual, with 41.6 ± 11.3 crossover per oocyte[38]. This study revealed a selection against transmission of non-recombinant oocytes at meiosis II. These achiasmate products were more likely to be found in the second polar body instead of the transmitted oocyte. This evidence outlines a potential mechanism by which non-recombinant, potentially aneuploid oocytes could be eliminated from the germ line. Furthermore, recombination, thought to be limited to prophase I, was shown to influence events in meiosis II, much later than previously thought.

1.6.5 Open questions in sexual dimorphism in recombination

Taken together, this evidence reflects a tremendous variation in levels of sexual dimorphism and heterochiasmy, both across and within species, as well as on an individual basis. From recombination surveys of various species, it is clear that heterochiasmy manifests itself in a number of different ways. Cytological evidence shows that differences in recombination rate are related to changes at the molecular level, a result of the differing mechanisms for completing meiosis between males and females. In addition, aging effects on the recombination process point to a carefully regulated process that has potential to degrade with time. One possibility is that biological differences in the meiotic process could drive the differences in recombination outcomes. Alternatively these differences could arise as a product of the differing requirements imposed upon recombination between the sexes by external forces such as natural selection. Though much work has been done, both theoretical and experimental, no clear explanations have emerged, and the question of the cause of heterochiasmy and differences between the sexes in the recombination process remain unanswered. These questions will be addressed within this thesis, for humans in Chapters 2 and 3, and for dogs in Chapter 4.

1.7 Recombination and disease

1.7.1 Non-disjunction and the role of recombination.

In humans, aneuploidy is a major cause of pregnancy loss and disease. An estimated 30% of fertilized eggs have aneuploid chromosomes[15]. The incidence of aneuploidy remains at a low baseline level in individuals in their twenties, with males having 1-4% and females 2-3% of their sperm or eggs with aneuploidy[105,106]. This number rises steeply with age in females, up to 35% or more[106]. Aneuploidy in newborns (0.3%) and stillbirths (4%) is relatively rare compared to the rate of aneuploidy in spontaneous abortions, which can be more than 35%[106].

Multiple lines of evidence have suggested that meiotic errors are responsible in many cases. Non-disjunction, the abnormal segregation of chromosomes, can occur at either of the two cell divisions in meiosis. Meiosis errors can occur in both males and females, but errors are much more frequent in females[15,106]. Meiosis I non-disjunction is thought to involve recombination, either in a failure to

resolve chiasmata, or a lack of chiasmata entirely. Several genomic disorders have been linked to meiosis I errors, which are a common cause of trisomies of chromosomes 15, 16, 21, and 22 in females[15,106]. Meiosis II non-disjunction typically results from a failure of the sister chromatids to separate, which is frequently caused by achiasmate chromosomes. This complete lack of crossover is a frequent cause of female trisomy 18[106]. In males, meiosis I errors are frequent in trisomy 2, and in Klinefelter syndrome (47,XXY), while meiosis II errors contribute to trisomies of chromosomes 14 and 15[101].

A reduced recombination rate has been observed in human genetic maps generated from viable meiosis I aneuploidies (15, 16, 18, 21, XXX, XXY)[12,15]. This indicates that recombination rate must remain above a certain level to prevent non-disjunction. This is supported by the finding that many trisomies involve achiasmate chromosomes, in which recombination is absent in that chromosome[106].

This supports the idea that chiasmata, which serve to tether homolgous chromosomes together after the dissolution of the SC, and through the end of anaphase I, provide a crucial tension that serves to inhibit non-disjunction. Research suggests that there is a requirement of one chiasma per chromosome to prevent non-disjunction, but that there may be a backup mechanism to enable chromosomes to properly segregate even without any chiamata[107].

Meiotic arrest of oocytes in mammals has been the subject of much discussion because it may be associated with increased aneuploidy in humans, in particularly in older mothers. Several biological possibilities have been proposed to explain what happens to oocytes during meiotic arrest. One is the production line hypothesis, first proposed in 1968[108], which proposes that oocytes exit meiosis and are ovulated in the order in which they enter. An implication of this is that an oocyte produced early in the fetal stage, which will thus exit early, must have more robust crossover connections and are less prone to aneuploidy when compared to later oocytes. This may explain why aneuploidy occurs more frequently in older mothers. Furthermore, the production line hypothesis suggests that chiasmata frequency would decrease with age in females, as the rate of aneuploidy increases. Several tests of the production line hypothesis have been done in mammals. There is evidence to support the existence of a production line in mice[109], suggesting that oocytes exit meiosis in the order in which they enter. However several studies in humans contradict the assumption of a decrease in crossover count in older mothers[68,103]. In a direct test of the production line hypothesis in humans, Rowsey *et al.*[110] examine more than 8,000 fetal oocytes, finding no significant change in crossover count.

Thus, it appears that the number of crossovers in a given oocyte is not controlled as a function of order of meiotic entry, and there is a lack of evidence for any effects of the production line hypothesis on crossover count.

Another possibility, suggested by a study in the Icelandic population [68], is oocytes that have higher recombination rates are more likely to survive to become successful embryos. Since a higher rate of recombination is linked to a lower incidence of aneuploidy, it is possible that oocytes with lower crossover counts were more likely to be aneuploid. These aneuploid oocytes would therefore be discarded somehow, by a cell cycle checkpoint mechanism for example. These could not be observed in a study looking at presumably healthy, or at least viable, offspring and would giving the impression of a recombination rate increase.

1.7.2 Genomic instability

The most direct association of recombination with disease is with aneuploidy due to meiosis I errors [15,101]. Otherwise, since recombination has such a high impact on the structure of the genome, it also has the possibility to cause genomic instability if errors occur during the break and repair of DNA. Defects in recombination have been associated with a number of disorders caused by genomic instability. Errors in the repair of breaks initiated during recombination are a major cause of structural variants, leading to disease [111]. Nonallelic homologous recombination (NAHR), also known as ectopic exchange results in structural genomic rearrangements and is a major cause of recombination-associated copy number alteration. This occurs in locations in the genome containing segmental duplications, also referred to as low copy repeats (LCRs). These blocks of sequence evolved recently in primate evolution and are highly homologous. Unequal exchanges can be caused due to non-allelic recombination between different, paralogous LCRs. For example, 22q11.2 deletion syndrome is believed to be caused by improper pairing and crossing over between LCR22s that flank the region, causing it to be deleted in the recombinant chromosome [112].

Ectopic exchange contributes to a number of other similar types of genomic diseases associated with LCR-mediated rearrangements [113,114]. For example, Pentao *et al.* [115] found that rearrangements between repeat regions associate with the deletion of a 1.5 Mb region, associated with Charcot-Marie-Tooth disease type 1A (CMT1A). In addition, deletion of the 7q11 region in sperm causes Williams-Beuren syndrome [116].

PRDM9 has also been implicated in genomic instability associated disease. PRDM9 hotspot motifs have been found in regions associated with disorders of genomic instability[47]. Berg *et al.*[65] reported that variation found at the PRDM9 locus contributed to genome instability and rearrangement. Using a sperm-typing approach, non-A alleles of PRDM9 were found to be protective against the risk of rearrangements leading to another LCR-mediated set of diseases on chromosome 17, which include CMT1A, and hereditary neuropathy with liability to pressure palsies (HNPP)[65]. Here, a protective effect against genome rearrangement was seen in men homozygous for the N/N allele, with a lesser effect in heterozygous A/N individuals. In addition, PRDM9 has been associated with children with B-cell precursor acute lymphoblastic leukemia (B-ALL)[117]. Here, a rare PRDM9 allele (C) was found in a number of families, and was tied to the abnormal location of crossover events, and associated with B-ALL. Coupled with the finding of a healthy PRDM9 knockout in humans[118], these disease association raise questions as to the requirement and effects of PRDM9 in meiosis.

1.8 Interference

1.8.1 Crossover interference

Crossover interference, also known as chiasma interference, is another mechanism that affects the placement on crossover events on a chromosome. Interference was first observed in flies in 1913[18], but not formally named until 1916 when Hermann Muller coined the term, as stated,

> In a sense, then, the occurrence of one crossing-over interferes with the coincident occurrence of another crossing-over in the same pair of chromosomes, and I have accordingly termed this phenomenon "*interference*".

When two or more events occur on the same chromosome during the same meiosis, crossover interference effects the spacing of events. Positive interference causes events to be placed further apart that expected, while negative interference results in events being placed closer together. Under positive interference one crossover occurring on a chromosome is thought to "interfere" with a second, and inhibit its placement nearby. Crossover interference affects the linkage patterns between genes, has possible implications for ensuring disjunction of chromosomes.

Crossover interference models

Several biological explanations have been proposed to account for the action of crossover interference in the genome, and are reviewed here. An early suggestion was that steric interference was responsible. In this case, the cluster of proteins that are necessary to facilitate resolution of the DSB would bind to the DNA, preventing further attachment at nearby sites. This idea was mostly discounted after cytological studies failed to observe complexes of sufficient size to enable interference over long enough distances.

Models seek to account for two key aspects governing crossover placement, crossover homeostasis, and crossover assurance. Crossover homeostasis is somewhat related, but deals with the ratio of crossovers to non-crossovers. Crossover assurance captures the idea that as least one crossover is required per chromosome, which is thought to be needed to prevent non-disjunction. In a typical meiotic progression, multiple DSBs are created in the DNA and only a small fraction of them resolve to crossovers, with the remainder repaired as non-crossover gene conversions. Studies in yeast have shown that the final number of crossovers remains at the same level even when reducing the number of precursor DSBs[119]. Thus, it appears that there is some regulatory mechanism in place to ensure at least one crossover per chromosome. Another implication of crossover homeostasis is that it has the potential to change the ratio of crossovers to gene conversion events. Since crossovers tend to be maintained at a constant level, the number of gene conversions decreases if there are fewer DSBs.

The mechanical stress model. Kleckner *et al.* [120] proposed a model in which interference is governed by mechanical forces. This model starts from the basis that chromatin configuration changes over the course of meiosis, undergoing cycles of expansion and contraction. These expansions and contractions create waves of periodic increased or decreased stress that propagate along the chromosome. DSBs create “cracks” in the DNA, reducing the amount of stress directly near the crack. The amount of relief slowly decreases with greater distance away from the break until the stress returns to normal levels.

This model has a number of properties that fit into generally accepted properties of recombination. First, it ensures that at least one crossover will occur on each chromosome, as long as enough stress is generated. Second, it allows for specification of the location of stress and therefore DSBs and

crossover, based on, for example, chromatin configuration. Finally, this model allows for interference to act at multiple time points in meiosis using the same mechanism. However, the mechanical stress model assumes that all crossovers are interfering, and does not explain the observation that some organisms appear to have two pathways of interference.

The polymerization model. The polymerization model describes the maturation of DSBs to crossovers in terms of a time-dependent polymerization event spreading along the synaptonemal complex[121]. The progression of DSBs to chiasmata to mature crossovers is described in terms of a precursor structure binding, which triggers this maturation. In the model, early recombination precursors form randomly along the chromosome while it is tied into the synaptonemal complex. Some of these early events mature into chiasmata and become crossovers. As this maturation occurs, a polymerization event is triggered that moves in both directions away from the crossover. The expanding polymer blocks the binding (or forces detachment) of further precursor structures that would lead to crossover, thus creating an interference effect. The ejected precursors can then re-bind at another location on the chromosome.

This model makes a distinction between early and late crossover events, guarantees at least one crossover occurs, and captures the distance-dependency of the interference effect. However, no evidence of a polymer structure has been observed in cytological studies. Furthermore, studies in mice have shown that the formation of the synaptonemal complex is not required for interference to act[122].

The counting model. The counting model does not depend on SC length, but instead proposes that crossovers must be separated by a discrete number of non-crossover gene conversion in between[123,124]. In this model, a fixed number of events (presumed to correspond to DSBs) are placed randomly upon a chromosome. Then, the events are classified in one of two ways, either crossovers or gene conversions. Crossovers form the minority of the mature events with a specific number of gene conversions in between. Therefore, no two crossovers can occur directly beside one another, and some degree of distance-dependent spacing is maintained.

Data in Drosophia were found to fit this model well, but it was found that it poorly predicted interference in budding yeast and humans[124]. Therefore, extensions to this model were made to allow a

second class of non-interfering crossovers to exist along-side those that are interfering[125] (see "The simple gamma model" and "The two pathway model" below).

A key element of the counting model is that interference here depends on genetic distance rather than physical, or SC distance. This allows interference to be estimated in different species with differing genomic properties, such as chromosome size. This also serves to "normalize" the enormous differences in recombination rates per physical distance that are seen in nature. Even within species, due to sex-differences in recombination rate, any comparison between males and females must be done using the genetic position (in cM) of each crossover, which takes into account differences the genetic map length.

The simple gamma model. The gamma model is an extension of the counting model, in which inter-crossover distances are modeled instead of counts, which are not observed in inferential studies of interference. The gamma model starts with the assumption that all crossovers are capable of interfering with each other, and chromatid interference is neutral. The location of chiasmata on a tetrad bundle are determined by a stationary renewal process. Each chiasmata is considered an "arrival" that resets the probability controlling the distance from the previous arrival. The inter-chiasmata distances are represented by a gamma distribution with a single parameter, ν, representing the strength of interference. In this model, $\nu < 0$ corresponds to negative interference, $\nu = 0$ no interference, and $\nu > 0$ positive interference. The shape and rate parameters of the gamma distribution are not independent, but instead represented with shape of ν and rate 2ν. Therefore, the gamma distribution has a mean of 0.5 Morgans, and a standard deviation of $1/(2\sqrt{\nu})$.

In typical recombination studies, the locations of the chiasmata are unknown and only the crossover locations are observed. From a four strand tetrad, crossing over occurs twice as often as what is observed on a single product, so chiasmata are thinned to become crossovers. Since chromatid interference is neutral, each chiasma has a 0.5 probability of becoming a crossover. The mean inter-chiasma distance of 0.5 therefore translates into an average distance of 1 Morgan between crossovers.

This model allows crossover locations from transmitted genotype data to be used to estimate the parameters of a gamma distribution that represent the strength of crossover interference.

The two pathway model. The two pathway model is an extension of the gamma model that allows a non-interfering crossovers to exist in a mixture with those that are affected by interference. After a growing body of data suggesting that there may be two distinct pathways of crossing over in a number of humans, Housworth and Stahl [125] proposed the two pathway model of interference in 2003.

Here, the gamma model is extended to a mixture model. The class of interfering crossovers continue to be represented by the gamma distribution with a positive ν, and the non-interfering crossovers by a simplified gamma distribution where ν =1. A second parameter is added, p, which controls the mixture between these two cases. Interfering crossovers occur with proportion $(1 - p)$, and non-interfering crossovers with proportion p. Thus, p represents the proportion of non-interfering crossover in the distribution. These crossovers are said to "escape" the effects of crossover interference.

Data on crossover interference

Cytological measurement of interference Cytological methods have been used to study interference, relying on the identification of markers for DSBs, crossovers, and the SC during meiosis. A cytological study in mice measured distances between foci marking crossover events in terms of relative distance along the SC. These distances were modeled using a gamma distribution [126]. The researchers looked at MLH1 foci, which mark crossovers at the pachytene stage, as well as MSH4 loci, which occur just prior, in the zygotene stage. Positive interference was observed in both foci at both stages, however it was much stronger in the later stage, marked by MLH1 foci at pacytene. A follow-up to this study found that interference was not affected by the lack of an intact SC, and that interference remained present even without complete synapsis [122] Furthermore, interference can act between DSBs and is not limited to the subset that resolve as crossovers [10].

These studies suggest that, at least in mice, interference acts across at least two stages of meiotic prophase, and is not dependent on the full assembly of the synaptonemal complex. Furthermore, these observations appear to support the mechanical stress model of interference. Since this study, this method has been used in number of other organisms including mice [127], dogs [128], pigs [129], cattle [130], wildebeest [130], cats [131], shrews [132], and mink [133]. Many of these found similar levels of crossover interference, suggesting that interference is a conserved feature. In addition these studies provide evidence that the centromere is not a barrier to interference in any of the studied species

and acts over the entire chromosome, a finding reinforced through inferential studies from human pedigrees[70,107].

Strength across the genome. Conflicting results have been reported on how the strength of interference varies across the genome. That is, does the chromosome length govern the strength of interference? A cytological study of crossover interference in human males[134] found that interference strength was high in smaller chromosome, and decreased with larger chromosomes. However, this data was re-analyzed by Housworth and Stahl[135], this time using the two-pathway model, who found that interference strength was constant across all chromosomes.

Interference inferred from pedigree studies. The recombination initiation maps generated by Pratto *et al.*[39] provided important data regarding the initiation of DSBs that will lead to crossover events. One suggestion from this study is that interference could act between nearby hotspots, inhibiting the formation of a second DSB nearby, supporting the idea of DSB-DSB interference found in mice[10].

The crossovers identified from single oocytes by Hou *et al.*[37] allows valuable data to be inferred regarding both crossover interference, affecting the spacing between pairs of crossovers. Comparing to previously published sperm data[35], Hou *et al.*[37] look at crossover spacing as a function of both physical distance (in bp), and synaptonemal complex length (µm), concluding that the strength of interference is equivalent in males and females when considering the synaptonemal complex length, which is longer in females, reflecting their higher recombination rate. Somewhat puzzling was the omission of genetic distance (in centimorgans) in this analysis, a factor that controls for the 1.6 fold higher recombination rate in females over males. A reanalysis of this data using genetic distance is presented in this thesis in Chapter 3.

1.8.2 Chromatid interference

Another type of interference that is somewhat more difficult to observe, and therefore less studied, is chromatid interference. Chromatid interference affects which chromatid is involved in crossing over within the tetrad. Chromatid interference occurs when a crossover between two chromatids changes

the probability of those same two chromatids to be involved in further crossover during the same meiosis.

By identifying which chromatid is involved in each crossover Hou *et al.* [37] found evidence for weak and negative chromatid interference in human oocytes. This means that once an initial crossover is established between two chromatids (e.g. labeled 1 and 2), a second crossover is more likely to form using at least one of these first chromatids (1 and 2), and less likely to form using neither (between 3 and 4). Positive chromatid interference corresponds to a decrease in the probability of a second crossover involving the same chromatids as the first, that is, the first crossover "interferes" or prevents the placement of the second on the same chromatids.

In another study, Fledel-Alon *et al.* [107] used the observed crossovers to infer the chiasma count in tetrads using human pedigrees. Under their model, chromatid interference could account for the observed transmission of nullichiasmic, non-recombinant chromatids, however there was not enough evidence to conclude this definitively. Weak negative chromatid interference has also been found in yeast [136].

1.9 Recombination in non humans

Recombination in humans is of great interest to us as a species, and much effort has been focused here. However it is valuable to learn about recombination in other species to put human recombination in an evolutionary context. Studies of recombination have been done in a wide variety of organisms to date, a partial list of which is summarized in Table 1.2. Recombination maps are available in a number of non-human species, but many are limited to LD studies, or low-resolution linkage analysis, due to the high resource requirements of pedigree studies, and the lack of availability of high quality genome builds or assay methods.

Chimpanzees, the most recent common ancestor to humans, have a LD-based recombination map [137], but no pedigree maps are yet available, leaving open questions regarding sex differences. This sex averaged map shows that chimpanzee recombination is broadly similar to that of humans, with increased rates near the telomeres. Recombination in chimpanzees is strongly influenced by hotspots, although there is a notable absence of a strong DNA motif for PRDM9 binding, in contrast to humans [137]. One possibility is that chimpanzee PRDM9 has less specific binding, and targets a

much wider variety of target sequences than humans.

Pedigree maps have been generated in mice[138], the most recent of which uses 3,546 meioses, but a low density of markers[75]. Mice contain approximately 15,000-20,000 hotspots, also under the regulation of PRDM9[139,140] More recently, data from the Collaborative Cross[141], an inbred population generated from eight founder strains, has been used to generate sex-specific maps within the mouse genome[142]. The researchers here leveraged the breeding funnel approach from the Collaborative Cross, gathering genotype data from sibling pairs, and using computational techniques to infer recombination events.

Hotspots have been discovered in a number of other species, both with and without PRDM9. PRDM9 has been subject to rapid evolution across a wide variety of species and taxa, contributing to rapidly diverging hotspot locations between species[143,144]. Humans and chimpanzees have a complete absence of hotspot sharing, despite a high degree of overall DNA sequence identity[137,145,146]. Evidence points specifically to the rapidly evolving zinc finger DNA binding array to explain the lack of shared hotspots between humans and chimpanzees[49], and between a wide variety of other mammals[143,144,147]. Even within different human populations, there are substantial differences in hotspot specification and usage, driven primarily by differences in PRDM9 alleles (discussed above).

PRDM9 appears to be an essential component of recombination in a large number of species, however it is not a ubiquitous feature for meiosis. PRDM9 is absent in a number of species, including birds, lizards, amphibians, dogs, and fruit flies[143,144]. Intriguingly, a recent study in humans identified a healthy mother carrying a homozygous knockout of PRDM9 predicted to render the protein inactive[118]. In PRDM9 knockout mice, meiosis is not able to complete properly[139]. However, this mother had three healthy children, one of which carried the mutation. In this transmission, crossovers at PRDM9 binding locations were reduced in number but recombination seemed otherwise normal. This observation by Narasimhan *et al.* [118] raises the possibility of a backup mechanism for the completion of meiosis in the absence of PRDM9 in the human genome.

Perhaps most interesting and relevant to this thesis is recombination in dogs (investigated further in Chapter 4). The canid family provides an interesting subject for recombination studies, since evidence exists that multiple disrupting mutations occurred in the canine version of PRDM9, rendering the gene inactive. Linkage maps in dogs have been available for a number of decades[76,77], but the first sign that dog PRDM9 might be missing came with the publication of the first draft sequence of the

domestic dog genome, in a boxer, in 2005[148]. Since then a number of studies have looked at dogs and their close relative within the family Canidae to determine when and how PRDM9 became inactivated. PRDM9 was found to be disrupted in the closest relative of dogs, wolves, as well as coyotes[149], revealing that inactivation was not a result of domestication, or a limited event. Additional studies found multiple PRDM9 mutations in both the Island Fox and Andean Fox[150], but not the cat or panda[151]. This indicates that the mutations must have happened at some point after the divergence of canids from the panda, which occurred approximately 49 Mya[143,151]. Despite the loss of this gene, canids are able to successfully complete meiosis and recombination and produce fertile offspring, raising questions as to the requirement of PRDM9 in meiosis. Evidence for hotspots has been found in dogs and these hotspots are characteristically different from those found in humans. Dog hotspots appear to have a lowered intensity and occupy a wider range (4-18 kb)[150,151] when compared to humans. Direct comparisons of these results between the two species must be made with care, however, as the LD approach is sensitive to population genetic parameters. In particular, an accurate estimate of the recombination rate depends directly on accurate measurements of the effective population size, which is well characterized in humans, but not dogs. Additionally, dog hotspots appear to be localized near gene promoter regions[150], a seemingly common feature of PRDM9-absent species.

Differences in timing of meiosis in dogs. Meiosis in dogs differs from that of humans in some key respects. Meiosis in female dogs begins later, starting in the neonatal period[152]. The meiotic arrest occurs at the same dictyotene stage in both species, but is shorter in dogs, given the later onset of meiosis in dogs as well as a reduced lifespan. In addition, while meiosis exits the arrest period prior to ovulation in humans, dogs ovulate immature, primary oocytes, which only mature to fertility 48-60 hours after ovulation[153,154].

Hotspots in the absence of PRDM9. Evidence suggests that, in the absence of PRDM9, hotspots continue to persist within the genome. Most recently, two studies of recombination were released in yeast[155] and birds[156], two species lacking PRDM9. These studies provided an evolutionary perspective on recombination initiation and hotspot evolution. Hotspot locations of four species of yeast were compared, and it was found that hotspots were frequently shared, with a high overlap, providing evidence for hotspot sharing that spans millions of years of evolutionary divergence[155]. Without the

rapid changes in hotspot specification driven by PRDM9 evolution, these hotspots tend to be stable in evolutionary time. In addition, in two species of birds, the zebra finch and the long-tailed finch, hotspots were again found to be shared, despite several million years of divergence[156].

1.10 Gene conversion

Due to the small size and nature of gene conversion events, their detection within population genetic data has proven to be difficult. Given the small size of gene conversion events of 50-1000 bp, it is unlikely that a given event would overlap a typed SNP. Instead the event would occur in a region in which the donor and recipient homologues have the same sequence. The resulting conversion will result in no change to the genome, and would not be detectable outside of molecular methods to observe the event while it occurs. Furthermore, even if a gene conversion overlaps a SNP, it would only be observed if the SNP is heterozygous. Within the human genome, even the highest density SNP arrays, consisting of ~2 million markers, will cover sites spaced on average 1,500 bp apart. Not all of these positions are heterozygous, and therefore is a further limitation. Most gene conversions therefore will be missed with this method.

In addition, genotyping error is a major factor that can complicate the detection of these events. Genotyping error rates for SNP arrays are typically under 0.05%[157], however at this rate, with an array of 2 million SNPs, 1,000 of these will be improperly typed.

Despite the limitations, it is possibly to use genome wide pedigree data to detect gene conversion, as demonstrated in a recent study. Williams *et al.*[158] use 34 three generation pedigrees to study and detect gene conversion within 98 meioses using first SNP arrays, with follow up validation via sequencing. The use of three generation pedigrees provides a way to overcome the risk of false positive calls arising from genotyping error. First, in order to confidently call an event, the conversion must be detected in the first generation, and be transmitted through to the grandchildren. In addition, the parent must transfer the alternate allele from a putative gene conversion site to one of the other children. This latter requirement ensures that both of the parent's alleles are correctly typed. Gene conversion tract lengths were estimated at 100-1000 bp, but these estimates may be biased towards longer lengths due to the low SNP array density. In addition, gene conversion was found to cluster within 20-30 kb intervals in several cases, a feature not previously seen.

Williams *et al.* [158] demonstrated the success of a pedigree approach to gene conversion detection, providing valuable information on transmission across single generations. However, the sensitivity of this approach is limited. From a set of 98 meioses, with approximately 20-60 crossovers expected in each, one would expect to find ~3,500 crossovers. Given that gene conversion occurs 10 times more frequently that crossover[10,159,160], this results in ~35,000 gene conversion events in the entire dataset. Even accounting for the proportion of invisible events, only ~100 gene conversions were detected. The analysis of gene conversion in pedigree data provides valuable data, however it is incomplete and should be combined with results from other approaches.

Sperm typing has proven to be a powerful technique with which to study gene conversion events. Jeffreys and May [159] described three human hotspots in detail. This study found that gene conversion occurs 4-15 time as frequently as crossover with the genome. In addition gene conversion events occupied a tract that was estimated to range from 55 to 290 bp, highlighting the variability in length. The gene conversion tract length was further estimated to be around 500 bp [160] in follow up studies.

1.11 Description of approach

1.11.1 Methods used here to study recombination

Here, I outline the methods used to study recombination in humans (Chapters 2, 3) and dogs (Chapter 4). For the study of crossover, the focus is on data inferred through family pedigrees, where genome wide genotypes are available for both parents and children. As outlined above, crossovers can be located by comparing the genotypes within the family to determine the transmission pattern. This yields an interval, flanked by informative markers, within which the crossover must have occurred. For gene conversion, I focus on a HMM procedure, building upon previous models, to detect gene conversion events in admixed population genetic data in humans.

Hidden Markov models

Hidden Markov models (HMMs) have several applications in biology and I will use this approach more than once in this thesis, so a general description here is necessary, which I will provide using the Li and Stephens model as an example. A HMM is a statistical model that can apply to any data that can be defined as a Markov process. Markov processes have “states,” which define different classifications

the process can exist in at each observation. In addition Markov processes are memoryless, that is, transitions occur from one state to the next and depend only on the current state, and earlier states are not taken into account. DNA is an excellent candidate for HMMs, as it easily satisfies the criterion for a Markov process. Each site on a DNA molecule can be considered a discrete observation which can be considered by the HMM sequentially. Additionally, it makes biological sense to associate positions on the DNA with additional information (epigenetic state, mutation, recombination, etc.). These classifications are the "hidden" states or output of the HMM. In the case of DNA, each site on the sequence can be considered an observation, with a number of possible hidden states underlying each observation. The goal of the HMM in this case is to reveal the sequence, or path, through the hidden states that best explains the observed data.

The Li and Stephens model

The application of HMMs to genomics data was given a major boost with the development of a model by Li and Stephens [161] that models recombination using a collection of haplotype segments. This methods uses an HMM approach that models an unknown haplotype as a patchwork of sections from the previously observed haplotypes. The haplotypes are broken into segments by recombination and each segment is assigned an identifier indicating the parental haplotype it was copied from. The hidden state of this model is the path through the existing parental haplotypes.

The goal of this model, and those previous, was to estimate the population scale recombination rate, ρ, using LD patterns within a given population of haplotypes. Previous to this model, a variety of approaches were attempted, however most were too computationally complex to be of practical use in population genetic data in humans [162,163].

The Li and Stephens model (hereafter referred to as LS) examines a collection of haplotypes, $h_1, h_2, \ldots, h_n$, and determines LD patterns for a given recombination rate, ρ, allowing variation in recombination between sites. The model defines an unknown probability distribution, which is the probability of observing all of the haplotyes given the underlying ρ. This is represented as a product of conditional distributions for the population of haplotypes at each site:

$$\Pr(h_1, \ldots, h_n|\rho) = \Pr(h_1|\rho) Pr(h_2|h_1;\rho) \ldots \Pr(h_n|h_1, \ldots, h_{n-1};\rho). \quad (1.1)$$

The main advance from this model was the use of an approximation to the conditional distributions,

which allows inference to be performed efficiently. This was written using π to denote the approximate conditional distributions:

$$\Pr(h_1,\ldots,h_n|\rho) \approx \hat{\pi}\Pr(h_1|\rho)\hat{\pi}Pr(h_2|h_1;\rho)\ldots\hat{\pi}\Pr(h_n|h_1,\ldots,h_{n-1};\rho). \tag{1.2}$$

This product of approximate conditionals (PAC) enables inference to be efficiently performed using this model with little decrease in accuracy[161].

In the LS model the recombination parameter depends on each previous haplotype observed. For a subpopulation of k haplotypes, the conditional distribution should reflect the underlying distribution. Using $h_1,\ldots,h_n$ haplotypes with genotypes at S sites, the object is to model the next haplotype, given we have already observed k haplotypes so far, with the π approximation reflecting the next observation: $\Pr(h_{k+1}|h_1,\ldots,h_k)$. This problem can be described in terms of a HMM, with X_j as a Markov chain representing which haplotype, k, was copied at site j.

The model can start in any state since we have no information prior to the first site. This is the probability of starting in state 1:

$$\alpha_1(x) = \Pr(X_1 = x) = 1/k. \tag{1.3}$$

Transitions are possible between the k previously observed haplotypes, taking into account d, the distance in base pairs between site j and $j+1$. We move from state j to state $j+1$ with a transition probability (α) equal to:

$$\alpha_{j+1}(x) = \Pr(X_{j+1} = x'|X_j = x) = \begin{cases} e^{-\rho_j d_j/k} + (1-e^{-\rho_j d_j/k})\frac{1}{k} & \text{if } x' = x \\ (1-e^{-\rho_j d_j/k})\frac{1}{k} & \text{otherwise .} \end{cases} \tag{1.4}$$

Additional haplotypes h_{k+1} are modeled as an imperfect mosaic of the previous $h_1,\ldots,h_k$ haplotypes. The emissions probabilities depend on a match at the observed site on haplotype $k+1$ to the sites on a haplotype, c, among all previously observed haplotypes. Miscopying, or mutation, is allowed by the model by including a mutation parameter θ:

$$e_k(j|X_j) = \Pr(h_{k+1,j} = c|X_j = x, h_1,\ldots,h_k) = \begin{cases} \frac{\theta}{2(kL+\theta)} & \text{if } h_{k+1,j} = h_{c,j} \\ \frac{2kL+\theta}{2(kL+\theta)} & \text{if } h_{k+1,j} \neq h_{c,j} \text{ .} \end{cases} \tag{1.5}$$

Several algorithms exist to efficiently solve the most optimal state path of the model. The simplest of these is the Viterbi algorithm, which finds the most probable path through the state space, yielding

the state with the maximum probability at each site. However, the Viterbi path finds only the single, most probable path through the data, which limits further inference on the data.

Forward algorithm

The forward algorithm moves forward along the sequence, and calculates the probability of the observed sequence up to (and including) each site. Here we calculate the probability to transition from state l to state m at each site, j. This is given by:

$$f_l(j) = \Pr(h_{k+1,\leq j}, X_j = x). \tag{1.6}$$

The forward algorithm is a recursive operation, with special cases given at the starting and end of the sequence. The full algorithm is:

$$\begin{aligned}
&\text{Initialization}(j = 0): \quad f_0 = 1, f_m(0) = 0 \text{ for } m > 0 \\
&\text{Recursion}(j = 1 \ldots S): f_k(j) = e_k(x_j) \sum_l f_l(j-1)\alpha_{lm} \\
&\text{Termination}: \quad \Pr(x) = \sum_l f_l(S)\alpha_l
\end{aligned}$$

Backward algorithm

The backward algorithm starts from the end of the sequence, and calculates the probability of producing the entire observed sequence at each site onward. This is given by:

$$b_l(j) = \Pr(h_{k+1,j}, \ldots, h_{k+1,S}, X_j = x). \tag{1.7}$$

The full algorithm is:

$$\begin{aligned}
&\text{Initialization}(j = S): \quad b_l(S) = \alpha_{l0} \text{ for all } l \\
&\text{Recursion}(j = S-1, \ldots, S): b_l(j) = \sum_m \alpha_{lm} e_l(x_{j+1}) b_m(j+1) \\
&\text{Termination}: \quad \Pr(x) = \sum_m \alpha_{0m} e_m(x_1) b_m(1)
\end{aligned}$$

It is often useful to find a posterior probability of a particular state emitting from one or more sites in the sequence: $\Pr(X_i = k|x)$. This posterior probability can be useful for a number of downstream applications, such as identifying probability of observing a particular state at a each site, for example. The forward and backward algorithms can be used to efficiently calculate this posterior probability.

This is an efficient way to obtain the full probability of the observed data, which is the sum of the probabilities for all possible paths.

Having calculated the probabilities for all paths through the state sequence of the model using the forward-backward algorithm, the posterior probability can be obtained. Using the results from the forward-backward algorithm at each site, this is:

$$\Pr(X_j = k|x) = \frac{f_k(j)b_k(j)}{\Pr(x)} \tag{1.8}$$

where $P(x)$ is the result of the forward or backward algorithm.

Extensions of the Li and Stephens model. This model has been widely used in a number of applications in population genetics. In one example, known as the HAPMIX algorithm[164], the Li and Stephens model is modified to include two separate groups of reference populations, corresponding to ancestral populations. From these distinct reference populations, segments of an admixed haplotype can inferred to belong to one or the other population. It is thus possible to deconvolute the admixed sample, and determine the recombination breakpoints between different ancestral haplotype segments. HAPMIX starts with a collection of reference haplotypes in two populations, with the goal of identifying ancestry breakpoints in a single admixed sample. This is accomplished by assuming the admixed sample must have undergone an admixture event some number of generations in the past. Recent recombination since the time of admixture is modeled separately, alongside more ancient recombination. This method allows for accurate detection of ancestry breakpoints.

In another model, Gay *et al.*[165] extend the Li and Stephens model to include gene conversion. Here, a gene conversion chain, G, is modeled separately from crossover. The transitions for crossovers are essentially the same as the Li and Stephens model: either continue copying from the same haplotype, or switch to another. There are five transitions for the gene conversion chain, outlined in Table 1.3. The key innovation of this model was to consider gene conversion and crossover chains independently and simultaneously:

Transition prob.	site $j+1$	site j
$\Pr(G_{j+1} = 0 \mid G_j = 0)$	no GC	no GC
$\Pr(G_{j+1} = g \mid G_j = 0)$	GC tract	no GC
$\Pr(G_{j+1} = 0 \mid G_j = g)$	no GC	GC tract
$\Pr(G_{j+1} = g \mid G_j = g)$	GC tract	GC tract
$\Pr(G_{j+1} = g' \mid G_j = g)$	new GC	previous GC

Table 1.3: **Gene conversion transition states.** GC, gene conversion.

$$\Pr(X_{j+1}, G_{j+1}|X_j, G_j) = \Pr(X_{j+1}|X_j)\Pr(G_{j+1}|G_j)\ . \tag{1.9}$$

These two models will be revisited in Chapter 5.

Pedigree analysis

To infer recombination within a pedigree, one must define a way of calculating the likelihood of the data as a function of the recombination rate. Specifically it is desirable to find the recombination rate at each marker that maximizes the likelihood of the data.

Lander-Green algorithm. The Lander-Green algorithm[166] considers all individuals jointly at each site along the chromosome, and seeks to identify the inheritance pattern at each site using a HMM. The inheritance vector represents the parental haplotype being copied from at each site, and recombination events are represented as transitions between states of the HMM. It proceeds in two steps, the first collecting the inheritance information at each marker, and the second combining this inheritance information for multiple markers.

The Lander-Green algorithm defines, and attempts to solve an inheritance vector that defines the alleles transmitted from parent to child within a pedigree. For the pedigree in Figure 1.4A, a genotype vector can be defined for each of the two alleles belonging to the father, mother, female child, and male child by $X_0 = (\{1,0\},\{1,1\},\{1,1\},\{1,1\})$. The inheritance vector at the first site is defined for the children as $I_0 = (\{0,1\},\{0,0\})$, where 0 represents the first parental chromosome, and 1, the second. Here, the female child inherits the first paternal chromosome (0), and the second maternal chromosome (1), while the male child inherits the first chromosome from each parent (0 and 0). At the second site, $X_1 = (\{1,0\},\{1,1\},\{1,1\},\{0,1\})$, and $I_1 = (\{0,1\},\{1,0\})$.

First, all possible inheritance patterns are defined, then each site is evaluated to generate $P(X_j|I)$. This gives the most likely inheritance pattern, from which recombination events can be identified. The Lander-Green algorithm grows in complexity in linear time with the number of markers, and exponentially with the number of individuals.

Haplotype phasing and crossover calling. Inferring locations of crossover events within pedigree data can also be done with another method that first phases the individuals, then uses pedigree structure to correct phasing errors and infer crossover locations.

Current commonly used methods of determining genotypes across the genome include microarray genotyping, or whole genome sequencing. These methods produce genotype data in which the phase is unknown, so that maternal and paternal origin is not determined. Phasing is a method that classifies each of the two alleles at every SNP into haplotypes, representing the transmitted paternal and maternal chromosomes. This problem has been extensively studied, and a number of methods have been devised that work with varying degrees of accuracy.

For my thesis work presented in Chapter 4, I used a combination of tools to accurately determine haplotype phase and recombination events. The first, SHAPEIT[167], performs general phasing using a hidden Markov model. HMMs have been previously used in the estimation of haplotypes from genetic data, notably in the algorithm Phase[168], which produces accurate results and has been widely used over the past decade.

SHAPEIT improves upon previous HMM approaches by substantially reducing the computation time. In a typical solution to the HMM, the forward-backward algorithm is used, requiring a calculation for each state at every site, which can lead to a long computation time. In SHAPEIT, redundant calculations are avoided by considering the haplotyes in terms of a binary tree (or haplotype tree)[167]. This effectively omits a substantial proportion of the calculations that would occur in the forward-backward algorithm. This data structure compacts all possible haplotypes into a branching structure, which represents all possible paths and switches. The calculations for homozygous SNPs within a putative haplotype are skipped entirely, since the haplotype must obviously continue past these SNPs. This method has been used to infer phased haplotypes in a number of previous studies with accurate results, including the 1000 Genomes project[169].

SHAPEIT has been shown to produce accurate haplotype with a minimum amount of error, although some degree of switch errors still occur in the final result. Following haplotype estimation via SHAPEIT, duoHMM[170] is then used to identify recombination events in parent-child duos. The procedure for duoHMM involves multiple steps. First, duoHMM is used together with a known pedigree structure to correct any switch errors that occurred during phasing with SHAPEIT2. Second, each parent-child duo is considered separately, and phasing information is taken into account to identify haplotype switches that represent recombination events. In addition, duoHMM can be used to identify sites with probable genotyping errors, which can be removed from the analysis. When compared to a Lander-Green approach, duoHMM has been shown to have a reduced error rate, and a high

sensitivity for detecting recombination events[170]. The duoHMM approach produces similar data to that of the Lander-Green algorithm. Crossovers are represented as a genomic interval, bounded by informative markers, within which the crossing over must have occurred.

1.12 Rationale

Recombination is an essential component of meiosis and serves multiple functions, both to reshuffle genetic variation to generate new combination of alleles that can be acted upon by natural selection, and to serve as a physical connection between the chromosomes during meiosis, which serves to prevent non-disjunction. Both crossover and gene conversion are important in shaping the LD structure of the genome, which affects inheritance. Additionally recombination has an important role in genome wide associations studies (GWAS), which often genotype a subset of variation within the genome, and rely on linked SNPs to infer associations. Of the subset of recombination events that are repaired as crossovers, there is tremendous variation in the placement and frequency within the genome. This variation in crossover properties has been shown to occur between individuals, sexes, and populations, and extends across species. A recent major advance was the discovery of the PRDM9 protein, which acts to funnel recombination events into concentrated regions of the genome known as hotspots. PRDM9 is essential to recombination in most mammals, including mice and primates and much research has focused on further characterizing its function and effects. Dogs, however, are the only known mammals to not use PRDM9, having lost it through mutation millions of years previous. This therefore makes dogs an interesting species in which to study recombination, to learn about the evolutionary effects of the loss of PRDM9.

In this thesis, I will investigate and characterize differences in crossover properties between the sexes in humans (Chapters 2 and 3), and in dogs (Chapter 4). In humans, I will present a large scale pedigree analysis, which enables the characterization of recombination properties on a fine scale. In addition, I will investigate how the recombination process changes with age, which has a connection to the incidence of aneuploid pregnancies that increase in older mothers.

In dogs (Chapter 4), I will use a pedigree approach to characterize differences crossover properties between the sexes and across the genome. Through a comparison to human data and a study of observed differences, this will shed light on how mammalian recombination has changed in the

absence of PRDM9.

As much as 90% of all DSBs that are created during the recombination process are repaired as non-crossover gene conversions, of which relatively little is known compared to crossovers. Current methods are limited to molecular characterization of gene conversion, which is limited to males. In recent years statistical methods have emerged that provide a promising framework to use for the further study of gene conversion in human population genetic data. In Chapter 5, I will generate a new statistical model for the detection of gene conversion in humans, which uses data from admixed individuals to increase the contrast to detect these events.

1.13 References

1. Bell, S. P. and Dutta, A. DNA replication in eukaryotic cells. *Annual review of biochemistry* 71:333–74 (2002). doi:10.1146/annurev.biochem.71.110601.135425.

2. Yang, F. and Wang, P. J. The Mammalian synaptonemal complex: a scaffold and beyond. *Genome Dynamics* 5:69–80 (2009). doi:10.1159/000166620.

3. de Boer, E. and Heyting, C. The diverse roles of transverse filaments of synaptonemal complexes in meiosis. *Chromosoma* 115(3):220–34 (2006). doi:10.1007/s00412-006-0057-5.

4. Oliver-Bonet, M., Turek, P. J., Sun, F., Ko, E., and Martin, R. H. Temporal progression of recombination in human males. *Molecular human reproduction* 11(7):517–22 (2005). doi:10.1093/molehr/gah193.

5. Gruhn, J. R., Rubio, C., Broman, K. W., Hunt, P. A., and Hassold, T. Cytological studies of human meiosis: sex-specific differences in recombination originate at, or prior to, establishment of double-strand breaks. *PloS one* 8(12):e85075 (2013). doi:10.1371/journal.pone.0085075.

6. Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. The double-strand-break repair model for recombination. *Cell* 33(1):25–35 (1983).

7. Baudat, F., Imai, Y., and de Massy, B. Meiotic recombination in mammals: localization and regulation. *Nature Reviews Genetics* 14(11):794–806 (2013). doi:10.1038/nrg3573.

8. de Massy, B. Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. *Annual Review of Genetics* 47:563–99 (2013). doi:10.1146/annurev-genet-110711-155423.

9. Baker, S. M., Plug, A. W., Prolla, T. A., Bronner, C. E., Harris, A. C., *et al.* Involvement of mouse Mlh1 in DNA mismatch repair and meiotic crossing over. *Nature genetics* 13(3):336–42 (1996). doi:10.1038/ng0796-336.

10. Baudat, F. and de Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Research* 15(5):565–77 (2007). doi:10.1007/s10577-007-1140-3.

11. Kauppi, L., Barchi, M., Baudat, F., Romanienko, P. J., Keeney, S., *et al.* Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science* 331(6019):916–20 (2011). doi:10.1126/science.1195774.

12. Lynn, A., Ashley, T., and Hassold, T. Variation in human meiotic recombination. *Annual Review of Genomics and Human Genetics* 5:317–49 (2004). doi:10.1146/annurev.genom.4.070802.110217.

13. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics* 1(1):40–7 (2000). doi:10.1038/35049558.

14. Venn, O., Turner, I., Mathieson, I., de Groot, N., Bontrop, R., *et al.* Strong male bias drives germline mutation in chimpanzees. *Science* 344(6189):1272–5 (2014). doi:10.1126/science.344.6189.1272.

15. Hassold, T. and Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics* 2(4):280–91 (2001). doi:10.1038/35066065.

16. Schmerler, S. and Wessel, G. M. Polar bodies–more a lack of understanding than a lack of respect. *Molecular reproduction and development* 78(1):3–8 (2011). doi:10.1002/mrd.21266.

17. Morgan, T. H. The Application of the Conception of Pure Lines to Sex-Limited Inheritance and to Sexual Dimorphism. *The American Naturalist* 45(530):65–78 (1911). doi:10.1086/279195.

18. Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* 14(1):43–59 (1913). doi:10.1002/jez.1400140104.

19. Creighton, H. B. and McClintock, B. A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. *Proceedings of the National Academy of Sciences of the United States of America* 17(8):492–7 (1931).

20. Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics* 32(3):314–31 (1980).

21. Morton, N. E. Parameters of the human genome. *Proceedings of the National Academy of Sciences of the United States of America* 88(17):7474–6 (1991).

22. Matise, T. C., Perlin, M., and Chakravarti, A. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nature genetics* 6(4):384–90 (1994). doi:10.1038/ng0494-384.

23. Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380(6570):152–4 (1996). doi:10.1038/380152a0.

24. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American Journal of Human Genetics* 63(3):861–869 (1998). doi:10.1086/302011.

25. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., *et al.* The sequence of the human genome. *Science (New York, N.Y.)* 291(5507):1304–51 (2001). doi:10.1126/science.1058040.

26. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., *et al.* Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921 (2001). doi:10.1038/35057062.

27. Auton, A. and Mcvean, G. Estimating Recombination Rates from Genetic Variation in Humans. *Evolutionary Genomics* 856:217–237 (2012). doi:10.1007/978-1-61779-585-5.

28. McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* 304(5670):581–4 (2004). doi:10.1126/science.1092500.

29. Auton, A. and McVean, G. Recombination rate estimation in the presence of hotspots. *Genome*

Research 17(8):1219–27 (2007). doi:10.1101/gr.6386707.

30. Auton, A., Myers, S., and McVean, G. Identifying recombination hotspots using population genetic data. *arXiv* (2014). doi:arXiv:1403.4264v1.

31. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61 (2007). doi:10.1038/nature06258.

32. Cui, X. F., Li, H. H., Goradia, T. M., Lange, K., Kazazian, H. H., *et al.* Single-sperm typing: determination of genetic distance between the G gamma-globin and parathyroid hormone loci by using the polymerase chain reaction and allele-specific oligomers. *Proceedings of the National Academy of Sciences of the United States of America* 86(23):9389–93 (1989).

33. Jeffreys, A. J., Ritchie, A., and Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Human molecular genetics* 9(5):725–33 (2000).

34. Jeffreys, A. J., Kauppi, L., and Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* 29(2):217–22 (2001). doi:10.1038/ng1001-217.

35. Lu, S., Zong, C., Fan, W., Yang, M., Li, J., *et al.* Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338(6114):1627–30 (2012). doi:10.1126/science.1229112.

36. Wang, J., Fan, H. C., Behr, B., and Quake, S. R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 150(2):402–12 (2012). doi:10.1016/j.cell.2012.06.030.

37. Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., *et al.* Genome analyses of single human oocytes. *Cell* 155(7):1492–506 (2013). doi:10.1016/j.cell.2013.11.040.

38. Ottolini, C. S., Newnham, L. J., Capalbo, A., Natesan, S. A., Joshi, H. A., *et al.* Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. *Nature Genetics* 47(7):727–35 (2015). doi:10.1038/ng.3306.

39. Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., *et al.* Recombination initiation maps of individual human genomes. *Science* 346(6211):1256442–1256442 (2014). doi:10.1126/science.1256442.

40. Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., *et al.* A high-resolution recombination map of the human genome. *Nature Genetics* 31(3):241–7 (2002). doi:10.1038/ng917.

41. Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099–103 (2010). doi:10.1038/nature09525.

42. Coop, G., Wen, X., Ober, C., Pritchard, J. K., and Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans.

Science 319(5868):1395–8 (2008). doi:10.1126/science.1151851.

43. Bleazard, T., Ju, Y. S., Sung, J., and Seo, J.-S. Fine-scale mapping of meiotic recombination in Asians. *BMC genetics* 14(1):19 (2013). doi:10.1186/1471-2156-14-19.

44. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310(5746):321–4 (2005). doi:10.1126/science.1117196.

45. Jeffreys, A. J., Holloway, J. K., Kauppi, L., May, C. A., Neumann, R., *et al.* Meiotic recombination hot spots and human DNA diversity. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 359(1441):141–52 (2004). doi:10.1098/rstb.2003.1372.

46. Arnheim, N., Calabrese, P., and Nordborg, M. Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *American Journal of Human Genetics* 73(1):5–16 (2003). doi:10.1086/376419.

47. Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics* 40(9):1124–9 (2008). doi:10.1038/ng.213.

48. Spencer, C. C. A., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., *et al.* The influence of recombination on human genetic diversity. *PLoS Genetics* 2(9):e148 (2006). doi:10.1371/journal.pgen.0020148.

49. Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., *et al.* Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327(5967):876–9 (2010). doi:10.1126/science.1182363.

50. Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327(5967):836–40 (2010). doi:10.1126/science.1183439.

51. Grey, C., Baudat, F., and de Massy, B. Genome-wide control of the distribution of meiotic recombination. *PLoS biology* 7(2):e35 (2009). doi:10.1371/journal.pbio.1000035.

52. Parvanov, E. D., Ng, S. H. S., Petkov, P. M., and Paigen, K. Trans-regulation of mouse meiotic recombination hotspots by Rcr1. *PLoS biology* 7(2):e36 (2009). doi:10.1371/journal.pbio.1000036.

53. Parvanov, E. D., Petkov, P. M., and Paigen, K. Prdm9 controls activation of mammalian recombination hotspots. *Science* 327(5967):835 (2010). doi:10.1126/science.1181495.

54. Hayashi, K., Yoshida, K., and Matsui, Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature* 438(7066):374–8 (2005). doi:10.1038/nature04112.

55. Schwartz, J. J., Roach, D. J., Thomas, J. H., and Shendure, J. Primate evolution of the recombination regulator PRDM9. *Nature communications* 5:4370 (2014). doi:10.1038/ncomms5370.

56. Berg, I. L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N. J., *et al.* 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):12378–83 (2011). doi:10.1073/pnas.1109531108.

57. Jeffreys, A. J. and Neumann, R. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Human molecular genetics* 14(15):2277–87 (2005). doi:10.1093/hmg/ddi232.

58. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., *et al.* A common inversion under selection in Europeans. *Nature genetics* 37(2):129–37 (2005). doi:10.1038/ng1508.

59. Kong, A., Thorleifsson, G., Stefansson, H., Masson, G., Helgason, A., *et al.* Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science* 319(5868):1398–401 (2008). doi:10.1126/science.1152422.

60. Chowdhury, R., Bois, P. R. J., Feingold, E., Sherman, S. L., and Cheung, V. G. Genetic analysis of variation in human meiotic recombination. *PLoS Genetics* 5(9):e1000648 (2009). doi:10.1371/journal.pgen.1000648.

61. Fledel-Alon, A., Leffler, E. M., Guan, Y., Stephens, M., Coop, G., *et al.* Variation in human recombination rates and its genetic determinants. *PloS one* 6(6):e20321 (2011). doi:10.1371/journal.pone.0020321.

62. Reynolds, A., Qiao, H., Yang, Y., Chen, J. K., Jackson, N., *et al.* RNF212 is a dosage-sensitive regulator of crossing-over during mammalian meiosis. *Nature Genetics* 45(3):269–78 (2013). doi:10.1038/ng.2541.

63. Kong, A., Thorleifsson, G., Frigge, M. L., Masson, G., Gudbjartsson, D. F., *et al.* Common and low-frequency variants associated with genome-wide recombination rate. *Nature Genetics* 46(1):11–16 (2014). doi:10.1038/ng.2833.

64. Bhalla, N., Wynne, D. J., Jantsch, V., and Dernburg, A. F. ZHP-3 acts at crossovers to couple meiotic recombination with synaptonemal complex disassembly and bivalent formation in C. elegans. *PLoS genetics* 4(10):e1000235 (2008). doi:10.1371/journal.pgen.1000235.

65. Berg, I. L., Neumann, R., Lam, K.-W. G., Sarbajna, S., Odenthal-Hesse, L., *et al.* PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics* 42(10):859–63 (2010). doi:10.1038/ng.658.

66. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., *et al.* The landscape of recombination in African Americans. *Nature* 476(7359):170–5 (2011). doi:10.1038/nature10336.

67. Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., *et al.* Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genetics* 8(7):e1002854 (2012). doi:10.1371/journal.pgen.1002854.

68. Kong, A., Barnard, J., Gudbjartsson, D. F., Thorleifsson, G., Jonsdottir, G., *et al.* Recombination rate and reproductive success in humans. *Nature Genetics* 36(11):1203–6 (2004). doi:10.1038/ng1445.

69. Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D., and Auton, A. Escape from crossover interference increases with maternal age. *Nature Communications* 6:6260 (2015). doi:10.1038/ncomms7260.

70. Broman, K. W. and Weber, J. L. Characterization of human crossover interference. *American Journal of Human Genetics* 66(6):1911–26 (2000). doi:10.1086/302923.

71. Bhérer, C., Campbell, C. L., and Auton, A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Unpublished - under review* (2016).

72. Haldane, J. B. S. Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics* 12(2):101–109 (1922). doi:10.1007/BF02983075.

73. Huxley, J. Sexual difference of linkage in Gammarus chevreuxi. *Journal of Genetics* 20:145–156 (1928).

74. Trivers, R. Sex differences in rates of recombination and sexual selection. In R. Michod and B. Levin, editors, *The evolution of sex*, pages 270–286. Sinauer Press, Sunderland, Massachusetts (1988).

75. Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., *et al.* A new standard genetic map for the laboratory mouse. *Genetics* 182(4):1335–44 (2009). doi:10.1534/genetics.109.105486.

76. Mellersh, C. S., Langston, A. A., Acland, G. M., Fleming, M. A., Ray, K., *et al.* A linkage map of the canine genome. *Genomics* 46(3):326–36 (1997). doi:10.1006/geno.1997.5098.

77. Neff, M. W., Broman, K. W., Mellersh, C. S., Ray, K., Acland, G. M., *et al.* A second-generation genetic linkage map of the domestic dog, Canis familiaris. *Genetics* 151(2):803–20 (1999).

78. Wong, A. K., Ruhe, A. L., Dumont, B. L., Robertson, K. R., Guerrero, G., *et al.* A comprehensive linkage map of the dog genome. *Genetics* 184(2):595–605 (2010). doi:10.1534/genetics.109.106831.

79. Crawford, A. M., Dodds, K. G., Ede, A. J., Pierson, C. A., Montgomery, G. W., *et al.* An autosomal genetic linkage map of the sheep genome. *Genetics* 140(2):703–24 (1995).

80. Kappes, S. M., Keele, J. W., Stone, R. T., McGraw, R. A., Sonstegard, T. S., *et al.* A second-generation linkage map of the bovine genome. *Genome Research* 7(3):235–249 (1997). doi:10.1101/gr.7.3.235.

81. Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., *et al.* Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLoS Genetics* 11(11):e1005387 (2015). doi:10.1371/journal.pgen.1005387.

82. Archibald, A. L., Haley, C. S., Brown, J. F., Couperwhite, S., McQueen, H. A., *et al.* The PiGMaP consortium linkage map of the pig (Sus scrofa). *Mammalian genome : official journal of the International Mammalian Genome Society* 6(3):157–75 (1995).

83. Marklund, L., Johansson Moller, M., Høyheim, B., Davies, W., Fredholm, M., *et al.* A comprehensive linkage map of the pig based on a wild pig-Large White intercross. *Animal genetics*

27(4):255–69 (1996).

84. Samollow, P. B., Kammerer, C. M., Mahaney, S. M., Schneider, J. L., Westenberger, S. J., *et al.* First-generation linkage map of the gray, short-tailed opossum, Monodelphis domestica, reveals genome-wide reduction in female recombination rates. *Genetics* 166(1):307–29 (2004).

85. Berset-Brändli, L., Jaquiéry, J., Broquet, T., Ulrich, Y., and Perrin, N. Extreme heterochiasmy and nascent sex chromosomes in European tree frogs. *Proceedings. Biological sciences / The Royal Society* 275(1642):1577–85 (2008). doi:10.1098/rspb.2008.0298.

86. Singer, A., Perlman, H., Yan, Y., Walker, C., Corley-Smith, G., *et al.* Sex-Specific Recombination Rates in Zebrafish (Danio rerio). *Genetics* 160(2):649–657 (2002).

87. Sakamoto, T., Danzmann, R. G., Gharbi, K., Howard, P., Ozaki, A., *et al.* A Microsatellite Linkage Map of Rainbow Trout (Oncorhynchus mykiss) Characterized by Large Sex-Specific Differences in Recombination Rates. *Genetics* 155(3):1331–1345 (2000).

88. Danzmann, R. G., Cairney, M., Davidson, W. S., Ferguson, M. M., Gharbi, K., *et al.* A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae). *Genome / National Research Council Canada = GeÌĄnome / Conseil national de recherches Canada* 48(6):1037–51 (2005). doi:10.1139/g05-067.

89. Pigozzi, M. I. and Solari, A. J. Equal frequencies of recombination nodules in both sexes of the pigeon suggest a basic difference with eutherian mammals. *Genome / National Research Council Canada = GeÌĄnome / Conseil national de recherches Canada* 42(2):315–21 (1999).

90. Groenen, M. A. M., Wahlberg, P., Foglio, M., Cheng, H. H., Megens, H.-J., *et al.* A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome research* 19(3):510–9 (2009). doi:10.1101/gr.086538.108.

91. Hansson, B., Akesson, M., Slate, J., and Pemberton, J. M. Linkage mapping reveals sex-dimorphic map distances in a passerine bird. *Proceedings. Biological sciences / The Royal Society* 272(1578):2289–98 (2005). doi:10.1098/rspb.2005.3228.

92. Morgan, T. H. No Crossing over in the Male of Drosophila of Genes in the Second and Third Pairs of Chromosomes (1914).

93. Basu-Roy, S., Gauthier, F., Giraut, L., Mézard, C., Falque, M., *et al.* Hot regions of noninterfering crossovers coexist with a nonuniformly interfering pathway in Arabidopsis thaliana. *Genetics* 195(3):769–79 (2013). doi:10.1534/genetics.113.155549.

94. Lenormand, T. The evolution of sex dimorphism in recombination. *Genetics* 163(2):811–22 (2003).

95. Lenormand, T. and Dutheil, J. Recombination difference between sexes: a role for haploid selection. *PLoS biology* 3(3):e63 (2005). doi:10.1371/journal.pbio.0030063.

96. Otto, S. P., Scott, M. F., and Immler, S. Evolution of haploid selection in predominantly diploid organisms. *Proceedings of the National Academy of Sciences of the United States of America* 112(52):15952–7 (2015). doi:10.1073/pnas.1512004112.

97. Dadoune, J.-P., Siffroi, J.-P., and Alfonsi, M.-F. Transcription in haploid male germ cells. *International review of cytology* 237:1–56 (2004). doi:10.1016/S0074-7696(04)37001-4.

98. Lercher, M. J. and Hurst, L. D. Imprinted chromosomal regions of the human genome have unusually high recombination rates. *Genetics* 165(3):1629–32 (2003).

99. Tease, C. and Hultén, M. A. 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):208–15 (2004). doi:10.1159/000080599.

100. Codina-Pascual, M., Campillo, M., Kraus, J., Speicher, M. R., Egozcue, J., *et al.* Crossover frequency and synaptonemal complex length: their variability and effects on human male meiosis. *Molecular human reproduction* 12(2):123–33 (2006). doi:10.1093/molehr/gal007.

101. Hassold, T., Hall, H., and Hunt, P. The origin of human aneuploidy: where we have been, where we are going. *Human Molecular Genetics* 16(R2):R203–R208 (2007). doi:10.1093/hmg/ddm243.

102. Hussin, J., Roy-Gagnon, M.-H., Gendron, R., Andelfinger, G., and Awadalla, P. Age-dependent recombination rates in human pedigrees. *PLoS Genetics* 7(9):e1002251 (2011). doi:10.1371/journal.pgen.1002251.

103. Martin, H. C., Christ, R., Hussin, J. G., O'Connell, J., Gordon, S., *et al.* Multicohort analysis of the maternal age effect on recombination. *Nature communications* 6:7846 (2015). doi:10.1038/ncomms8846.

104. Cohen, P. E., Pollack, S. E., and Pollard, J. W. Genetic analysis of chromosome pairing, recombination, and cell cycle control during first meiotic prophase in mammals. *Endocrine reviews* 27(4):398–426 (2006). doi:10.1210/er.2005-0017.

105. Hassold, T. and Hunt, P. Maternal age and chromosomally abnormal pregnancies: what we know and what we wish we knew. *Current opinion in pediatrics* 21(6):703–8 (2009). doi:10.1097/MOP.0b013e328332c6ab.

106. Nagaoka, S. I., Hassold, T. J., and Hunt, P. A. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nature Reviews Genetics* 13(7):493–504 (2012). doi:10.1038/nrg3245.

107. Fledel-Alon, A., Wilson, D. J., Broman, K., Wen, X., Ober, C., *et al.* Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genetics* 5(9):e1000658 (2009). doi:10.1371/journal.pgen.1000658.

108. Henderson, S. A. and Edwards, R. G. Chiasma frequency and maternal age in mammals. *Nature* 218(5136):22–8 (1968).

109. Polani, P. E. and Crolla, J. A. A test of the production line hypothesis of mammalian oogenesis. *Human Genetics* 88(1):64–70 (1991). doi:10.1007/BF00204931.

110. Rowsey, R., Gruhn, J., Broman, K. W., Hunt, P. A., and Hassold, T. Examining variation in recombination levels in the human female: A test of the production-line hypothesis. *American Journal of Human Genetics* 95(1):108–112 (2014). doi:10.1016/j.ajhg.2014.06.008.

111. Carvalho, C. M. B. and Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nature reviews. Genetics* 17(4):224–238 (2016). doi:10.1038/nrg.2015.25.

112. Emanuel, B. S. Molecular mechanisms and diagnosis of chromosome 22q11.2 rearrangements. *Developmental disabilities research reviews* 14(1):11–8 (2008). doi:10.1002/ddrr.3.

113. Stankiewicz, P. and Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* 18(2):74–82 (2002).

114. Liu, P., Carvalho, C. M. B., Hastings, P. J., and Lupski, J. R. Mechanisms for recurrent and complex human genomic rearrangements. *Current opinion in genetics & development* 22(3):211–20 (2012). doi:10.1016/j.gde.2012.02.012.

115. Pentao, L., Wise, C. A., Chinault, A. C., Patel, P. I., and Lupski, J. R. 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 (1992). doi:10.1038/ng1292-292.

116. Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., *et al.* Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature genetics* 40(1):90–5 (2008). doi:10.1038/ng.2007.40.

117. Hussin, J., Sinnett, D., Casals, F., Idaghdour, Y., Bruat, V., *et al.* Rare allelic forms of PRDM9 associated with childhood leukemogenesis. *Genome research* 23(3):419–30 (2013). doi:10.1101/gr.144188.112.

118. Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* page aac8624 (2016). doi:10.1126/science.aac8624.

119. Martini, E., Diaz, R. L., Hunter, N., and Keeney, S. Crossover homeostasis in yeast meiosis. *Cell* 126(2):285–95 (2006). doi:10.1016/j.cell.2006.05.044.

120. Kleckner, N., Zickler, D., Jones, G. H., Dekker, J., Padmore, R., *et al.* A mechanical basis for chromosome function. *Proceedings of the National Academy of Sciences of the United States of America* 101(34):12592–7 (2004). doi:10.1073/pnas.0402724101.

121. King, J. S. and Mortimer, R. K. A polymerization model of chiasma interference and corresponding computer simulation. *Genetics* 126(4):1127–38 (1990).

122. de Boer, E., Dietrich, A. J. J., Höög, C., Stam, P., and Heyting, C. Meiotic interference among MLH1 foci requires neither an intact axial element structure nor full synapsis. *Journal of cell science* 120(Pt 5):731–6 (2007). doi:10.1242/jcs.003186.

123. Foss, E., Lande, R., Stahl, F. W., and Steinberg, C. M. Chiasma interference as a function of genetic distance. *Genetics* 133(3):681–91 (1993).

124. Foss, E. J. and Stahl, F. W. A test of a counting model for chiasma interference. *Genetics* 139(3):1201–9 (1995).

125. Housworth, E. and Stahl, F. Crossover Interference in Humans. *American Journal of Human Genetics* 73(1):188–197 (2003). doi:10.1086/376610.

126. de Boer, E., Stam, P., Dietrich, A. J. J., Pastink, A., and Heyting, C. Two levels of interference in mouse meiotic recombination. *Proceedings of the National Academy of Sciences of the United States of America* 103(25):9607–12 (2006). doi:10.1073/pnas.0600418103.

127. Barchi, M., Roig, I., Di Giacomo, M., de Rooij, D. G., Keeney, S., *et al.* ATM promotes the obligate XY crossover and both crossover control and chromosome axis integrity on autosomes. *PLoS genetics* 4(5):e1000076 (2008). doi:10.1371/journal.pgen.1000076.

128. Basheva, E. A., Bidau, C. J., and Borodin, P. M. General pattern of meiotic recombination in male dogs estimated by MLH1 and RAD51 immunolocalization. *Chromosome Research* 16(5):709–19 (2008). doi:10.1007/s10577-008-1221-y.

129. Mary, N., Barasc, H., Ferchaud, S., Billon, Y., Meslier, F., *et al.* Meiotic recombination analyses of individual chromosomes in male domestic pigs (Sus scrofa domestica). *PloS one* 9(6):e99123 (2014). doi:10.1371/journal.pone.0099123.

130. Vozdova, M., Sebestova, H., Kubickova, S., Cernohorska, H., Vahala, J., *et al.* A comparative study of meiotic recombination in cattle (Bos taurus) and three wildebeest species (Connochaetes gnou, C. taurinus taurinus and C. t. albojubatus). *Cytogenetic and genome research* 140(1):36–45 (2013). doi:10.1159/000350444.

131. Borodin, P. M., Karamysheva, T. V., and Rubtsov, N. B. Immunofluorescent analysis of meiotic recombination and interference in the domestic cat. *Tsitologiia* 50(1):62–6 (2008).

132. Borodin, P. M., Karamysheva, T. V., Belonogova, N. M., Torgasheva, A. A., Rubtsov, N. B., *et al.* Recombination map of the common shrew, Sorex araneus (Eulipotyphla, Mammalia). *Genetics* 178(2):621–32 (2008). doi:10.1534/genetics.107.079665.

133. Borodin, P. M., Basheva, E. A., and Zhelezova, A. I. Immunocytological analysis of meiotic recombination in the American mink (Mustela vison). *Animal genetics* 40(2):235–8 (2009). doi:10.1111/j.1365-2052.2008.01808.x.

134. Lian, J., Yin, Y., Oliver-Bonet, M., Liehr, T., Ko, E., *et al.* Variation in crossover interference levels on individual chromosomes from human males. *Human molecular genetics* 17(17):2583–94 (2008). doi:10.1093/hmg/ddn158.

135. Housworth, E. A. and Stahl, F. W. Is there variation in crossover interference levels among chromosomes from human males? *Genetics* 183(1):403–5 (2009). doi:10.1534/genetics.109.103853.

136. Zhao, H., McPeek, M. S., and Speed, T. P. Statistical analysis of chromatid interference. *Genetics* 139(2):1057–65 (1995).

137. Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., *et al.* A fine-scale chimpanzee genetic map from population sequencing. *Science* 336(6078):193–8 (2012). doi:10.1126/science.1216872.

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

139. Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. Genetic recom-

bination is directed away from functional genomic elements in mice. *Nature* 485(7400):642–5 (2012). doi:10.1038/nature11089.

140. Smagulova, F., Gregoretti, I. V., Brick, K., Khil, P., Camerini-Otero, R. D., *et al.* Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* 472(7343):375–8 (2011). doi:10.1038/nature09869.

141. Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* 190(2):389–401 (2012). doi:10.1534/genetics.111.132639.

142. Liu, E. Y., Morgan, A. P., Chesler, E. J., Wang, W., Churchill, G. A., *et al.* High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. *Genetics* 197(1):91–106 (2014). doi:10.1534/genetics.114.161653.

143. Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., *et al.* Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genetics* 5(12):e1000753 (2009). doi:10.1371/journal.pgen.1000753.

144. Ponting, C. P. What are the genomic drivers of the rapid evolution of PRDM9? *Trends in Genetics* 27(5):165–71 (2011). doi:10.1016/j.tig.2011.02.001.

145. Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., *et al.* Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics* 37(4):429–34 (2005). doi:10.1038/ng1529.

146. Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308(5718):107–11 (2005). doi:10.1126/science.1105322.

147. Thomas, J. H., Emerson, R. O., and Shendure, J. Extraordinary molecular evolution in the PRDM9 fertility gene. *PloS one* 4(12):e8505 (2009). doi:10.1371/journal.pone.0008505.

148. Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069):803–19 (2005). doi:10.1038/nature04338.

149. Muñoz-Fuentes, V., Di Rienzo, A., and Vilà, C. Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. *PloS One* 6(11):e25498 (2011). doi:10.1371/journal.pone.0025498.

150. Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., *et al.* Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genetics* 9(12):e1003984 (2013). doi:10.1371/journal.pgen.1003984.

151. Axelsson, E., Webster, M. T., Ratnakumar, A., Ponting, C. P., and Lindblad-Toh, K. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Research* 22(1):51–63 (2012). doi:10.1101/gr.124123.111.

152. Freixa, L., García, M., and Egozcue, J. The timing of first meiotic prophase in oocytes from female domestic dogs (Canis familiaris). *Genome* 29(1):208–210 (1987). doi:10.1139/g87-036.

153. Tsutsui, T. Gamete physiology and timing of ovulation and fertilization in dogs. *Journal of reproduction and fertility. Supplement* 39:269–75 (1989).

154. Chastant-Maillard, S., Viaris de Lesegno, C., Chebrout, M., Thoumire, S., Meylheuc, T., *et al.* The canine oocyte: uncommon features of in vivo and in vitro maturation. *Reproduction, Fertility, and Development* 23(3):391–402 (2011). doi:10.1071/RD10064.

155. Lam, I. and Keeney, S. Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* 350(6263):932–937 (2015). doi:10.1126/science.aad0814.

156. Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., *et al.* Stable recombination hotspots in birds. *Science* 350(6263):928–932 (2015). doi:10.1126/science.aad0843.

157. Imai, K., Kricka, L. J., and Fortina, P. Concordance Study of 3 Direct-to-Consumer Genetic-Testing Services. *Clinical Chemistry* 57(3):518–521 (2010). doi:10.1373/clinchem.2010.158220.

158. Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* 4:e04637 (2015). doi:10.7554/eLife.04637.

159. Jeffreys, A. J. and May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics* 36(2):151–6 (2004). doi:10.1038/ng1287.

160. Cole, F., Keeney, S., and Jasin, M. Preaching about the converted: how meiotic gene conversion influences genomic diversity. *Annals of the New York Academy of Sciences* 1267:95–102 (2012). doi:10.1111/j.1749-6632.2012.06595.x.

161. Li, N. and Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165(4):2213–33 (2003).

162. Wall, J. D. A Comparison of Estimators of the Population Recombination Rate. *Molecular Biology and Evolution* 17(1):156–163 (2000). doi:10.1093/oxfordjournals.molbev.a026228.

163. Fearnhead, P. and Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* 159(3):1299–318 (2001).

164. Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* 5(6):e1000519 (2009). doi:10.1371/journal.pgen.1000519.

165. Gay, J., Myers, S., and McVean, G. Estimating meiotic gene conversion rates from population genetic data. *Genetics* 177(2):881–94 (2007). doi:10.1534/genetics.107.078907.

166. Lander, E. S. and Green, P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* 84(8):2363–7 (1987).

167. Delaneau, O., Zagury, J.-F., and Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* 10(1):5–6 (2013). doi:10.1038/nmeth.2307.

168. Stephens, M., Smith, N. J., and Donnelly, P. A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal of Human Genetics* 68(4):978–989 (2001). doi:10.1086/319501.

169. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526(7571):68–74 (2015). doi:10.1038/nature15393.

170. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* 10(4):e1004234 (2014). doi:10.1371/journal.pgen.1004234.

Chapter 2

Escape from crossover interference increases with maternal age

Christopher L. Campbell[1], Nicholas A. Furlotte[2], Nick Eriksson[2,†], David Hinds[2], and Adam Auton[1]

This manuscript has been published:
Campbell, C. L. et al. Escape from crossover interference increases with maternal age. Nat. Commun. 6:6260 doi: 10.1038/ncomms7260 (2015).
PMCID: PMC4335350

[1] Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, New York 10461, USA.
[2] 23andMe Inc., Mountain View, California 94043, USA.
[†] Present address: Coursera, 381 East Evelyn Avenue, Mountain View, California 94041, USA.

Correspondence and requests for materials should be addressed to
A.A. (email: adam.auton@einstein.yu.edu).

Abstract

Recombination plays a fundamental role in meiosis, ensuring the proper segregation of chromosomes and contributing to genetic diversity by generating novel combinations of alleles. Here, we use data derived from direct-to-consumer genetic testing to investigate patterns of recombination in over 4,200 families. Our analysis reveals a number of sex differences in the distribution of recombination. We find the fraction of male events occurring within hotspots to be 4.6% higher than for females. We confirm that the recombination rate increases with maternal age, while hotspot usage decreases, with no such effects observed in males. Finally, we show that the placement of female recombination events appears to become increasingly deregulated with maternal age, with an increasing fraction of events observed within closer proximity to each other than would be expected under simple models of crossover interference.

2.1 Introduction

Recombination is a fundamental meiotic process that is required to ensure the proper segregation of chromosomes. In mammals and other eukaryotes, at least one crossover is normally required to ensure proper disjunction, and failures in recombination can result in deleterious outcomes such as aneuploidy. As such, the recombination process is highly regulated to ensure that sufficient numbers of crossovers occur. The placement of crossover events along a chromosome is also tightly regulated. At the fine scale, the majority of crossovers tend to occur within localized regions of $\sim$2 kb in width known as recombination hotspots. At broader scales, interference between crossovers appears to increase spacing between events occurring on the same chromosome during meiosis.

As relatively few crossover events occur within a single meiosis, quantifying the recombination landscape requires the observation of large numbers of meioses. In this study, we adopt a pedigree approach to study the properties of recombination in over 18,000 meioses using data derived from families genotyped via direct-to-consumer genetic testing. Our approach enables hundreds of thousands of recombination events to be localized, and allows us to investigate how the frequency and placement of recombination changes as a function of sex and parental age.

2.2 Results

To investigate properties of crossover placement in humans, we collected data from pedigree families contained within the database of 23andMe Inc. (Mountain View, CA). Our data set consists of 4,209 families contributing a total of 18,302 informative meioses genotyped at over 515,972 sites. To preserve the privacy of the participants, families were removed if the age of the mother was greater than 40 years at the time of childbirth, the age of the father was greater than 45 years or the difference between the parental ages was greater than 15 years (Supplementary Fig. 2.S1). The majority of the data is derived from family quartets (Supplementary Table 2.S1), accounting for 78.6% of the families, and is also predominately composed of individuals of European ancestry (Supplementary Table 2.S2). Ancestral populations are assigned to each individual by comparison with a set of reference populations (see Supplementary Methods).

To infer recombination events in nuclear families, we applied the Lander-Green algorithm as implemented within Abecasis *et al.*[1]. To guard against genotyping error, we curated the data to remove nearby recombination events that could be indicative of genotyping error (see Supplementary Methods; Supplementary Fig. 2.S2). This approach allowed us to identify over 645,000 well-supported crossover events, with the median event being localized to 28.2 kb (Supplementary Fig. 2.S3).

We inferred a mean of 41.6 autosomal recombination events per gamete in females (95% confidence interval (CI): 41.4-41.9) and 26.6 in males (95% CI: 26.5-26.7, Fig. 2.1a). The genetic map constructed from our data agrees well with those generated by previous studies (Fig. 2.1b; Supplementary Fig. 2.S4; Supplementary Table 2.S3). At the 5-Mb scale, the Pearson correlation between our map and that of deCODE[2] is $r^2 = 0.975$ and 0.983 for females and males, respectively. Likewise, our sex-averaged map has a correlation of $r^2 = 0.955$ with the HapMap map inferred from patterns of linkage disequilibrium (LD)[3]. At the chromosome scale, the map length is well predicated by the physical chromosome length ($r^2 = 0.991$ in females and 0.945 in males; Supplementary Fig. 2.S5).

Treating the overall recombination rate as a phenotype, we replicate genetic associations at genome-wide significance for RNF212, which is known to be essential for crossover-specific complexes[4], and within the vicinity of TTC5, which appears to replicate an association with CCNB1IP1 (Kong *et al.*[5]). Another association near SMEK1 also replicates discoveries elsewhere[5], but not at genome-wide significance (Supplementary Table 2.S4).

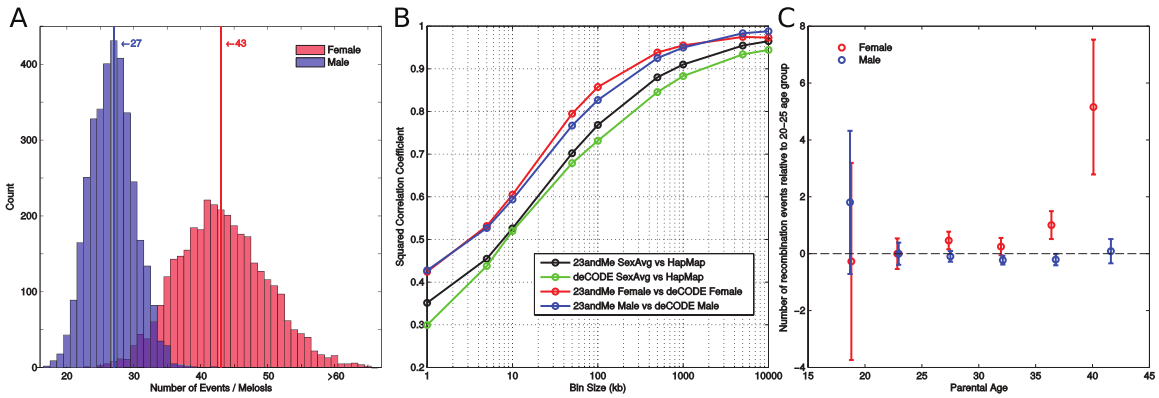

Figure 2.1: **Properties of recombination partitioned by sex and age.** (a) The number of events per meiosis for females (red, n = 9152) and males (blue, n = 9150), with median values indicated by a vertical line. For phase-unknown individuals, the average number of events per meiosis was used. (b) Squared Pearson correlation between the 23andMe map, the deCODE map and the HapMap map, as a function of scale. (c) The number of recombination events as a function of parental age for females (red, n = 9152) and males (blue, n = 9150), relative to parents of between 20 and 25 years of age. Parents were grouped into 5-year age bins, and the mean number of events estimated. Error bars show a 95% confidence interval for each group.

Previous reports have suggested increased recombination rates in older females[6,7]. Using linear regression (Supplementary Fig. 2.S6), we obtain a similar result with an additional 0.067 events per year being observed in females ($P = 0.002$, F-test), and no such effect being observed in males ($P = 0.30$, F-test). The female effect appears to be driven by sharp increase in the number of recombination events for older mothers (Fig. 2.1c). Fitting the piecewise-linear model with a single change point infers a rapid increase in the female recombination rate after 38.8 years, increasing from 0.047 events per year to 2.990 events per year. On average, mothers of 39 years and over have an additional 2.51 events compared with younger mothers ($P = 0.0005$, Mann-Whitney U).

One possible interpretation of the increasing number of recombination events with maternal age is that mothers with higher recombination rates can maintain fertility until a later age[6] . To investigate this possibility, we focused on 776 mothers (providing 2,184 meioses) that were part of larger families and could have recombination events assigned to specific children. After subtracting off the average age and average number of recombination events for each mother, the resulting regression does not find a significant association with age ($P = 0.11$, F-test), although we estimate our power to detect an effect size of an additional 0.067 events per year in this subsample to be no more than 30%.

Both pedigree and LD studies have suggested that ~60-70% of crossover events occur within recombination hotspots[7,8]. Our data confirm this result with 62.7% of events occurring within LD-defined hotspots in females, and 67.3% occurring within hotspots in males (Fig. 2.2a; Supplementary Fig. 2.S7A). The 4.6% difference between the two sexes is highly significant ($P = 1.1 \times 10^{-69}$, Mann-Whitney U), suggesting differences in the regulation of crossover placement between the sexes. The result remains significant after thinning the female data to match the crossover density of the male data ($P < 2.2 \times 10^{-16}$, Mann-Whitney U), and does not appear to be driven by increased male recombination rates near the telomeres (see Supplementary Methods).

Hotspot localization is believed to be under the control of the zinc-finger protein PRDM9, which recognizes and binds specific DNA motifs[9–12]. We find single-nucleotide polymorphisms (SNPs) in the vicinity of PRDM9 to be strongly associated with the degree of hotspot usage, as has previously been reported[5,11]. The most strongly associated SNP is rs73742307 achieving a P value of 7.9×10^{-184} (Reynolds *et al.* [4]), with no other region achieving a genome-wide significant association with this phenotype (Supplementary Table 2.S5).

Variation within the PRDM9 DNA-binding domain can result in changes to the recognized motif

and hence lead to differences in hotspot localization between individuals. While the major allele of PRDM9 (allele A) is present at high frequency in most human populations, a large number of low-frequency alleles have been observed, particularly within African populations[10,12]. Consistent with this, we find hotspot usage to be significantly lower within individuals of African ancestry (Fig. 2.2b; Supplementary Table 2.S6), which reflects the fact that the LD-defined hotspots are expected to mostly represent the common PRDM9 allele. Notably, while over 75% of our data are derived from individuals of European ancestry, hotspot usage is higher for males than females across all ancestries.

We find a weak association between hotspot usage and maternal age (Supplementary Fig. 2.S7B). Using logistic regression, we estimate a decrease in hotspot usage corresponding to ~1% over a 10-year period ($\beta_1 = -0.0042$, s.e. $= 9.6 \times 10^{-4}$, $P = 1.2 \times 10^{-5}$, F-test). To ensure this effect is not driven by differences in parental ancestry within the sample, we repeated the analysis only using individuals of European ancestry. In this case, the effect size remains similar ($\beta_1 = -0.0033$, s.e. $= 0.0013$), but is only marginally significant ($P = 0.0101$, F-test). Including the number of events as an additional predictor variable within the regression leaves age as a weakly significant predictor ($P = 0.0106$, F-test), but not the number of events ($P = 0.74$, F-test). Despite the small size of the estimated effect, we note that no such age-related effects were observed in males.

To learn more about interactions between recombination events, we used the high number of crossover locations in our data to better characterize the phenomenon of crossover interference. By considering the distribution inter-crossover distances, we fit three models to describe the distribution of inter-crossover distances: a model without interference between crossovers (also known as the gamma model of crossover interference[13]), and a mixture model in which a subset of events come from a process that exhibits no interference (also known as the Housworth-Stahl model[14]). To fit these models, we used existing methods for families in which recombination events could be assigned to specific individuals, and extended these methods for smaller families where recombination events cannot be simply assigned to a specific individual (see Supplementary Methods).

In agreement with previous reports[14,15], the Housworth-Stahl interference escape model provides a much better fit to our data than either the gamma simple interference model or the interference-free model (Fig. 2.3a). Under this model, the estimates of the strength of crossover interference are similar to previous reported using smaller data sets[15]. The degree of interference is inferred to be lower in

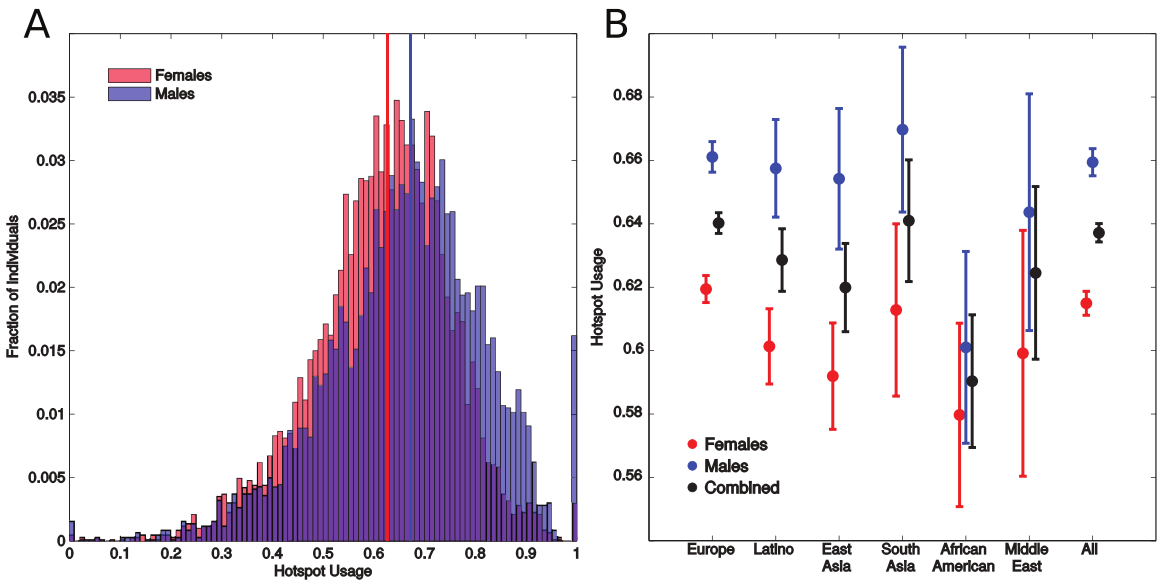

Figure 2.2: **Sex differences in recombination hotspot usage.** (a) Hotspot usage for female (red, n = 9152) and male (blue, n = 9150) meioses. Median values for each sex are shown by vertical lines. (b) Mean hotspot usage, subdivided by parental population. Females are shown in red, males in blue and a combined estimate in black. Error bars indicate a 95% confidence interval.

females than in males (ν_{female} = 7.19 vs $\nu male$ = 8.93). In addition, 7.8%/6.7% of female/male events are inferred to escape interference. We therefore conclude that a non-negligible fraction of crossovers occur in the absence of crossover interference.

We find evidence that both the degree of interference and interference escape varies across chromosomes (Fig. 2.3b,c; Supplementary Table 2.S7). The strength of interference is reasonably well predicted by the chromosome map length ($r^2 = 0.565$, $P = 6.4 \times 10^{-9}$), although the relationship is only significant in females when considering the sexes separately ($r^2_{female} = 0.69$, $P = 1.7 \times 10^{-6}$ and $r^2_{male} = 0.172$, $P = 0.06$; Supplementary Fig. 2.S8). In contrast, the fraction of events escaping interference shows no relationship with chromosome map length ($r^2 = 0.001$, $P = 0.84$). Certain chromosomes appear to have high degrees of escape, with chromosomes 8, 9 and 16 (in females) being notable outliers.

To investigate whether crossover interference changes with parental age, we subdivided our data into 10 quantiles on the basis of age, and fit the Housworth-Stahl interference escape model to each group independently. We observe a striking increase in the proportion of events that escape interference with maternal age (Fig. 2.4a), rising from 6.7% for mothers under 25 years to 9.5% for mothers over 35 years. No such correlation is observed for the interference parameter in females, and no correlation is observed for either parameter in males (Supplementary Fig. 2.S9). The effect is robust to different subdivisions of the data (Supplementary Figs 2.S10 and 2.S11).

A potential concern is that the detected increase in interference escape could be driven by the observed increased number of crossovers in older mothers. If the number of crossovers is increased, then the distances between them are necessarily shorter, which may in turn influence the interference parameter estimates. To account for this possibility, we performed stratified sampling of individuals to control for the number of events within each quantile. The observed increase in the escape parameter with maternal age is still observed (Supplementary Fig. 2.S12), indicating that it is not driven by changes in the overall recombination rate.

To further investigate the differences between old and young parents, we plotted the distribution of inter-crossover distances for young and old parents (Fig. 2.4b,c). The interference escape effect in females appears to be predominately driven by an increase in the number of very tightly clustered events, generally separated by less than ~5 cM. These tightly clustered events are not well captured by the Housworth-Stahl interference escape model (Supplementary Fig. 2.S13), and a major concern

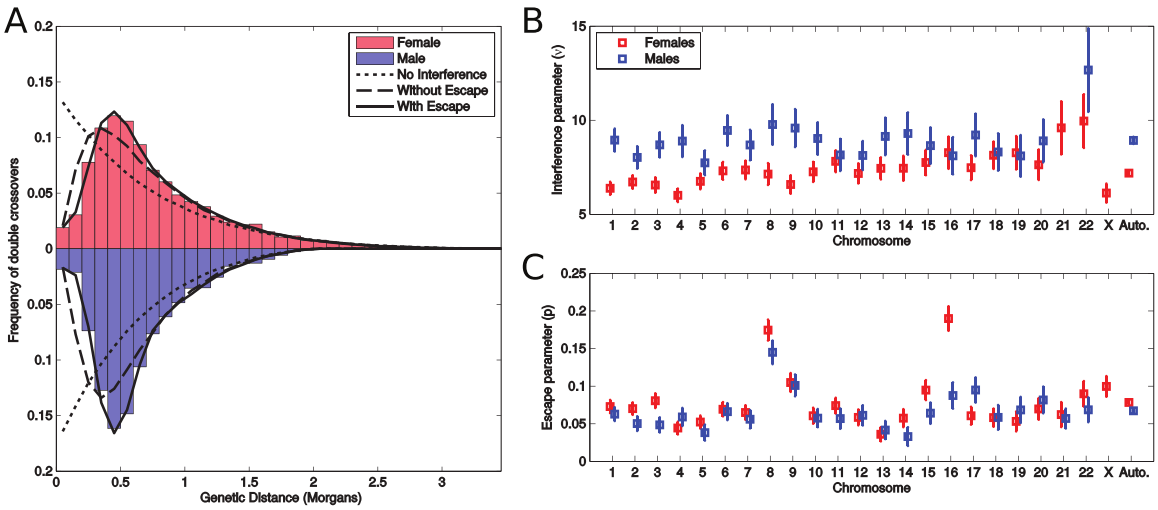

Figure 2.3: **Estimation of crossover interference parameters.** (a) Fit of three models of interference to the inter-crossover distances observed on chromosome 1, derived from phase-known mothers (red, n = 2184) and fathers (blue, n = 2092). The interference-free model is shown as a dotted line, the gamma simple interference model is shown as a dashed line and the Housworth-Stahl interference escape model is shown as a solid line. (b) Per-chromosome estimates of the interference parameter as estimated from the Housworth-Stahl interference escape model. Error bars indicate a 95% confidence interval. Note that chr21 in males is excluded due to an extremely high estimate. (c) Per-chromosome estimates of the proportion of events escaping interference. Error bars indicate a 95% confidence interval.

therefore is that these tightly clustered events represent false-positive calls arising from genotyping error. However, the effect remains even if we apply much stricter filtering of the crossover events (Supplementary Fig. 2.S14), and in addition we believe genotyping error is unlikely to explain the association between the escape parameter and maternal age because (a) the effect is not seen in males, and (b) it would imply increased genotyping error for older mothers (but not fathers).

In terms of meiosis, a major difference between the sexes is that female meiosis starts during fetal development, but does not complete until adulthood. As such, while male gametes are produced throughout adulthood and promptly proceed through meiosis, oocytes remain arrested in a late stage of prophase (dictyotene) for many years, if not decades. Presuming our observation of increasing crossover interference escape with maternal age is not due to some obscure form of genotyping error, our observations add to similar evidence of increasing rates of recombination[6] and aneuploidy[16] in aging females. Although these phenomena are presumably related, the biological mechanisms by which they occur are unclear, and we can think of at least three possibilities. First, given chromatids remain physically proximal during the extended period of female meiotic arrest, one possible explanation is that additional recombinations are initiated during this time, perhaps in response to DNA damage. However, as recombination is believed to have completed by the time of dictyotene, such an explanation appears unlikely. A second possibility, previously invoked to explain the increasing recombination rate with maternal age[6], suggests oocytes with additional recombination events could be at reduced risk of nondisjunction, and hence would be more likely to lead to viable embryos in older mothers. However, it is not clear that this mechanism would explain the increased clustering of events observed in our data. Finally, a third possibility is related to the so-called “production line” hypothesis, in which oocytes are selected for maturation sequentially in the same order as their generation, and later oocytes have therefore potentially undergone additional mitotic divisions prior to entering meiosis[17]. However, the existence of a production line has been debated for many years[17–19], and so the likelihood of this explanation is unclear.

2.3 Methods

Sample genotyping. Samples were collected and genotyped at the consumer genetics company 23andMe Inc., as described previously[20]. Briefly, genotyping was performed on genomic DNA ex-

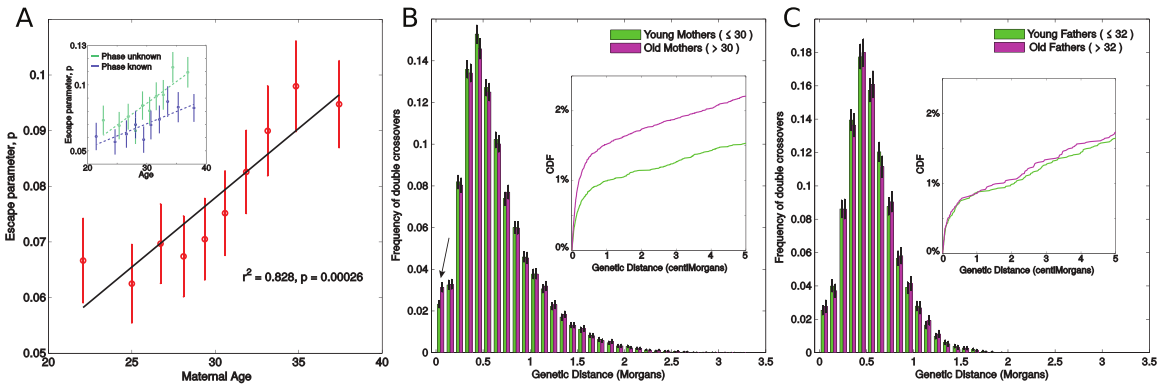

Figure 2.4: **Departures from simple crossover interference.** (a) Inferred escape parameter as a function of maternal age. Mothers were divided into 10 approximately equal-sized deciles on the basis of age, and the Housworth-Stahl interference escape model was fitted for each group separately. The inset shows the estimates of the escape parameter when considering phase-known (blue, n = 2184) and phase-unknown (green, n = 6968) individuals separately. Estimates for ν show no correlation with age (Supplementary Fig. 2.S9). Error bars indicate 95% confidence intervals. (b) Distribution of inter-crossover distances for young and old mothers, where the boundary between young and old is taken as median maternal age (30 years). Error bars represent a 95% confidence interval assessed via 1000 bootstrap samples, and the arrow highlights a significant difference between the young and old groups for tightly clustered events. The inset shows the cumulative distribution function (CDF) up to 5 cM. (c) Distribution of inter-crossover distances for young and old fathers, where the boundary between young and old is taken as median paternal age (32 years).

tracted from saliva samples. DNA was genotyped on one of two microarray platforms: the Illumina HumanHap550+ BeadChip platform, which includes more than 550,000 SNPs, or the Illumina HumanOmniExpress+ BeadChip, which has a base set of 730,000 SNPs augmented with ∼250,000 SNPs to obtain a superset of the HumanHap550+, as well as a custom set of about 30,000 SNPs.

Pedigree construction. Pedigrees were constructed first by identifying trios using estimated identity-by-decent relationships. Trios were then combined to form nuclear families, and nuclear families were joined based on the assumed relationships to form larger pedigrees. We identified trios by finding triplets of individuals in the 23andMe customer cohort that had estimated identity-by-decent relationships matching those expected in a true trio. Trios were accepted if both parents were at least 18 years old upon the birth of the child and one parent was male and the other female. We created nuclear families by identifying all trios with the same two parents and then by combining the children of these trios. Finally, larger pedigrees were created by simply joining the nuclear families based on the assumed relationships and by accounting for directionality given by the age of individuals. Any two individuals with more than one potential relationship were excluded along with the pedigrees they belonged to.

Calling of recombination events and data filtering. Prior to data filtering, the data set consisted of 4,270 pedigree families, with data pertaining to 18,647 informative meioses. This raw data set consisted of 692,876 recombination events, with a median of 45 and 28 events per meiosis in females and males, respectively.

The Merlin algorithm (version 1.1.2) used to detect recombination events does not account for genotyping error, and genotyping errors are therefore likely to result in spurious recombination event calls. To account for this issue, our first step was to only use high-confidence sites. First, we required the sites to have a call rate greater than 90% and Hardy-Weinberg P value $\leq 1 \times 10^{-20}$ (as calculated in the 23andMe cohort). Second, we excluded sites with minor allele frequencies differing from those of the 1000 Genomes Phase 1 reference panel[21]. This was achieved by constructing a 2×2 contingency table and comparing the 1000 Genomes European allele counts with those from 2,000 randomly selection 23andMe customers, and using a χ^2-test to identify significant deviations. Sites with P values less than 1×10^{-15} were removed.

Having applied these basic site filters, we next aimed to remove any weakly supported recombination events. This was achieved by first using the Merlin "error" feature to remove potential genotyping errors not consistent with gene flow within each pedigree. In addition, we excluded all recombination events supported by less than three recombination-informative sites on either side, where we define an informative site as a site that is called as heterozygotic in exactly two individuals out of each mother-father-child trio. Finally, we removed all pairs of events within each single family that occurred within the same SNP interval. Together, these filters removed 31,742 weakly supported events, which corresponded to 4.6% of the total number.

Preliminary inspection of the genetic maps identified a region on chromosome 10 where the 23andMe genetic map diverged substantially from that generated by deCODE[2]. This can be seen in a plot of the chromosome 10 genetic map at ~50 Mb (Supplementary Fig. 2.S2A).

Further investigation of this region revealed a large number of "double" crossovers in close proximity to each other (that is, pairs of recombination events occurring in close proximity within the same individual). While some such observations are expected through the action of gene conversion, such strong clustering of these events is not expected biologically. Instead, we believe the result is suggestive of misplacement of polymorphisms, mis-assembly of one or more reference contigs in the hg19 reference genome or of more complex types of error related to copy number polymorphism or array design. In any case, these double-recombination events represent a form of error that needed to be eliminated.

To better quantify this issue, we identified all pairs of recombination events occurring within a single individual that were within 1 Mb of each other. For each SNP in the genome, we estimated the number of these event pairs that span the SNP (Supplementary Fig. 2.S2B).

For the vast majority of the genome, there were very few such event pairs, and hence localized peaks likely represent data quality issues. We therefore identified all SNPs spanned by at least 14 event pairs (with this threshold being equivalent to the 99.9th percentile of the distribution). In this way, we identified 50 regions with strong enrichment of nearby event pairs (Supplementary Table 2.S8). Note that for this analysis we ignored the pseudoautosome, as a large number of events occurring in close proximity might be expected due to the extreme male recombination rate within this region.

The regions with high numbers of clustered events were themselves clustered into 13 regions across 8 chromosomes, and are often in the vicinity of chromosome centromeres, telomeres or ref-

erence assembly gaps. We removed all event pairs within 500 kb of the region boundaries described in Supplementary Table 2.S8, which resulted in the removal of 2,916 events (0.42% of the total). The removal of these events improved the concordance between the 23andMe and deCODE maps (Supplementary Fig. 2.S2C).

Previous research using well-curated data in 728 meioses reported an average of 39.6 autosomal events per gamete in females (95% CI 38.5-40.6), and 26.2 autosomal events per gamete in males (95% CI 25.6-26.7)7. The minimum/ maximum number of observed autosomal events in any given meiosis in this data was 19/71 for females, and 16/43 for males (Graham Coop, personal communication).

Preliminary analysis of our data revealed a small subset of individuals had biologically unrealistic numbers of recombination events. Our first filtering step was to remove the pedigrees containing these individuals. Specifically, we removed individuals (and their containing pedigrees) that were more than 5 s.d. from the (sex specific) median number of recombination events. To guard against outliers, we used a robust estimate of the s.d. taken as $\sigma = 1.4826$ MAD, where MAD represents the median absolute deviation.

Before filtering, the median number of recombination events was 43 and 27 for females and males, respectively (including chrX and the pseudoautosome). Using the $\pm 5\sigma$ thresholds, we removed pedigrees containing any female with fewer than 10 or more than 76 events per meiosis, or any male with fewer than 9 or more than 45 events. These filters removed a total of 52 pedigrees.

Summary of the filtered data set. After applying the filtering steps described above, the filtered data set consists of 4,209 pedigrees containing 18,302 informative meioses, of which 9,152 are from females and 9,150 are from males. Of the families included in the study, 78.6% are family quartets, 14.3% are larger one-generation families, and 7.1% are two-generation families (Supplementary Table 2.S1).

Due to the structure of the pedigrees included in the study, certain recombination events can be identified as having occurred within a specific child, whereas others cannot. For example, in family quartets, it is generally unclear which child has the recombinant haplotype, and we therefore refer to these events as “phase unknown”. Conversely, the child containing the recombinant haplotype can generally be identified in larger pedigree families in which the parental haplotype can be confidently

phased, and we therefore refer to these events as “phase known”.

In total, 4,276 meioses are derived from phase-known individuals, whereas 14,026 are derived from phase-unknown individuals. Of the female meioses, 2,184 are derived from phase-known mothers and 6,968 are from phase-unknown mothers. Of the male meioses, 2,092 are derived from phase-known fathers, and 7,058 are derived from phase-unknown fathers.

Individuals were assigned to high-level population groups via comparison with a set of reference populations (see Supplementary Methods). The majority of individuals in the data set are of European descent, with ∼78% of the meioses in the sample occurring within a European individual (Supplementary Table 2.S2).

The parental age distribution for the filtered data set is shown in Supplementary Fig. 2.S1. The mean age was 30 years for females, and 32 for males.

The final filtered data set consists of 645,853 recombination events. Including the sex chromosomes, the mean number of recombination events was 43.47 for females (σ = 6.64, 95% CI 43.25-43.69), and 27.04 for males (σ = 3.28, 95% CI 26.94-27.16). For the autosomes alone, the mean number of recombination events was 41.64 for females (σ = 6.34, 95% CI 41.43-41.85) and 26.61 for males (σ = 3.26, 95% CI 26.51-26.73).

The distribution of interval sizes to which crossovers could be resolved (that is, the distance between informative markers on either side of the recombination event) is given in Supplementary Fig. 2.S3. Crossovers could be resolved within a median distance of 28.2 kb.

2.4 Acknowledgements

We would like to thank Hilary Martin and Julie Hussin for their constructive comments regarding earlier versions of this manuscript. C.L.C. was supported by the Training Program in Cellular and Molecular Biology and Genetics, T32 GM007491. Work by N.A.F., N.E. and D.H. was supported by NIH award 2R44HG006981-02.

2.5 Author contributions

A.A., D.H. and N.E. designed the study. A.A., C.L.C. and N.A.F. conducted analysis. A.A. and C.L.C. wrote the paper.

2.6 Additional information

Supplementary Information accompanies this paper at

http://www.nature.com/naturecommunications

Competing Financial Interests: N.A.F. and D.H. are current employees, and N.E. is a former employee of 23andMe Inc., and have private equity interest. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at

http://npg.nature.com/reprintsandpermissions/

How to cite this article: Campbell, C. L. *et al*. Escape from crossover interference increases with maternal age. *Nat. Commun.* 6:6260 doi: 10.1038/ncomms7260 (2015).

2.7 References

1. Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 30(1):97–101 (2002). doi: 10.1038/ng786.

2. Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099–103 (2010). doi:10.1038/nature09525.

3. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61 (2007). doi:10.1038/nature06258.

4. Reynolds, A., Qiao, H., Yang, Y., Chen, J. K., Jackson, N., *et al.* RNF212 is a dosage-sensitive regulator of crossing-over during mammalian meiosis. *Nature Genetics* 45(3):269–78 (2013). doi:10.1038/ng.2541.

5. Kong, A., Thorleifsson, G., Frigge, M. L., Masson, G., Gudbjartsson, D. F., *et al.* Common and low-frequency variants associated with genome-wide recombination rate. *Nature Genetics* 46(1):11–16 (2014). doi:10.1038/ng.2833.

6. Kong, A., Barnard, J., Gudbjartsson, D. F., Thorleifsson, G., Jonsdottir, G., *et al.* Recombination rate and reproductive success in humans. *Nature Genetics* 36(11):1203–6 (2004). doi:10.1038/ng1445.

7. Coop, G., Wen, X., Ober, C., Pritchard, J. K., and Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319(5868):1395–8 (2008). doi:10.1126/science.1151851.

8. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310(5746):321–4 (2005). doi:10.1126/science.1117196.

9. Berg, I. L., Neumann, R., Lam, K.-W. G., Sarbajna, S., Odenthal-Hesse, L., *et al.* PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics* 42(10):859–63 (2010). doi:10.1038/ng.658.

10. Berg, I. L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N. J., *et al.* 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):12378–83 (2011). doi:10.1073/pnas.1109531108.

11. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., *et al.* The landscape of recombination in African Americans. *Nature* 476(7359):170–5 (2011). doi:10.1038/nature10336.

12. Parvanov, E. D., Petkov, P. M., and Paigen, K. Prdm9 controls activation of mammalian recombination hotspots. *Science* 327(5967):835 (2010). doi:10.1126/science.1181495.

13. Broman, K. W. and Weber, J. L. Characterization of human crossover interference. *American Journal of Human Genetics* 66(6):1911–26 (2000). doi:10.1086/302923.

14. Housworth, E. A. and Stahl, F. W. Crossover interference in humans. *American Journal of Human Genetics* 73(1):188–97 (2003). doi:10.1086/376610.

15. Fledel-Alon, A., Wilson, D. J., Broman, K., Wen, X., Ober, C., *et al.* Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genetics* 5(9):e1000658 (2009). doi:10.1371/journal.pgen.1000658.

16. Hassold, T. and Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics* 2(4):280–91 (2001). doi:10.1038/35066065.

17. Reizel, Y., Itzkovitz, S., Adar, R., Elbaz, J., Jinich, A., *et al.* Cell lineage analysis of the mammalian female germline. *PLoS Genetics* 8(2):e1002477 (2012). doi:10.1371/journal.pgen.1002477.

18. Polani, P. E. and Crolla, J. A. A test of the production line hypothesis of mammalian oogenesis. *Human Genetics* 88(1):64–70 (1991). doi:10.1007/BF00204931.

19. Rowsey, R., Gruhn, J., Broman, K. W., Hunt, P. a., and Hassold, T. Examining variation in

recombination levels in the human female: a test of the production-line hypothesis. *American Journal of Human Genetics* 95(1):108–12 (2014). doi:10.1016/j.ajhg.2014.06.008.

20. Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., *et al.* Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genetics* 6(6):e1000993 (2010). doi:10.1371/journal.pgen.1000993.

21. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65 (2012). doi:10.1038/nature11632.

2.8 Supplementary Figures

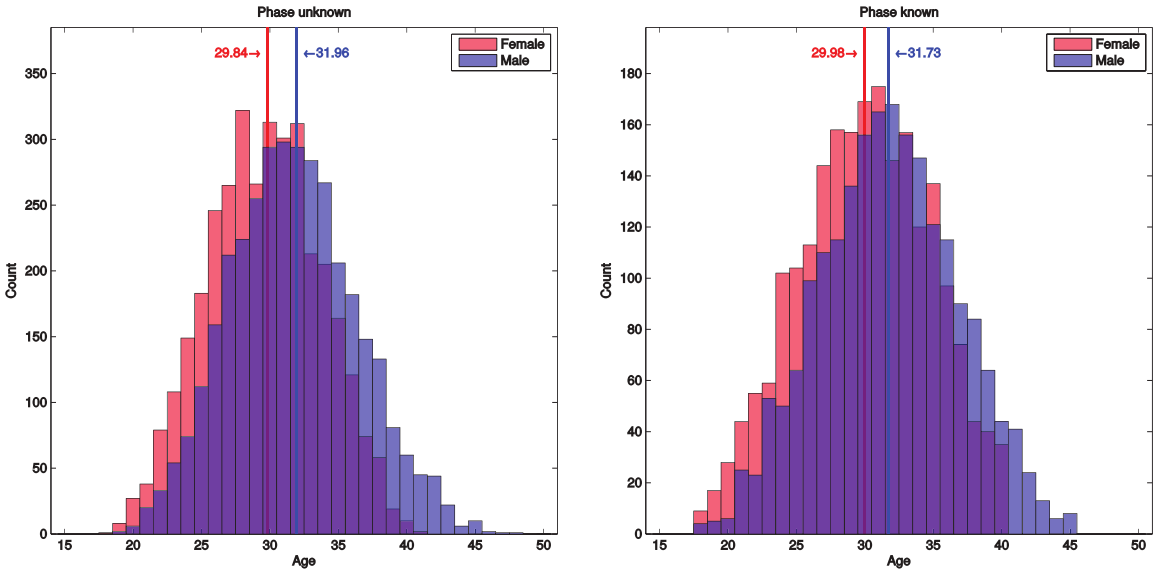

Figure 2.S1: **Age distributions within the filtered dataset.** The left hand panel shows the distribution for phase-unknown individuals, where the parental ages were averaged across children. The right hand panel shows data for the phase-known meioses where the parental age at the time of childbirth is known. Lines indicate the mean of each distribution. Note that some families were excluded from analysis by 23andMe on the basis age to protect privacy, as seen from the truncated distribution of maternal ages in the right hand panel.

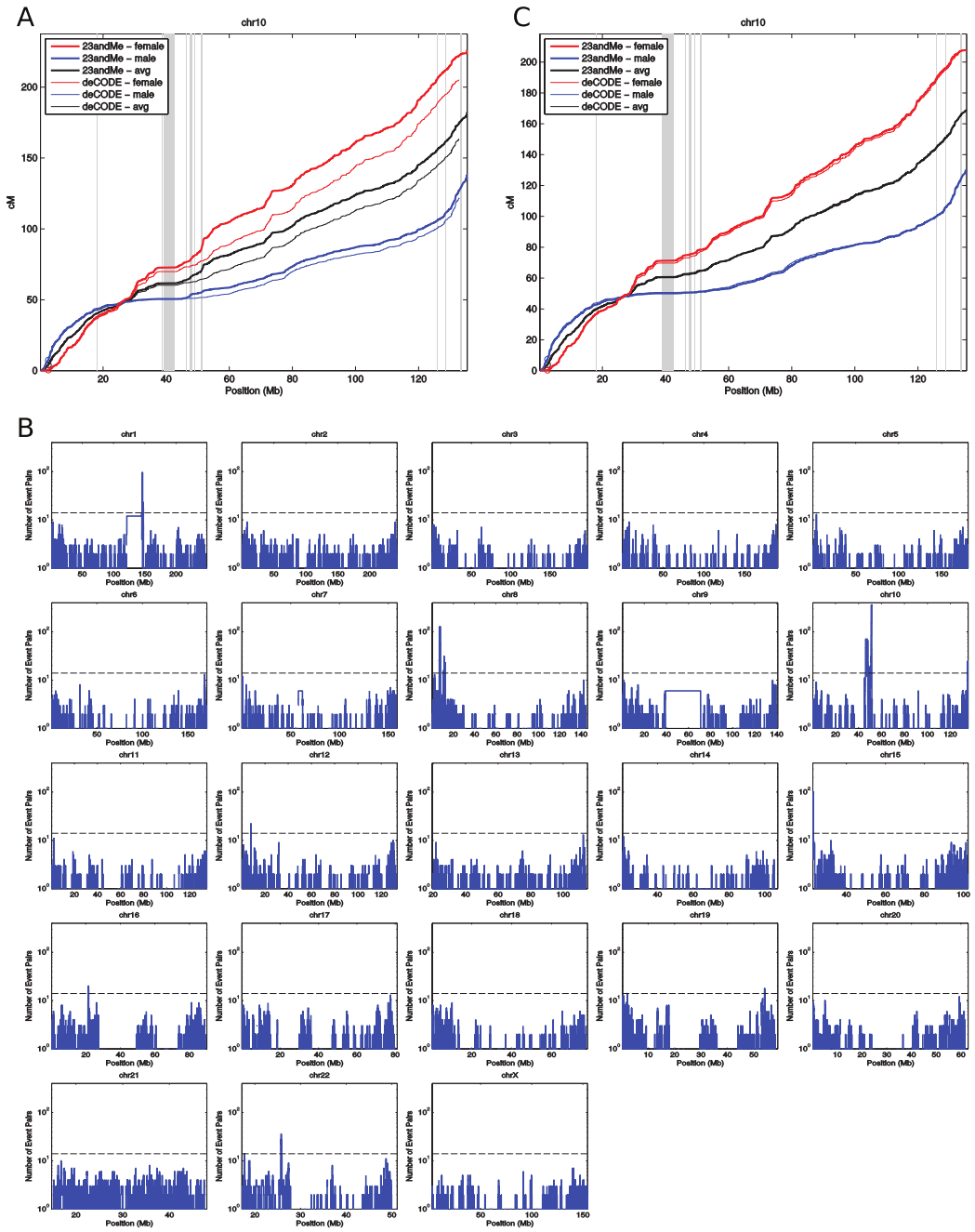

Figure 2.S2: **Data grooming.** A) Chromosome 10 map before filtering. Genetic maps from the 23andMe data are shown in bold lines, whereas the genetic maps from deCODE are shown as thin lines. Separate maps are shown for females(red), males (blue), and sex-averaged (black). Also shown are regions highlighted in grey that represent gaps in the reference assembly, the largest of which being the centromere at around 40Mb. B) Clustering of recombination events occurring within 1 Mb of each other within single individuals. Each plot shows the number of events within 1 Mb of each other on a $\log_{10}$ scale as a function of physical position on each chromosome. A large number of these event pairs can be observed on chromosome 10, although other large peaks can also be observed on, for example, chromosomes 8 and 15. The dashed line represents the 99.9% percentile of the distribution, and was used as a threshold for filtering. C) Chromosome 10 map after filtering.

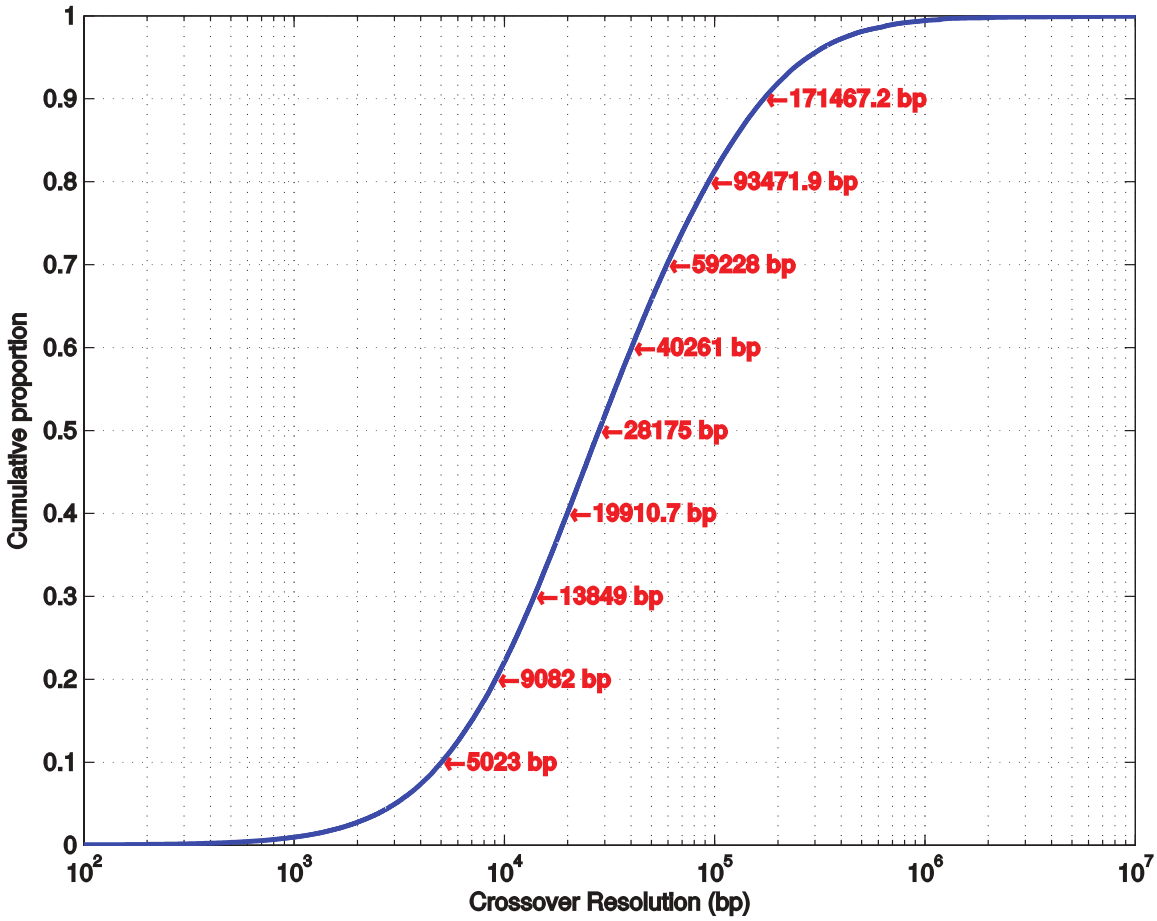

Figure 2.S3: **Empirical cumulative distance function of crossover localization distances.** Red labels indicate the interval distances at the distribution deciles.

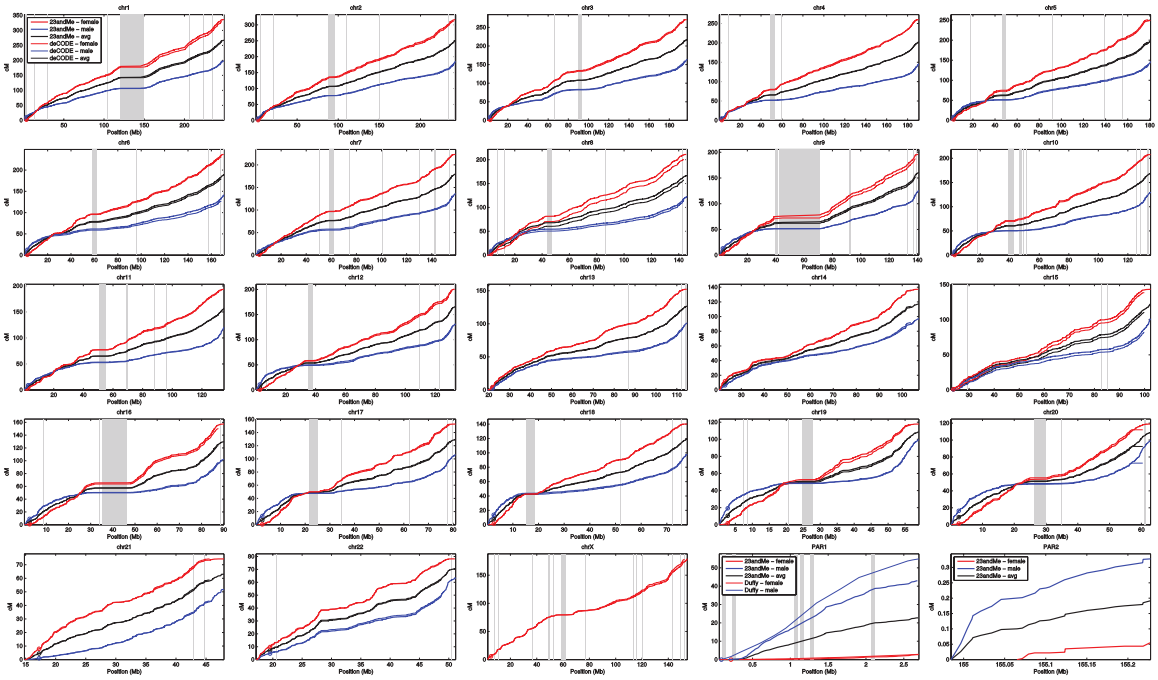

Figure 2.S4: **Genetic map estimated from 23andMe data.** Genetic maps from the 23andMe data are shown in bold lines, whereas the genetic maps from deCODE are shown as thin lines. Separate maps are shown for females (red), males(blue), and sex-averaged (black). Also shown are regions highlighted in grey that represent gaps in the reference assembly. For PAR1, we are showing data derived from Duffy[1] for comparison. As the deCODE maps cover a slightly smaller physical region than the 23andMe maps, the deCODE maps have been shifted slightly upwards to aid visual comparison. Specifically, the deCODE map has been aligned with the 23andMe map at the first physical position within the deCODE map. The locations of the alignments are indicated by small circles that can be most clearly seen on the smaller chromosomes.

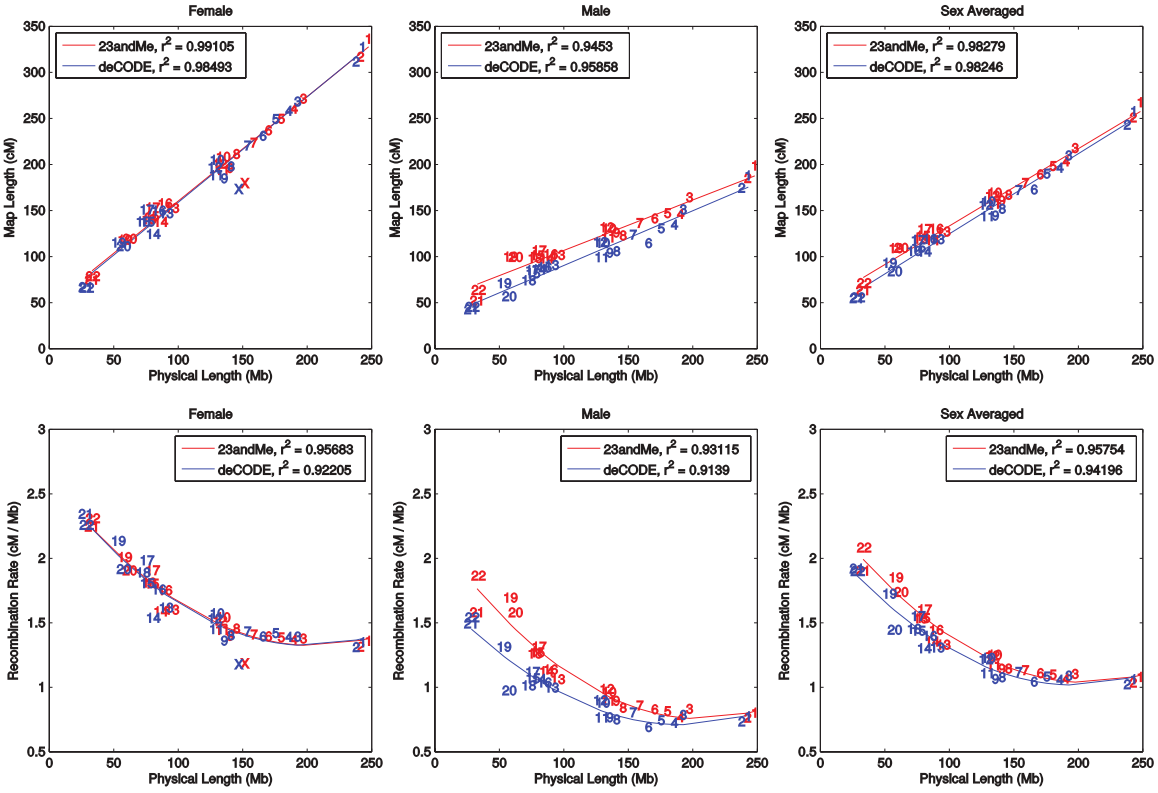

Figure 2.S5: **The relationship between chromosome length and recombination.** The top row shows the correlation between physical length and map length for females (left), males (center), and sex averaged (right), with a linear fit included for the 23andMe map (red) and the deCODE map (blue). The bottom row shows the relationship between physical length and average recombination rate with a quadratic fit. Note that chromosome X has been included in the female plots, but was excluded from the regressions.

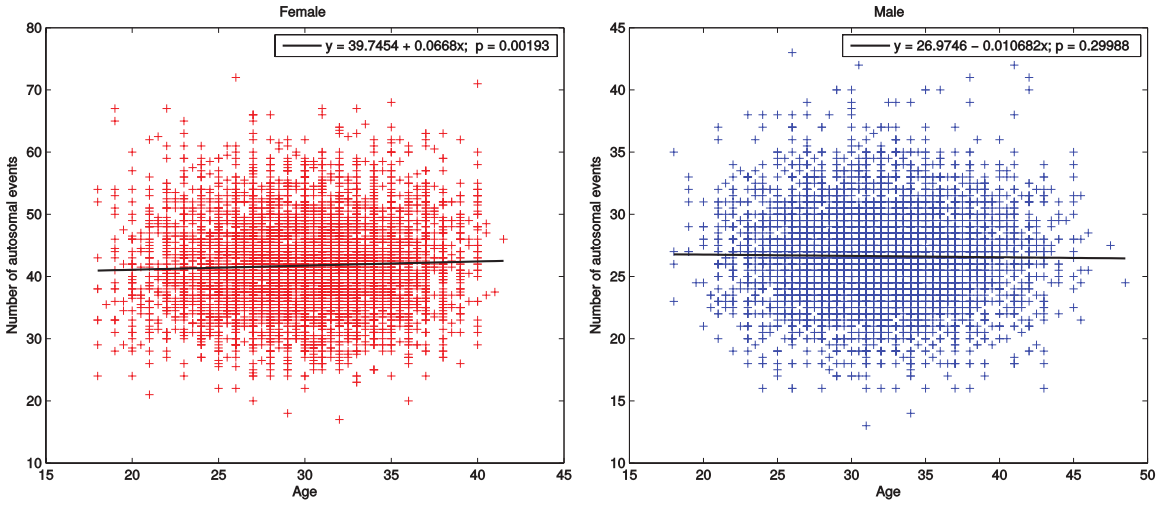

Figure 2.S6: **Number of autosome recombination events versus parental age** for females (left) and males (right). A linear least-squares fit is indicated by a black line. The least-squares fit equation given in the legend together with a p-value for the non-constant term.

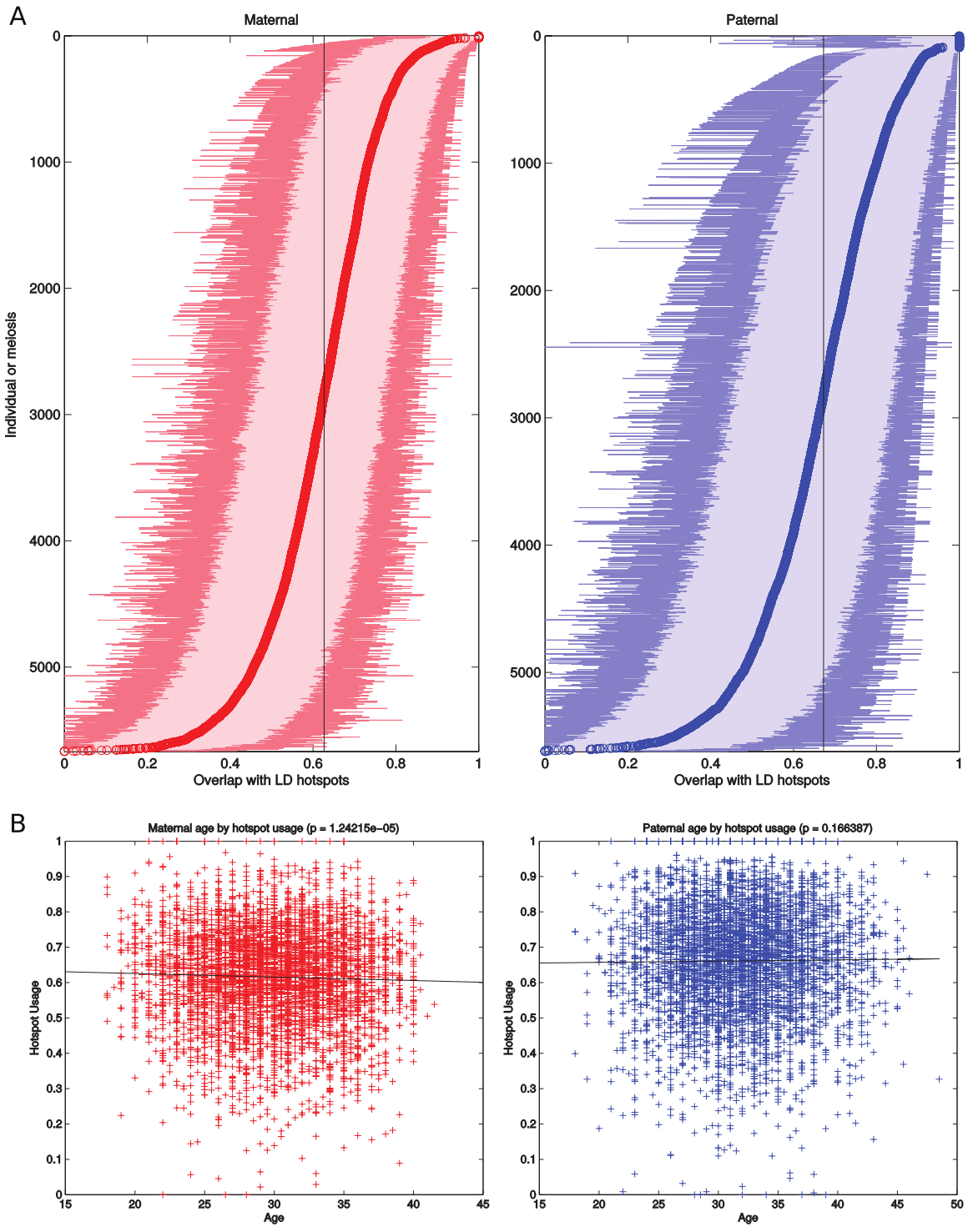

Figure 2.S7: A) Hotspot usage estimated in females (left) and males(right). The MLE estimate for each individual is indicated by a circle, with a 95% confidence interval indicated by the shaded area. The median MLE estimate for each sex is indicated by a vertical black line. B) Hotspot usage by parental age for females(left) and males (right). For each plot a logistic regression is also shown, with the p-value for the non-constant term given in the title.

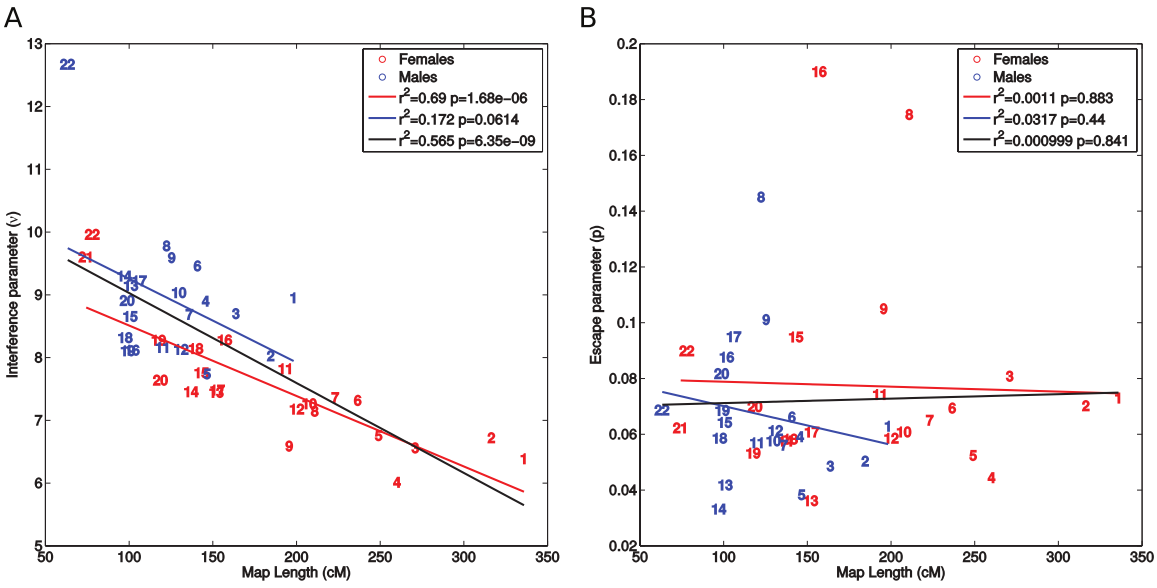

Figure 2.S8: A) The relationship between chromosome map length and the interference parameter, ν. B) The relationship between chromosome map length and the escape parameter, p. Linear fits are shown for females (red), males (blue), and the data combined across sexes (black). In both plots, the chr21 estimate in males has been excluded.

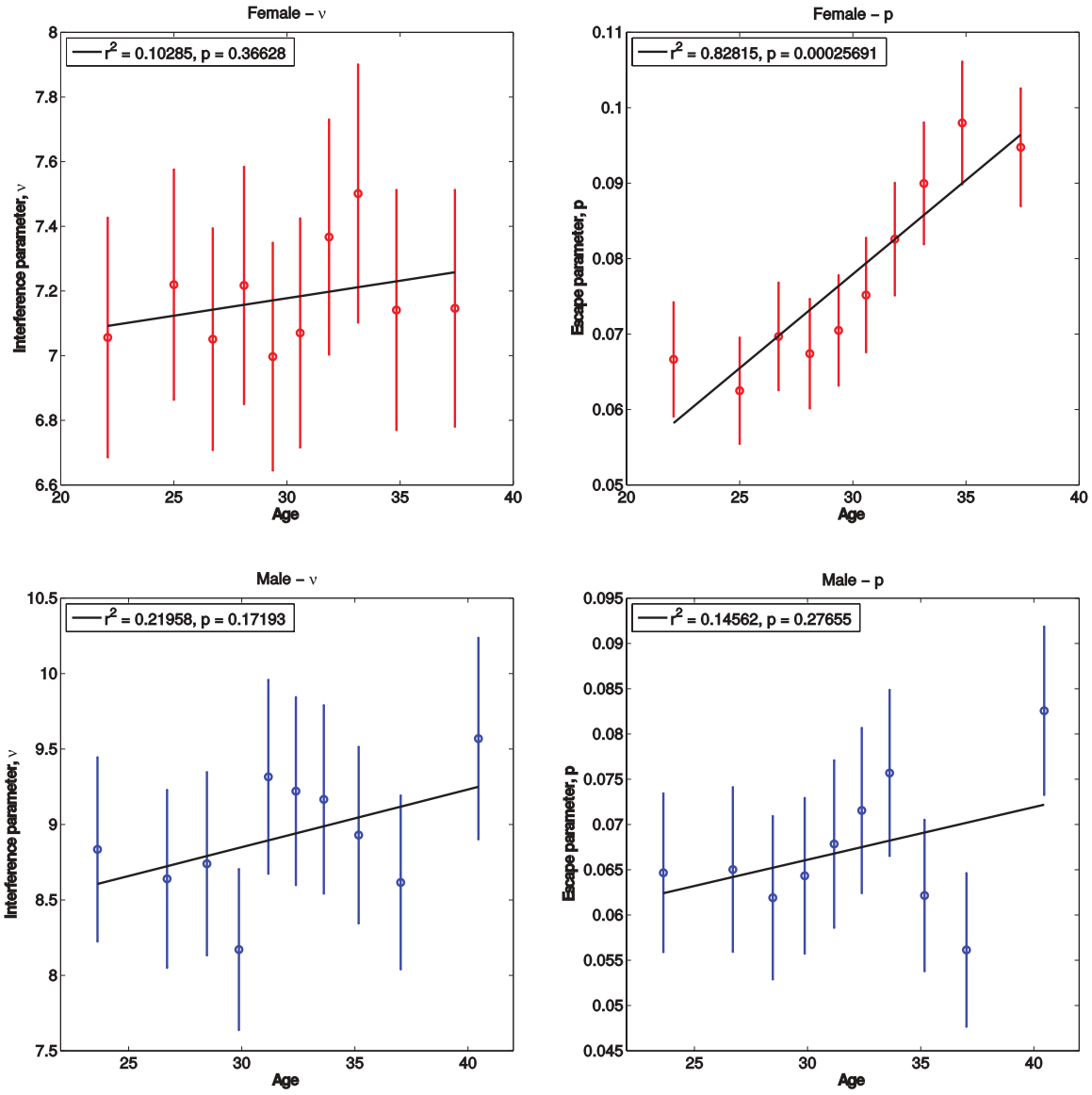

Figure 2.S9: **Interference parameters as a function of age.** Females and males are shown on the top and bottom rows respectively. Estimates of the interference parameter, ν, are shown on the left, whereas estimates of the escape parameter, p, are shown on the right. Error bars show 95% confidence intervals.

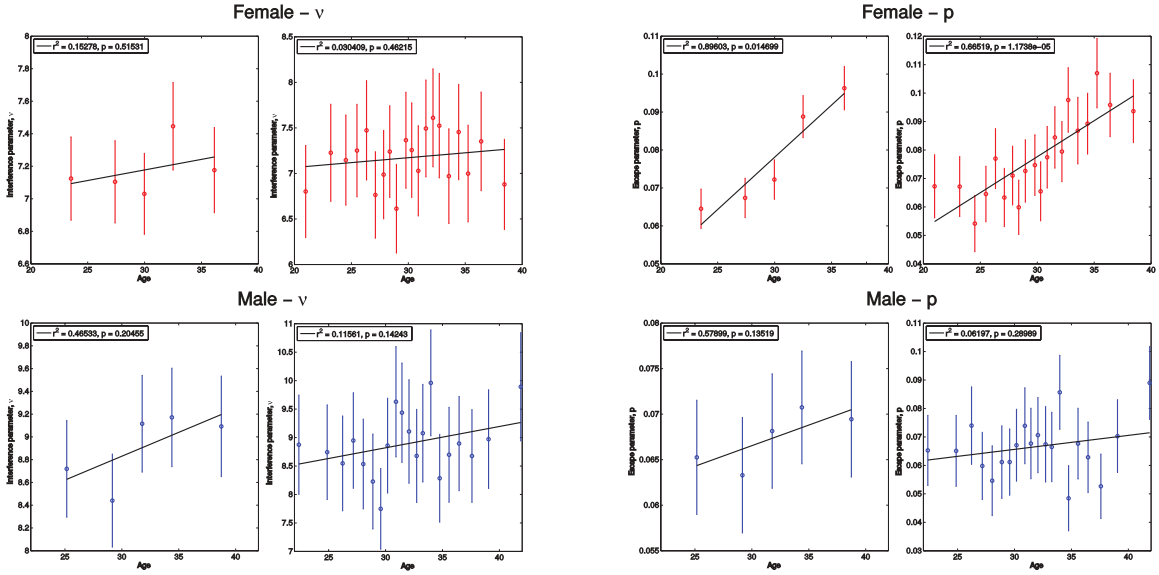

Figure 2.S10: **Interference parameters by age**, having divided the data in 5 or 20 age quantiles. Error bars show 95% confidence intervals.

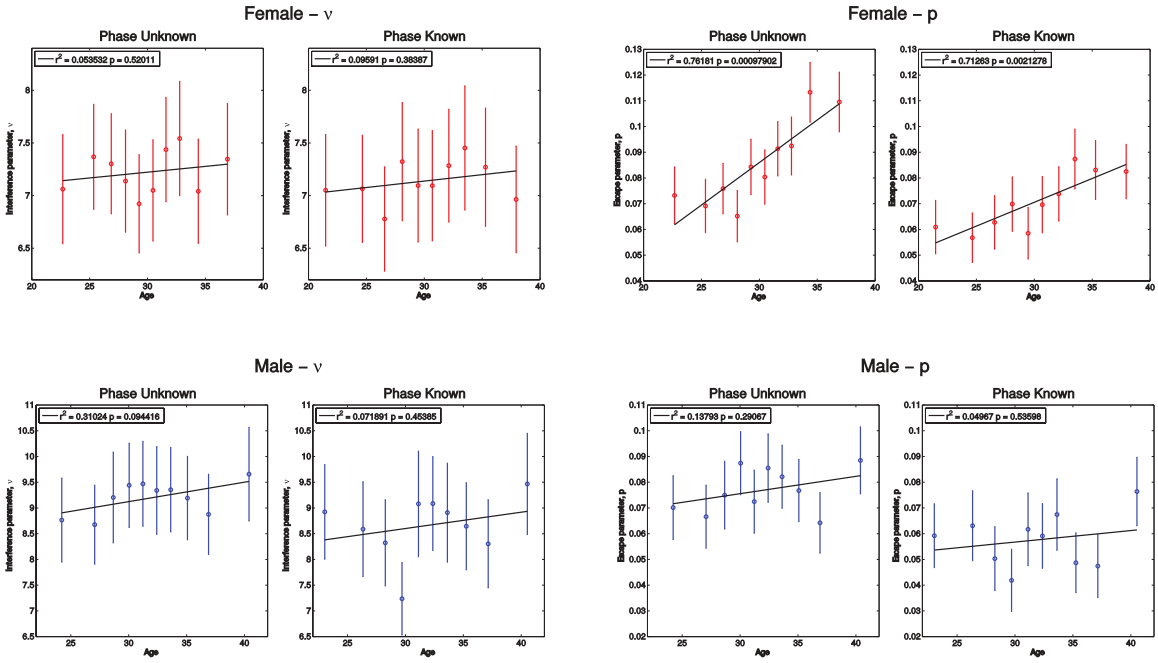

Figure 2.S11: **Interference parameters by age and phase.** Interference parameters by age, having estimated the interference parameters for phase-known and phase-unknown groups separately.

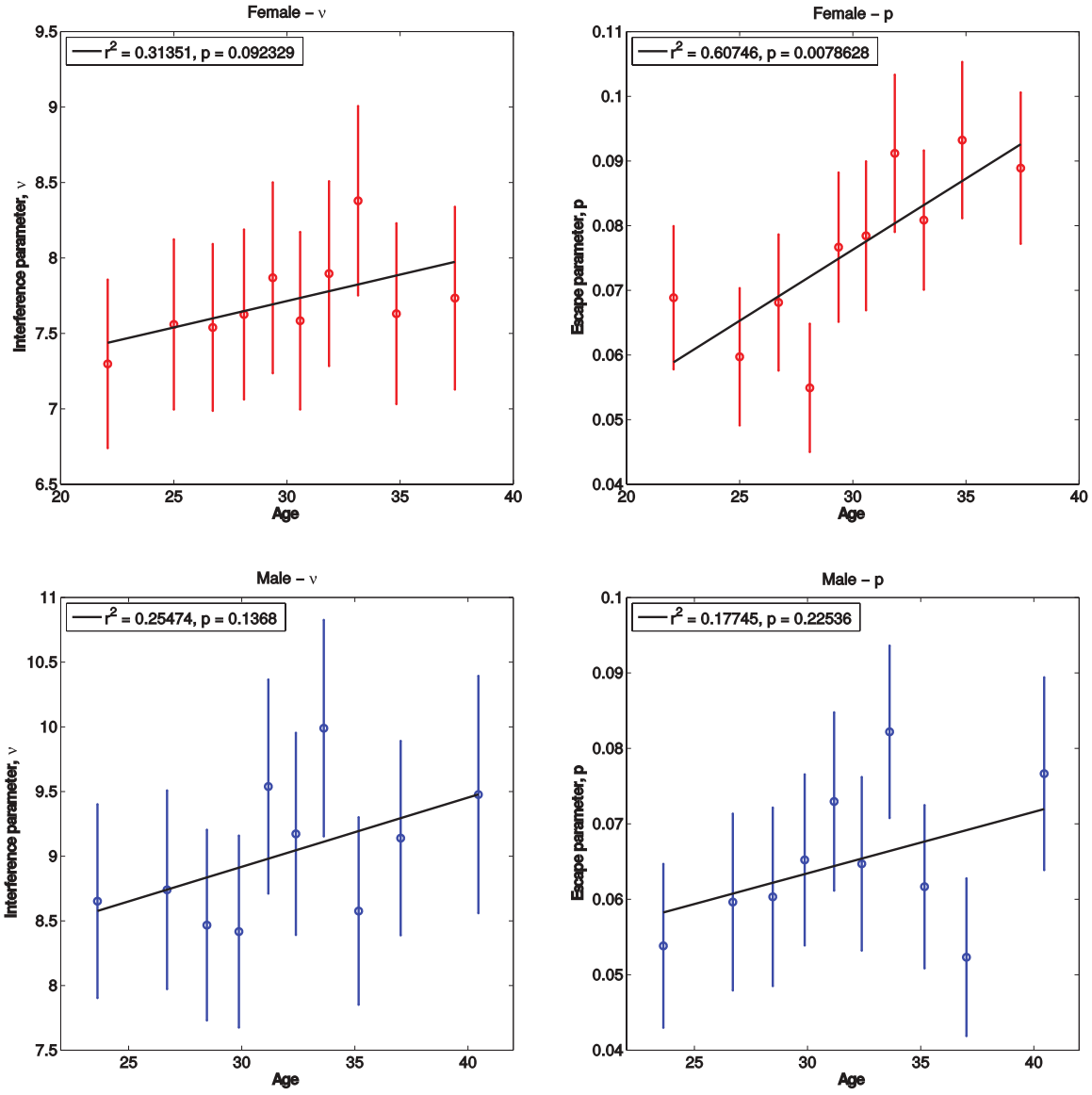

Figure 2.S12: **Interference parameters as a function of age, following stratified sampling.** Females and males are shown on the top and bottom rows respectively. Estimates of the interference parameter, ν, are shown on the left, whereas estimates of the escape parameter, p, are shown on the right. Error bars show 95% confidence intervals.

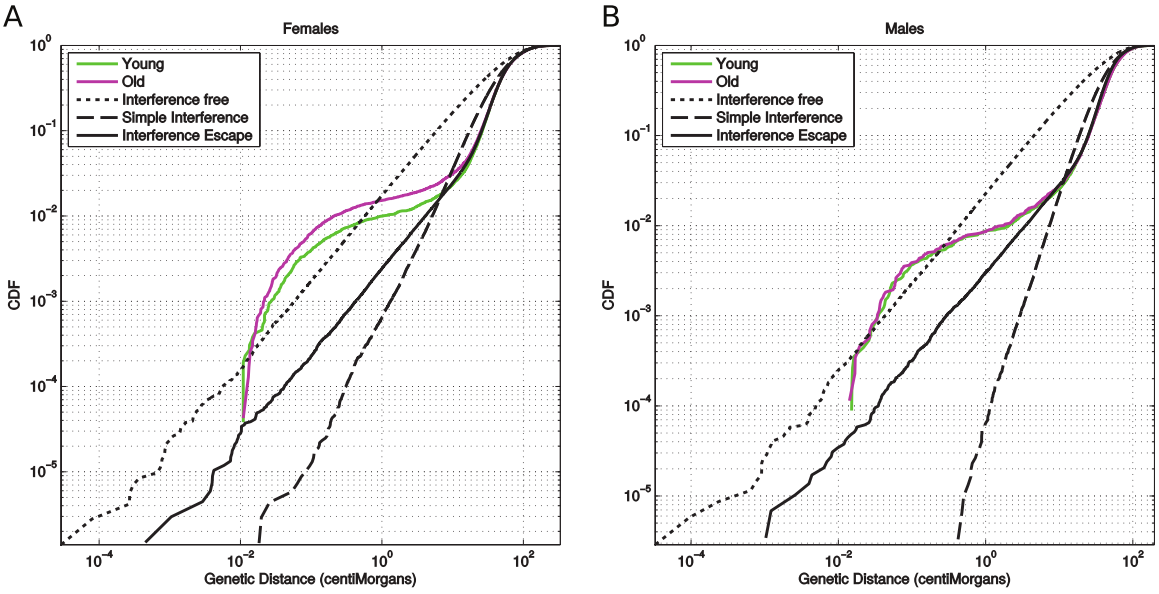

Figure 2.S13: **Model fit for tightly clustered events** in females (A) and males (B). The figure shows the empirical cumulative distribution function for young (green line) and old (magenta line) mothers/fathers, and compares to that obtained via simulation under the interference free model (black dotted line), the Gamma simple interference model (black dashed line), and the Housworth-Stahl interference escape model (solid black line), with parameters were taken from Supplementary Table 7. The figure is shown on a log-log scale to emphasize the short inter-crossover distances.

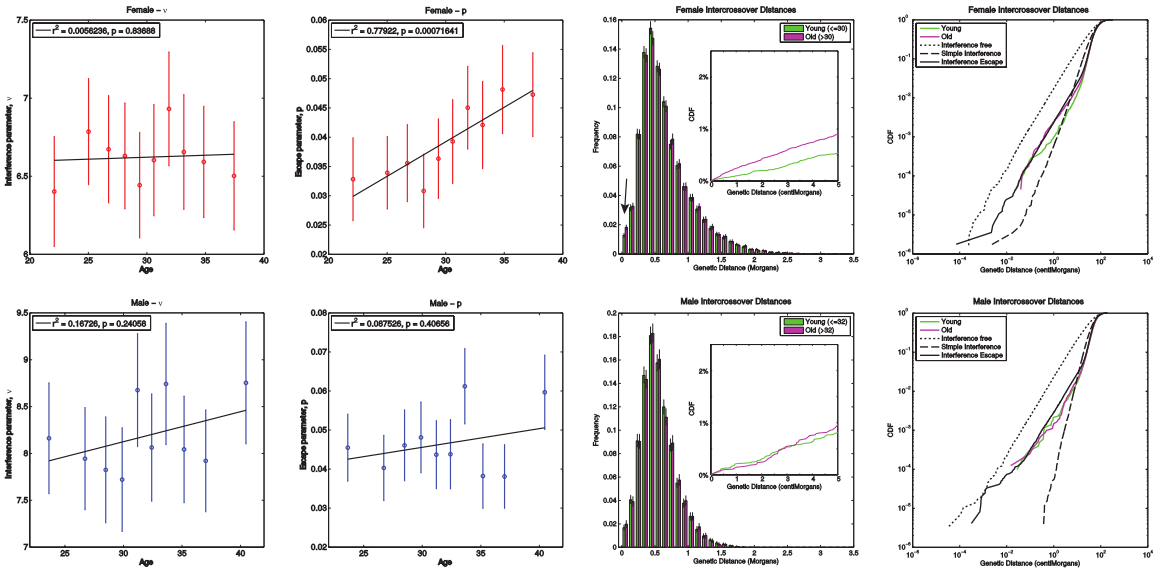

Figure 2.S14: **Interference parameters estimated for a strictly filtered dataset.** In this case, all crossover events were required at least 10 supporting informative sites (compared to 3 in the main dataset), no two events within a single family were allowed to be within 5 SNPs of each other (compared to 1 in the main dataset), and no more than 4 events within 1 Mb of each other were allowed across the whole dataset (and compared to 14 in the main dataset, which corresponds to the 99.9th percentile). After this very strict filtering, the deviation from the Housworth-Stahl interference escape model is much less pronounced at short scales (right hand panels), but the association between interference escape and maternal age remains strong (2nd panel from top left).

2.9 Supplementary Tables

Pedigree Type	Description	Before Filtering	After Filtering
1	2 parents, 2 children	3319	3307
2	2 parents, 3 children	560	523
3	2 parents, 4 children	89	80
4	Quartet, with 2nd generation trio	101	100
5	Trio, with 2nd generation quartet	201	199
	Total	**4270**	**4209**

Table 2.S1: **Summary of dataset, before and after filtering.**

Population	Female unphased	Male unphased	Female phased	Male phased	Total Meioses	Percentage
Europe	5382	5508	1789	1641	14320	78.24%
Latino	602	546	171	190	1509	8.25%
East Asia	380	308	88	74	850	4.64%
None/Other	198	268	68	109	643	3.51%
South Asia	178	176	19	20	393	2.15%
African American	152	152	34	36	374	2.04%
Middle East	76	100	15	22	213	1.16%
Total	**6968**	**7058**	**2184**	**2092**	**18302**	**100.00%**

Table 2.S2: **Description of parental ancestry for each meiosis within the sample.**

Chrom	First Position (bp)	Last Position (bp)	Physical Length (Mb)	Female Map Length (cM)	Female Mean Rate (cM/Mb)	Male Map Length (cM)	Male Mean Rate (cM/Mb)	SexAvg Map Length (cM)	SexAvg Mean Rate (cM/Mb)
chr1	1,031,540	249,170,711	248.14	335.9	1.36	198.30	0.80	267.05	1.08
chr2	118,913	242,763,542	242.64	316.45	1.31	184.64	0.76	250.52	1.03
chr3	152,592	197,759,785	197.61	270.98	1.37	163.85	0.83	217.4	1.1
chr4	167,596	190,787,660	190.62	260.11	1.37	145.79	0.76	202.93	1.06
chr5	184,702	180,673,228	180.49	249.13	1.38	146.66	0.81	197.87	1.1
chr6	188,937	170,777,087	170.59	236.64	1.39	140.88	0.83	188.74	1.11
chr7	67,365	159,042,351	158.97	223.17	1.41	136.04	0.86	179.55	1.13
chr8	200,898	146,235,564	146.03	210.94	1.45	122.41	0.84	166.64	1.14
chr9	215,269	141,004,945	140.79	195.69	1.4	125.54	0.89	160.58	1.14
chr10	162,102	135,402,200	135.24	207.86	1.54	129.91	0.96	168.86	1.25
chr11	244,552	134,872,342	134.63	193.59	1.44	120.21	0.89	156.88	1.17
chr12	216,039	133,684,321	133.47	200.36	1.51	131.20	0.98	165.75	1.24
chr13	19,458,371	114,998,076	95.54	152.26	1.6	101.19	1.06	126.71	1.33
chr14	20,445,905	107,233,999	86.79	137.22	1.59	97.29	1.12	117.24	1.35
chr15	22,763,396	102,381,360	79.62	143.39	1.8	100.85	1.27	122.11	1.53
chr16	143,503	90,102,384	89.96	157.29	1.75	102.03	1.13	129.64	1.44
chr17	84,782	81,025,393	80.94	152.87	1.9	106.23	1.31	129.53	1.6
chr18	218,695	77,955,378	77.74	140.06	1.81	97.80	1.26	118.91	1.53
chr19	288,246	59,058,083	58.77	117.8	2.01	99.42	1.69	108.59	1.85
chr20	100,699	62,892,739	62.79	118.9	1.9	99.00	1.58	108.93	1.73
chr21	14,807,136	47,978,421	33.17	74.34	2.24	51.76	1.58	63.04	1.9
chr22	17,152,611	51,165,664	34.01	78.16	2.31	63.30	1.86	70.71	2.08
chrX	2,737,282	154,408,041	151.67	179.02	1.18				
PAR1	178,624	2,689,575	2.51	2.73	1.16	42.94	17.17	22.75	9.06
PAR2	154,984,651	155,227,607	0.24	0.05	0.34	0.33	1.35	0.19	0.79
Genome			**2932.98**	**4354.91**	**1.48**	**2707.55**	**0.92**	**3441.11**	**1.17**

Table 2.S3: **Properties of the map estimated from 23andMe data.** Recombination fractions were converted to genetic map distances using the Haldane map function.

SNP	Chrom	Position	Alleles	P-value	Effect	95% CI	Gene Context
rs2001572	chr14	20,767,868	A/T	1.50E-08	0.503	[0.329,0.677]	[TTC5]
rs79621814	chr4	1,089,268	C/T	2.90E-08	-0.99	[-1.340,-0.640]	[RNF212]
rs11624006	chr14	91,961,188	C/T	2.80E-07	-0.478	[-0.660,-0.296]	[SMEK1]
rs72631326	chr17	65,769,087	C/T	4.40E-07	0.959	[0.587,1.331]	NOL11–[]–BPTF
rs11932663	chr4	184,458,083	A/G	5.10E-07	0.622	[0.380,0.865]	ING2–[]—RWDD4
rs17127442	chr8	18,779,787	C/T	5.10E-07	-0.537	[-0.746,-0.327]	[PSD3]
rs1879904	chr11	82,076,387	C/T	6.80E-07	-0.507	[-0.707,-0.307]	[]—FAM181B

Table 2.S4: **Variants associated with total number of recombination events.** Linear regression model tested as N_events $\sim$ sex + age + pc.0 + pc.1 + pc.2 + pc.3 + pc.4 + genotype. Association tests conducted using only individuals found to have $\geq$ 97% European ancestry.

SNP	Chrom	Position	Alleles	P-value	Effect	95% CI	Gene Context
rs73742307	chr5	23,534,421	C/T	7.90E-184	0.16	[0.149,0.170]	PRDM9-[]—CDH10
rs78474856	chr20	1,450,623	C/G	6.10E-07	-0.021	[-0.029,-0.013]	NSFL1C-[]-SIRPB2
rs62078596	chr17	53,906,496	C/T	8.50E-07	0.013	[0.008,0.018]	PCTP–[]—ANKFN1
rs8134126	chr21	28,401,705	C/T	1.00E-06	-0.01	[-0.013,-0.006]	ADAMTS5–[]
rs138108783	chr1	119,711,419	A/G	1.40E-06	0.274	[0.163,0.385]	WARS2–[]—HAO2

Table 2.S5: **Variants associated with hotspot usage.** Linear regression model tested as hotspot_usage $\sim$ sex + age + pc.0 + pc.1 + pc.2 + pc.3 + pc.4 + genotype. Association tests conducted using only individuals found to have $\geq$ 97% European ancestry.

Population	Female sample size*	Male sample size*	Female median hotspot usage	Male median hotspot usage	Difference	p-value (Mann-Whitney U)
Europe	3329	3325	62.96%	67.12%	4.16%	4.93E-40
Latino	362	341	61.15%	66.84%	5.68%	1.36E-09
East Asia	221	180	60.38%	67.56%	7.18%	5.67E-06
South Asia	97	95	61.65%	66.35%	4.71%	0.00494563
Middle East	88	88	59.52%	61.26%	1.74%	0.284789
African American	43	57	61.37%	65.37%	4.00%	0.135323
All	5668	5621	0.6268	0.67255	0.04575	1.06E-69

Table 2.S6: **Differences in hotspot usage between males and females**, partitioned by population. *The sample size represents the number estimated αs, with one estimate for each meiosis from phase-known parents, and a single estimate for phase-unknown parents.

Females									
	Gamma model (no escape)			**Escape model**					
	Phase known	Phase unknown	Weighted mean	Phase known		Phase unknown		Weighted mean	
Chrom	ν	ν	ν	ν	p	ν	p	ν	p
chr1	2.749	3.211	2.952	6.045	0.067	6.711	0.079	6.384	0.073
chr2	2.390	3.035	2.643	6.499	0.064	6.902	0.076	6.718	0.070
chr3	2.328	2.653	2.473	6.489	0.072	6.612	0.089	6.556	0.081
chr4	3.074	3.956	3.414	5.981	0.042	6.036	0.047	6.009	0.044
chr5	3.289	3.824	3.526	6.582	0.044	6.941	0.065	6.753	0.052
chr6	2.893	2.864	2.878	7.221	0.055	7.395	0.086	7.314	0.069
chr7	3.007	2.826	2.902	7.435	0.048	7.289	0.090	7.360	0.065
chr8	1.395	2.014	1.566	8.073	0.165	6.615	0.184	7.141	0.175
chr9	1.760	2.590	2.007	6.168	0.095	7.096	0.113	6.586	0.105
chr10	2.548	4.228	2.971	7.561	0.066	7.039	0.056	7.260	0.061
chr11	2.485	2.829	2.645	7.466	0.065	8.240	0.084	7.818	0.074
chr12	2.979	3.896	3.323	7.519	0.058	6.927	0.060	7.175	0.059
chr13	3.506	4.727	3.982	7.876	0.039	7.157	0.034	7.442	0.036
chr14	2.654	4.065	3.070	7.574	0.056	7.338	0.059	7.451	0.057
chr15	2.090	2.604	2.292	7.652	0.081	7.842	0.109	7.754	0.095
chr16	1.357	1.888	1.504	7.708	0.158	9.383	0.220	8.277	0.190
chr17	2.874	4.016	3.246	8.216	0.064	6.972	0.056	7.479	0.061
chr18	3.063	4.920	3.575	8.244	0.064	8.056	0.053	8.139	0.058
chr19	3.444	5.322	4.001	7.991	0.052	8.576	0.055	8.273	0.053
chr20	3.149	3.530	3.329	7.672	0.060	7.612	0.078	7.637	0.070
chr21	2.694	3.596	2.996	9.454	0.061	9.713	0.064	9.598	0.062
chr22	2.315	1.904	2.033	9.456	0.060	10.664	0.128	9.958	0.090
chrX	1.959	2.151	2.050	6.439	0.089	5.886	0.110	6.129	0.100
Autosomes	2.409	3.084	2.666	7.134	0.071	7.233	0.086	7.188	0.078
Males									
	Gamma model (no escape)			**Escape model**					
	Phase known	Phase unknown	Weighted mean	Phase known		Phase unknown		Weighted mean	
Chrom	ν	ν	ν	ν	p	ν	p	ν	p
chr1	3.240	3.289	3.266	8.515	0.047	9.419	0.082	8.949	0.063
chr2	4.081	3.972	4.019	7.567	0.038	8.439	0.063	8.024	0.050
chr3	3.640	4.381	3.977	9.123	0.045	8.376	0.053	8.695	0.049
chr4	4.469	4.256	4.343	8.516	0.046	9.217	0.072	8.895	0.059
chr5	4.425	5.232	4.795	7.593	0.030	7.847	0.047	7.737	0.038
chr6	3.255	3.388	3.324	9.828	0.055	9.199	0.077	9.456	0.066
chr7	3.266	5.311	3.873	8.297	0.057	8.991	0.055	8.685	0.056
chr8	2.197	1.816	1.946	10.760	0.119	9.216	0.173	9.775	0.145
chr9	2.137	3.642	2.490	9.253	0.108	9.845	0.096	9.587	0.101
chr10	4.323	4.823	4.564	8.575	0.047	9.556	0.071	9.031	0.058
chr11	3.693	4.879	4.160	7.422	0.055	8.794	0.058	8.158	0.057
chr12	3.228	4.430	3.666	8.269	0.060	8.025	0.063	8.126	0.061
chr13	5.706	4.058	4.467	8.387	0.029	10.051	0.058	9.142	0.042
chr14	4.647	5.348	4.969	9.479	0.028	9.083	0.042	9.295	0.033
chr15	2.579	3.596	2.932	8.127	0.065	9.244	0.064	8.652	0.064
chr16	3.485	2.641	2.875	7.675	0.064	8.492	0.105	8.114	0.088
chr17	3.278	2.092	2.339	8.735	0.063	9.582	0.125	9.220	0.095
chr18	4.587	3.191	3.538	8.380	0.050	8.278	0.066	8.314	0.058
chr19	3.808	4.607	4.156	7.423	0.061	8.975	0.074	8.104	0.068
chr20	3.184	3.478	3.333	8.205	0.079	9.601	0.084	8.905	0.082
chr21	2.485	5.772	2.841	100	0.074	100	0.049	100	0.057
chr22	2.467	3.414	2.786	10.442	0.059	16.799	0.074	12.670	0.069
Autosomes	3.346	3.591	3.470	8.608	0.058	9.184	0.077	8.931	0.067

Table 2.S7: **Interference parameter estimates for females (top) and males (bottom).** Estimates are given for phase-known and phase-unknown individuals separately. In addition, a combined estimate was calculated as a weighted average with weights taken to be the reciprocal of the variance.

Chrom	Start position (bp)	End position (bp)
1	144,954,851	145,394,955
1	145,547,963	146,508,934
1	146,997,245	147,093,887
1	147,162,445	147,205,770
1	147,210,993	147,222,372
1	147,375,981	147,782,284
8	6,881,638	8,119,716
8	11,088,131	11,096,553
8	11,251,705	11,256,184
8	11,330,364	11,332,026
8	11,354,933	11,359,638
8	11,363,950	11,372,141
8	11,406,175	11,476,726
8	11,486,220	11,496,193
8	11,501,265	11,503,333
8	11,514,144	11,516,373
8	11,533,384	11,570,036
8	11,722,125	11,755,513
8	11,763,932	11,799,654
8	11,830,877	11,846,482
8	11,857,317	12,559,475
10	46,076,235	47,597,927
10	47,611,631	48,324,245
10	48,368,273	48,380,952
10	48,400,458	48,427,246
10	48,440,744	48,471,020
10	48,489,541	48,508,137
10	48,512,114	48,545,527
10	50,122,109	50,163,975
10	50,382,038	50,382,478
10	50,451,843	50,471,176
10	50,568,814	50,585,177
10	50,615,087	50,615,806
10	50,623,895	50,643,498
10	50,821,243	50,824,244
10	50,824,619	51,559,469
10	135,160,950	135,195,332
10	135,202,594	135,257,091
10	135,347,727	135,349,367
10	135,351,362	135,352,100
12	8,000,912	8,021,932
15	22,876,889	22,908,392
15	22,909,207	22,918,657
15	22,932,511	23,053,839
16	21,327,273	21,620,270
19	2,098,015	2,099,820
19	54,077,870	54,106,839
19	54,107,686	54,111,568
22	17,729,044	17,731,977
22	25,650,406	25,848,811

Table 2.S8: **Locations of regions with high numbers of double recombination events.** Hg19 coordinates.

2.10 Supplementary Methods

2.10.1 Assessment of robustness to genotyping error

In order to understand how our results could be influenced by genotyping error we simulated data for each of the pedigree structures contained within our data. To do this, we generated haplotypes for the founder individuals using the coalescent simulation software ms[2]. Specifically, we generated 6 haplotypes (using: `ms 6 1 -t 2189.781`) and combined haplotypes at random to generate the genotypes of the founders. The population mutation rate was selected give an expected number of 5000 segregating sites. Children were then created by drawing haplotypes from each parent, and adding recombination as required.

To test MERLIN's ability to detect crossover events we placed one recombination event in the center of the sequence in one random parent, and passed this simulated pedigree data to MERLIN for haplotype analysis (option `--best`). This process is repeated to obtain 1000 total events per parent in each pedigree structure. Our results indicate that MERLIN is able to capture 99.6% of recombination events generated in this manner. The false negative calls resulted from low levels of heterozygosity (i.e. high relatedness) in the simulated haplotypes. The events placed in phase-known pedigrees were correctly assigned to the proper child in all cases. We repeated this simulation in the absence of any introduced recombination and find that in all cases, no events were called.

Estimates of the error rate of the Illumina HumanOmniExpress array used by 23andMe range from 0.01%[3] to 0.054%[4]. To test for robustness of our results to genotyping error, we next simulated pedigrees without recombination, but with a single genotyping error introduced into one of the individuals by switching one of the alleles at the middle site in the sequence. This procedure was repeated 1000 times in each of the five pedigree structures in our dataset. We looked for any events called by MERLIN and recorded the position in the sequence and the number of informative sites to the left and right of the event.

We estimated the number of false recombination events as a function of genotyping error. Without any filtering (and without using MERLIN's error detection functionality), we find MERLIN to be sensitive to genotyping error. For a dataset of our size and pedigree composition, a genotyping error rate of 0.001% would produce 15,000 false positive recombination events, rising to 150,000 for a 0.01%

genotyping error rate. However, the filters applied in the real dataset are effective at removing these simple false positives. After requiring at least 3 informative sites on both sides of a recombination event, we estimate that a dataset of our size would contain 74 spurious events with a 0.001% genotyping error rate, 739 with a 0.01% genotyping error rate, and 7,386 with a 0.1% genotyping error rate.

Although the assumptions of this simulation study are quite simplistic, given our dataset contains over 645,000 events these results would suggest that less than 1% of the events represent false positives. In addition, we note that in analysis of the real data, we used high-confidence sites and removed potential genotyping errors using MERLIN's error-detection feature (see Methods).

2.10.2 Individual Ancestral Assignment

Individuals were assigned to ancestral categories by quantifying the genetic variation they share with a set of representative reference populations.Chromosomal segments are assigned to geographic regions using 23andMe's Ancestry Composition tool[5]. Informally, Ancestry Composition assigns regions of an individual's genome to 31 reference populations constructed from public reference datasets as well as private 23andMe cohort data[6]. Individuals are assigned to genomic regions by first splitting the genome into short non-overlapping segments, and assigning each segment to the reference population with the highest degree of similarity. Given this assignment, it is straightforward to compute the percentage of an individual's DNA that originates from a certain sub-population. For example, if 200,000 out of 400,000 total segments are predicted to come from an African background, then the global percentage of African ancestry is 50%. Given this global percentage, individuals are assigned to high-level categories (European, Middle Eastern, East Asian or South Asian) if their total percentage of ancestry in that category exceeds 97%. For individuals of admixed ancestry, 23andMe uses a logistic classifier trained on the segment length distributions of individuals who have self-identified as African American or Latino. In order to define the final population label for a given individual, we first determined if they had at least 97% European, Middle Eastern, East Asian or South Asian ancestry. If so, then their category was determined. If the 97% threshold was not met, but the individual had a total global percentage of at least 97% when summing contributions from European, African and Native American, then the logistic classifier was applied. If neither of these conditions were met, then the individual was categorized as "Other".

2.10.3 Estimation of hotspot usage

To estimate the degree of hotspot usage by an individual, we adopted the method of Coop et al.[7]. In brief, this method estimates the fraction of recombination events that overlap with known LD-based hotspots while accounting for the uncertainty in the localization of the called recombination events. For convenience, we re-describe the approach here.

We aim to estimate the proportion, α, of events that occur within LD-based hotspots. Given a recombination event, r, the probability that the event overlaps with a hotspot is given by:

$$P(r \text{ overlaps a hotspot}) = \alpha + (1-\alpha)P(r \text{ overlaps a hotspot by chance})$$

To estimate $P(r \text{ overlaps a hotspot by chance})$, we randomly shift the recombination events by a normally distributed distance (mean 0, standard deviation 200kb) a total of 1,000 times, and calculated the fraction of these moves that result in the event overlapping a hotspot. The likelihood for α is given by:

$$L(\alpha|r) = \delta_r P(r \text{ overlaps a hotspot}) + (1-\delta_r)(1 - P(r \text{ overlaps a hotspot}))$$

where δ_r is an indicator function, taking the value 1 if r overlaps a hotspot and zero otherwise. For a set of k recombination events labeled $r_0, r_1, \ldots, r_{k-1}$, the likelihood of α for the whole dataset is given by:

$$L(\alpha|r_0, r_1, \ldots, r_{k-1} = \prod_{i=0}^{k-1} L(\alpha|r_i). \tag{2.1}$$

We used this method to estimate α for each mother and father (for phase unknown individuals), and each meiosis (for phase known individuals). As in Coop *et al.*[7], we used all events that were well localized to within 30kb, but note that our results are robust to larger values of this parameter. The likelihood of alpha was estimated over a uniformly spaced grid of 2,000 values between 0 and 1, with the MLE taken as the value of α with the maximum likelihood on this grid. A 95% confidence interval was constructed as being the set of values within two log likelihood units of the MLE.

For phase-known individuals for which recombination events could be assigned to specific children, a separate α was estimated for each meiosis. For phase-unknown individuals where such an assignment was not possible, α was estimated using all events that could be attributed to the parent.

Hotspot usage results

The estimates for hotspot usage are shown in Supplementary Figure 2.S7. The median hotspot usage estimate for females was 62.68% (95% C.I. 62.25% - 63.10%), whereas for males it was 67.26% (95% C.I. 66.85% - 67.69%), a difference of 4.6% ($p = 1.1 \times 10^{-69}$, Mann-Whitney U).

To ensure the difference between males and females is not driven by higher precision in females (resulting from higher numbers of events), we thinned the female data in order to match the number of events in males. Specifically, for each male, we randomly selected a female (without replacement) with a greater or equal number of events, and thinned the female events to match the number of male events. The resulting dataset contains an equal number of males and females, with each pair having an equal number of events. The estimates of hotspot usage for the two sexes were very similar to the previous estimates (62.2% for females, and 66.8% for males), and the difference in hotspot usage remains highly significant ($p < 2.2 \times 10^{-16}$).

To determine whether the observed differences in hotspot usage between males and females is dependent on the position within the chromosome (as males tend to have higher recombination rates towards the telomeres), we repeated the analysis having divided each chromosome into segments. Specifically, we split each chromosome into three windows, assigning the terminal 25% of sequence from each end to p- and q-arm bins, and keeping the central 50% of the sequence for the middle bin. For acrocentric chromosomes we omit the p-arm bin. We estimated the degree of hotspot usage in each of these bins. We observe that males use hotspots to a greater extent than females (Mann-Whitney U $p < 2.2 \times 10^{-16}$ for all three bins), suggesting that the difference in hotspot usage between males and females cannot be explained by telomere effects.

Due to variation in PRDM9, hotspot usage is expected to vary between populations[8,9]. The hotspots used in this study were identified from genome-wide Phase II HapMap linkage disequilibrium data[10], in which hotspots were called that were active in at least two of the three constituent populations (CEU, YRI, JPT+CHB). As such, one possibility for the observed difference between males and females is that the ancestry proportions within our data differ between the female and male samples. Inspection of the ancestry proportions within our data showed this not to be the case. In addition, if the analysis is partitioned by inferred ancestry, females have lower hotspot usage within all populations (Figure 2.2B), with the difference remaining significant in European, East Asian, Latino, and South

Asian populations(Supplementary Table 2.S6).

2.10.4 Description of age effect

Previous research has indicated a relationship between maternal age and the number of recombination events. In particular, research from the deCODE consortium used data from 14,140 meioses to report that the number of recombination events in females increase with age[11]. The reported effect size is reasonably modest, contributing 0.082 ($\pm$ 0.012 standard error) recombination events per year, depending on the analysis method used. This translates as approximately a 4% increase in the average maternal recombination rate over a period of 25 years. No such association was observed in males. A second study confirmed this effect using 728 meioses observed with from Hutterite families[7], observing that mothers over 35 years of age had approximately 3.1 extra recombination events compared to those under 25. Despite the small sample size, the effect size in this study was estimated to be 0.19 ($\pm$ 0.092 standard error) events per year. Again, no such effect was observed in males.

Conversely, a separate research group considering recombination events in 195 meioses reported a decrease in the number of recombination events with maternal age[12]. In this case, the effect size was larger, corresponding to between -0.49 and -0.42 crossovers per year, again with no such effect observed in males. Although the smallest of the three studies, the authors suggest that the discrepancy in the direction of the effect between studies could be due to marker density and/or true biological differences between populations.

Correlation between number of recombination events and parental age

To quantify the correlation between parental age and recombination rate, we first partitioned our data into phase-unknown parents for which recombination events could not be assigned to a specific child (or meiosis), and phase-known parents for which such an assignment was possible. For the phase-unknown parents group we used the maternal / paternal ages averaged across children, whereas for the phase-known group, we used the known parental ages at the time of the child's birth.

Using linear regression, we estimated the association between the number of autosomal events and parent age (Supplementary Figure 2.S6). A weak positive association between age and the number of recombination events was detected for females, but no such effect was observed for males.

The number of recombination events in females increased on average by 0.067 per year (standard error: ± 0.0215), which is similar to the estimate from deCODE.

We note that the observed effect is quite weak, and appears to be largely driven by an increase in the number of recombination events for mothers of 35 years or older (Figure 2.1C).

To ensure the observed effect is not confounded by population structure within the data, we first repeated the analysis for each population separately. In Europeans, for whom we have by far the largest sample size (accounting for ∼76%of individuals), a significant association with maternal age was still observed (0.087 extra events per year, $p = 3.2 \times 10^{-4}$). In all other populations (East Asian, Middle Eastern, Latino, African American, and South Asian), no significant association was observed, possibly due to insufficient power. No significant association with paternal age was observed within any population.

2.10.5 Inferring Crossover Interference

In the following text, we provide a description of the crossover interference models used within the main analysis.

The Gamma Model (a.k.a. the 'simple interference' model)

We follow the description of the Gamma model of crossover interference presented by Broman and Weber[13]. For clarity, we repeat the description of this the model below.

The Gamma model describes the locations of chiasmata on the four-strand bundle according to a stationary renewal process, with increments being drawn from a gamma distribution with shape ν and rate 2ν. As such, in this model the distances between chiasmata are independent with mean 0.5 Morgans, and a standard deviation of $(2\sqrt{\nu})^{-1}$. Under the assumption of no chromatid interference, the chiasmata are thinned such that each chiasmata becomes a crossover with probability 0.5. As such, this model satisfies the requirement that the average inter-crossover distance should be 1 Morgan.

The parameter ν is a unitless measure of the strength of interference. Specifically, ν = 1 corresponds to no interference between chiasmata, and $\nu > 1$ corresponds to positive interference (i.e. decreased variance in chiasmata spacing than would be expected under a Poisson model), and $\nu <$

1 corresponds to negative interference (i.e. increased variance in chiasmata spacing than expected under a Poisson model).

Let $x_0, x_1, x_2, \ldots$ be the genetic distances (in Morgans) between adjacent chiasmata, with x_0 being the distance from the p-terminal end of the chromosome to the first chiasma. Under the Gamma model, the chiasmata locations are generated according to a gamma renewal process, such that $x_1, x_2, \ldots$ are independent and follow a gamma distribution with shape ν and rate 2ν, where ν is a positive real number. Therefore, the density of x_i is given by $f(x;\nu) = (2\nu)^\nu e^{-2\nu x} x^{\nu-1}/\Gamma(\nu)$, for $i > 0$, and where $\Gamma(.)$ represents the gamma function. The density of x_0 is given by $g(x;\nu) = 2[1 - F(x;\nu)]$, where F is the cumulative distribution function (cdf) of f.

However, using transmitted genotype data, the actual chiasmata locations are not observed. Rather, only the crossovers derived from the chiasmata positions are observed. Assuming no chromatid interference, the probability that a chiasmata results in a crossover is $\frac{1}{2}$.

Let $y_0, y_1, y_2, \ldots$ be the genetic distances (in Morgans) between adjacent crossovers. Each y_i is independent, with density given by $f^*(y;\nu) = \sum_{k=1}^{\infty} (\frac{1}{2})^k f_k(y;\nu)$, where is the gamma distribution density with shape $k\nu$ and rate 2ν: $f_k(k;\nu) = (2\nu)^{k\nu} e^{-2\nu x} x^{k\nu-1}/\Gamma(k\nu)$, which is derived from the convolution of $f(y;\nu)$ with itself k times. The density of y_0 is given by $g^*(k;\nu) = 1 - F^*(y;\nu)$, where F^* is the cdf of f^*. Likewise, let G^* represent the cdf of g^*.

Given the above model, the contribution to the likelihood is:

$$Lk(\nu; y) = \begin{cases} 1 - G^*(L;\nu) & \text{if } m_i = 0 \\ g^*(y_0;\nu) g^*(y_1;\nu) & \text{if } m_i = 1 \\ g^*(y_0;\nu) \left[\prod_{j=1}^{m_i - 1} f^*(y_j;\nu) \right] g^*(y_m;\nu) & \text{otherwise} \end{cases} \tag{2.2}$$

The likelihood for the complete data may be obtained as the product over all individual contributions.

The Housworth-Stahl 'interference escape' model.

The Gamma model assumes that all crossover events are subject to the same interference process. The model has been shown to fit the data reasonably well for numerous organisms[13,14]. However, evidence from model organisms suggests the existence of a subset of events that are not subject to crossover interference[15], and statistical support of this finding has been seen in humans[16,17].

For this reason, we adopt the Housworth-Stahl model of interference, which models the distances between crossovers as being a mixture of two processes. In one process, crossovers are distributed according to the gamma model described above, whereas in the second process, crossovers are distributed without interference. We describe this model here, following Housworth and Stahl's 2003 paper[17], and refer to it as the 'interference escape' model.

Assume that we have a mixture of two independent types of crossover, such that one type occurs with probability q and has interference parameter ν, and the other type occurs with probability $p = 1 - q$ and is not subject to interference ($\nu = 1$). As for the Gamma model described above, let $x_0, x_1, x_2, \ldots$ be the genetic distances (in Morgans) between adjacent chiasmata, with x_0 being the distance from the p-terminal end of the chromosome to the first chiasma. The distances between chiasmata are given by a gamma distribution with shape ν and rate $2q\nu$. As such, the density of x_i is given by $f(x;\nu,2q\nu) = (2q\nu)^\nu e^{-2q\nu x} x^{\nu-1}/\Gamma(\nu)$, for $i > 0$. Likewise, the density of x_0 is given by $g(x;\nu,q) = 2q[1 - F(x;\nu,2q\nu)]$, where F is the cumulative distribution function (cdf) of f.

As described for the Gamma model, crossover events are determined by thinning the chiasmata positions, with each position retained with probability $\frac{1}{2}$. Let $y_0, y_1, y_2, \ldots$ be the genetic distances (in Morgans) between adjacent crossovers of this type. Each y_i is independent, with density given by $f^*(y;\nu,q) = \sum_{k=1}^{\infty} (\frac{1}{2})^k f(y;k\nu,2q\nu)$. The density of y_0 is given by $g^*(k;\nu,q) = q[1 - F^*(y;\nu,q)]$, where F^* is the cdf of f^*. Likewise, let G^* represent the cdf of g^*.

Now consider a dataset from a single meiosis where the intercrossover distances are given by $x_0, x_1, x_2, \ldots, x_n$, where $\sum_{i=0}^{n} x_i = L$. We assume these events are derived from two types of crossover. The interference-free type occurs with probability p and has $\nu = 1$. The second type is subject to interference and occurs with probability $q = 1 - p$. To calculate the likelihood of the data, we must sum over the 2^n possible ways to assign crossovers to the two types. Given one possible assignment, we split the data into two sets of intercrossover distances, $y_0, y_1, y_2, \ldots, y_j$ for the interference-free type, and $z_0, z_1, z_2, \ldots, z_k$ for the second 'interference' type, where $j + k = n + 1$.

The likelihood of the data in from the interference-free type is:

$$Lk(\nu = 1, q = p; y) = \begin{cases} 1 - G^*(L; 1, p) & \text{if } j = 0 \\ g^*(y_0; 1, p)[1 - F^*(y_1|1, p)] & \text{if } j = 1 \\ g^*(y_0; 1, p)\left[\prod_{i=1}^{j-1} f^*(y_i; 1, p)\right]\left[1 - F^*(y_j|1, p)\right] & \text{otherwise.} \end{cases} \quad (2.3)$$

The likelihood of the data from the interference type is:

$$Lk(\nu = t, q = 1 - p; z) = \begin{cases} 1 - G^*(L; t, 1 - p) & \text{if } j = 0 \\ g^*(y_0; t, 1 - p)[1 - F^*(y_1|t, 1 - p)] & \text{if } j = 1 \\ g^*(y_0; t, 1 - p)\left[\prod_{i=1}^{j-1} f^*(y_i; t, 1 - p)\right]\left[1 - F^*(y_j|t, 1 - p)\right] & \text{otherwise.} \end{cases} \quad (2.4)$$

To calculate the likelihood of the data, we sum over all 2^n possible assignments to the two types:

$$Lk'(\nu = t, q = p; x) = \sum_{\substack{(y_0, y_1, y_2, \ldots, y_j), \\ (z_0, z_1, z_2, \ldots, z_k)}} Lk(\nu = 1, q = p; y) Lk(\nu = t, q = 1 - p; z) \quad (2.5)$$

To calculate the likelihood over multiple individuals, one simply takes the product of the above likelihood.

In our implementation of the above formulas, we calculated f^* by summing over k from 0 to 25. Numerical integration was used to calculate G^* using the *integral* function in MATLAB.

Extension to interference escape model for phase-unknown data

The above description of the interference escape model assumes that the observed crossover events can be assigned to a specific meiosis. However, in the case of the phase-unknown individuals that make up the majority of our data, the observed crossovers cannot be assigned to specific children. As such, the above model cannot be used.

To extend the model for phase-unknown, we perform the same trick of summing over all possible assignments to each type, but this time also summing over all possible assignments to each meiosis. Although this procedure is somewhat naïve, both simulations and comparison of results between phased and unphased families have shown that it works well in practice (Supplementary Figure 2.S11).

Consider a family quartet. For each parent, the observed crossovers are the result of two independent meioses, which we will call M_1 and M_2 respectively. Let the intercrossover distances events in M_1 be $a_{i0}, a_{i1}, a_{i2}, \ldots$, and the intercrossover distances in M_2 be $b_{i0}, b_{i1}, b_{i2}, \ldots$, where a_{i0} and b_{i0} represent the distances between the first event and the p-terminal end of the chromosome in M_1 and M_2 respectively. If we could observe these intercrossover distances, we could apply the Housworth-Stahl model as described above. However, due to the nature of phase-unknown individuals, all we are unable to directly observe these distances, and can only observe crossovers derived from both meioses without knowing which event is from which meiosis.

Naïvely, we could be to sum over all possible assignments to each meiosis,and for each assignment apply the Housworth-Stahl model independently. However, this would be inefficient, as it would result in summing over 4^n possible assignments (as there are 2 crossover types in each of 2 meioses). Instead, we note that the same result can be achieved by combining the 'interference free' classes, allowing us to sum over 3^n possible assignments.

Let the n observed crossover positions assigned to a parent be $z_{i1}, z_{i2}, z_{i3}, \ldots, z_{in}$, which are derived from a superposition of the gamma renewal processes. In order to calculate the likelihood of this data, we treat the assignment of each event as either belonging to one of two inference classes, or to a single interference free class. Specifically, we calculate the likelihood as:

$$Lk'(\nu, q; z) = \sum_{\substack{(a_0, a_1, \ldots, a_i), \\ (b_0, b_1, \ldots, b_i), \\ (c_0, c_1, \ldots, c_i)}} \Big[Lk(\nu = t, q = 1 - p; a) Lk(\nu = t, q = 1 - p; b) Lk(\nu = 1, q = 2p; b) \Big] \quad (2.6)$$

where the summation is taken over all possible 3^n divisions of the n crossovers into the three classes. The likelihood for the complete dataset is given by taking the product of $Lk'(\nu, q; z)$ over all individuals.

Maximum likelihood estimation of ν and q was performed using a MATLAB implementation of the Nelder-Mead method[18], restricting the search space such that $\nu \in [0.1, 100]$, and $q \in (0, 0.5)$. Uncertainty in the MLE point estimates was obtained by using the inverse of the Fisher information matrix to estimate the covariance matrix.

We note that the mixture model lacks identifiability when ν is close to 1. In this situation, the estimates of q become uninformative. When performing likelihood maximization, we experimented with including a weakly informative prior on q that favors smaller values. Specifically, we set $P(q) =$

$1-q$, and performing maximum a posteriori estimation in place of maximum likelihood. In simulations, we found this method slightly improved results when ν is small, and has negligible effect otherwise. However, given the limited benefit of this approach,we did not pursue it further.

We validated the extension using simulations, and found it to give comparable results to those obtained from the original version for phase-known data. In addition, the per-chromosome estimates obtained from the real data were largely concordant between the phase-known and phase-unknown estimates (Supplementary Table 2.S7).

MATLAB code for performing inference of crossover interference parameters using this extension can be found at `https://github.com/auton1/interference/`.

Interference across the genome

We fitted the Gamma and interference-escape models for each chromosome separately, and also having combined data across the autosomes. As reported previously[16,17], we find the interference escape model to provide a much better fit to the data than the traditional Gamma model (Figure 2.3A), and therefore focus on parameter estimates from this model.

Across the whole genome, crossover interference is stronger in males than for females. The average interference parameter was estimated to be $\nu = 7.18$ in females, and $\nu = 8.93$ in males, which implies increased variance in crossover spacing for females relative to males. We infer that $p = 7.8\%$ and $p = 6.7\%$ of events escape interference in males and females respectively. We note that these estimates are quite similar to those obtained in Hutterites[16], where the estimates were reported as $\nu = 9.17$, $p = 8\%$ and $\nu = 6.96$, $p = 6\%$ in males and females respectively.

The results for each chromosome are shown in Figure 2.3B and C. In females, there is a clear trend of shorter chromosomes having higher interference parameter (ν) estimates, whereas any such effect is much weaker in males. In contrast, no such relationship is seen in the fraction of events that escape interference (p).

Of note in males, the estimate of the interference parameter for chromosome 21 appears to be extremely large, if not infinite (Supplementary Table 2.S7). This finding has been reported previously[13,16], and reflects the fact that very few paternal chromosomes exhibit more than one crossover. In our data, just 1.7% of paternal meioses have evidence of more than one crossover on chromosome 21, compared to 30.0% for chromosome 20 and 8.3% for chromosome 22.

The degree of interference on a chromosome is reasonably well predicted by the map length. Combining data across the sexes, the chromosome map length explains 57% of the variance in the interference parameter (Supplementary Figure 2.S8). When considering the sexes separately, the association is stronger in females (where 69% of the variance can be explained) than in males (where just 17.2% can be explained, and the fit does not achieve significance; $p = 0.061$). A multiple regression including sex as a predictor variable ($\nu_{chr} = \beta_0 + \beta_1 maplength_{chr} + \beta_2 sex_{chr}$) finds the β_2 to be marginally significant ($p = 0.0183$), but the model is not a significantly better fit than the model without including sex (Δ(deviance)= 3.54, $p = 0.0599$).

Analysis of interference by age

We divided our data into quantiles of approximate equal size on the basis of age. For each decile, we fitted the interference-escape model. The results are shown in Figure 2.4 and Supplementary Figure 2.S9 for 10 quantiles, and in Supplementary Figure 2.S10 for 5 and 20 quantiles. In females, the proportion of events escaping interference consistently increases with maternal age, and the pattern is consistent across both phase-known and phase-unknown individuals (Supplementary Figure 2.S11). There is no such correlation in the degree of interference, which appears to be constant across maternal ages. In contrast, no correlation is observed between paternal age and either parameter.

Stratified sampling to account for number of crossovers

One potential concern is that the inferred degree of interference may be influenced by a change in recombination rate. As the distribution of distances between crossovers depends on the number of crossovers (when there are more crossovers, they are necessarily more closely spaced), if there is a change in the recombination rate with age then this may influence the interference estimates.

We can address this concern by the use of stratified sampling. Specifically, for each age group, we subsampled individuals in order to ensure that each decile has the exact same distribution of the number of crossovers per meiosis. This was achieved as follows. First, for each age group i, we counted the number of individuals with x crossovers, which we call $N_i(x)$. For each x, we estimated the minimum $N_i(x)$ across all decile age groups, so that $n(x) = min_i(N_i(x))$. We then subsampled individuals within each decile by randomly selecting $n(x)$ individuals, without replacement, for each possible x.

Having performed this subsampling, we repeated the analysis. The results are shown in Supplementary Figure 2.S12. The results for females are largely identical to that obtained without stratified sampling, with a significant increase in the proportion of events escaping interference as maternal age increases.

2.11 Data Availability

Sex-specific genetic maps generated from this data are available at `http://autonlab.einstein.yu.edu/23andMe_recomb/`. To preserve the privacy of participants, access to other data associated with this study is controlled through the 23andMe Research Portal[19].

2.12 Supplementary References

1. Duffy, D. L. An integrated genetic map for linkage analysis. *Behavior Genetics* 36(1):4–6 (2006). doi:10.1007/s10519-005-9015-x.

2. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–8 (2002). doi:10.1093/bioinformatics/18.2.337.

3. Illumina. HumanOmniExpress datasheet. *http://www.illumina.com/Documents/products/datasheets/datasheet_human_omni_express.pdf* (2013).

4. Imai, K., Kricka, L. J., and Fortina, P. Concordance Study of 3 Direct-to-Consumer Genetic-Testing Services. *Clinical Chemistry* 57(3):518–521 (2010). doi:10.1373/clinchem.2010.158220.

5. 23andMe. Ancestry Composition: 23andMe's State-of-the-Art Geographic Ancestry Analysis. *http://www.23andme.com/ancestry_composition_guide/* (2014).

6. 23andMe. Reference Populations in Ancestry Composition. *http://customercare.23andme.com/entries/22584878-Reference-populations-in-Ancestry-Composition* (2012).

7. Coop, G., Wen, X., Ober, C., Pritchard, J. K., and Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319(5868):1395–8 (2008). doi:10.1126/science.1151851.

8. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., *et al.* The landscape of recombination in African Americans. *Nature* 476(7359):170–5 (2011). doi:10.1038/nature10336.

9. Berg, I. L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N. J., *et al.* 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):12378–83 (2011). doi:10.1073/pnas.1109531108.

10. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61 (2007). doi:10.1038/nature06258.

11. Kong, A., Barnard, J., Gudbjartsson, D. F., Thorleifsson, G., Jonsdottir, G., *et al.* Recombination rate and reproductive success in humans. *Nature Genetics* 36(11):1203–6 (2004). doi:10.1038/ng1445.

12. Hussin, J., Roy-Gagnon, M.-H., Gendron, R., Andelfinger, G., and Awadalla, P. Age-dependent recombination rates in human pedigrees. *PLoS Genetics* 7(9):e1002251 (2011). doi:10.1371/journal.pgen.1002251.

13. Broman, K. W. and Weber, J. L. Characterization of human crossover interference. *American Journal of Human Genetics* 66(6):1911–26 (2000). doi:10.1086/302923.

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

15. Baudat, F. and de Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Research* 15(5):565–77 (2007). doi:10.1007/s10577-007-1140-3.

16. Fledel-Alon, A., Wilson, D. J., Broman, K., Wen, X., Ober, C., *et al.* Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genetics* 5(9):e1000658 (2009). doi:10.1371/journal.pgen.1000658.

17. Housworth, E. A. and Stahl, F. W. Crossover interference in humans. *American Journal of Human Genetics* 73(1):188–97 (2003). doi:10.1086/376610.

18. D'Errico, J. fminsearchbnd, fminsearchcon - File Exchange - MATLAB Central. *http://www.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd–fminsearchcon* (2005).

19. 23andMe. 23andMe Research Portal. *http://www.23andme.com/researchportal/* (2013).

Chapter 3

Crossover interference varies by age and individual

Christopher L. Campbell[1] and Adam Auton[1*]

This chapter contains unpublished data.

[1] Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, New York 10461, USA.
[*] Former affiliation.

3.1 Introduction

Crossing over during meiotic recombination is not a random process but a carefully orchestrated sequence of events, with many factors controlling the placement, frequency, and spacing of recombination. While the crossover properties appear to be under evolutionary constraints, few (if any) of the features of crossover appear to be fixed, with many crossover properties having been shown to vary widely between individuals, sexes and populations.

One well characterized example of variation is that of recombination hotspot usage. Crossovers have been shown to cluster into hotspots of recombination[1,2], and the placement of these events has been shown to be under the control of the PRDM9 protein[3–5]. However, when looking at the location of crossovers derived from specific meioses, the overlap with known hotspot locations varies widely[6]. The degree to which a set of crossovers overlaps with a set of known hotspots is largely dependent on which particular PRDM9 allele an individual carries. As over 35 distinct PRDM9 human alleles have been identified to date, the degree of hotspots usage can vary widely between individuals and populations[3,7].

Many of the current methods for studying meiotic recombination are based on pedigree studies, which infer the locations of crossover events through indirect measurements from transmitted meiotic products. While these studies provide high-quality genome-wide data on recombination, they are limited by a requirement for large sample sizes. Furthermore, while such studies do provide data on a per-individual basis, the crossover products from a single meiosis generally number no more than 20-60[6,8]. As such, much of the value from pedigree studies is derived from combining data over a number of distinct meioses. Even in the case where multiple children are born within the same family, the number of crossovers that can be obtained per individual parent is perhaps a few hundred at most.

These limitation have left unanswered questions regarding recombination how varies within a single individual. To address such questions, researchers have generally relied on sperm typing methods, which are powerful but labor intensive and restricted to targeted regions of the genome[9]. A recent series of analyses has used single cell sequencing approaches to identify crossover events in sperm[10,11] and oocytes[12], and were therefore able to observe multiple recombination products from single individuals. These data have the potential to reveal information on recombination variance on

an individual level that cannot be achieved through human pedigree studies.

Nonetheless, it is clear that recombination is a dynamic process, even over the lifetime of a single individual. Specifically, there are a number of outstanding questions regarding how the properties of recombination vary as an organism ages. In humans, a number of studies have provided evidence that the crossover frequency increases with age in females[13,14], but there are conflicting studies that report the opposite[15,16].

In addition, in a recent study, presented in Chapter 2 of this thesis, we reported on an age-based effect relating to crossover interference. Crossover interference governs the spacing of events between crossovers within the same meiosis, and acts to space events further apart than would be expected by chance. Crossover interference appears to be well explained by a model that assumes two classes of crossovers[17,18]. Specifically, this model implies two classes of crossovers. The main type of crossover are spaced further apart than expected under a simple crossover model. However, these appear to coexist with those that "escape" the interference effect, and therefore appear to be spaced closer together. In our study, which was one of the largest pedigree-based studies of human recombination, older mothers were shown to have a higher proportion of crossovers that escape crossover interference and this proportion appears to increase linearly with maternal age[19]. The escaping crossovers appear to bypass the regulatory effect of crossover interference. However, beyond this study, no other reports have given any clues to changing interference properties with age in humans.

Here, we present an analysis of crossover interference using single cell data from a number of previously published studies, which used alternate methods to study interference. In addition, we re-analyze pedigree data obtained through our collaboration with 23andMe, but using an additional set of older parents that were not present in the original report. The data presented here provide further insight into the properties of recombination in humans and how these properties vary with age.

3.2 Methods

3.2.1 Extended pedigree data from older parents

Crossover data was obtained through a collaboration with the personal genomics company 23andMe, as described previously (Chapter 2, Campbell *et al.*[19]). In the original publication ages of older

parents were omitted for privacy reasons if mothers were over the age of 40 or if fathers were over the age of 45 when their children were born. Here, we obtained data with masked ages for an additional 399 meioses from mothers over the age of 40 and 398 meioses from fathers over the age of 45. All data collection procedures, genotyping, recombination event calling and filtering were performed as described previously.

3.2.2 Public data

Previously published crossover data was obtained from the following sources:

Oocyte data. Crossover calls obtained from female oocytes, described in Hou *et al.* [12], were obtained by request from the authors. The data are from 8 Asian female donors, and consist of crossover calls made from all four products of meiosis: polar body 1 (PB1), polar body 2 (PB2), and the female pronucleus (FPN).

Sperm data Crossover calls from single cell sperm data were obtained from two sources. Wang *et al.* [10] provide data on 2,075 crossover events from 91 single sperm cells from the same individual, a 40 year old Caucasian male. Data from Lu *et al.* [11] were obtained via request from the authors. In this study, 99 sperm cells from a single Asian male were analyzed and crossover events identified.

3.2.3 Calling crossover events

Crossover interference was modeled according to the two pathway model, also known as the gamma-escape model. In this model, two types of crossovers exist together in a mixture model. The distance between interfering crossovers are modeled according to a gamma distribution, with shape ν and rate 2ν. Non-interfering crossovers follow a similar distribution of inter-event distances, but with $\nu = 1$, representing no interference, or a random placement of events. A second parameter, p, represents the proportion of these non-interfering crossovers within the mixture. To estimate these interference parameters in this data we used a MATLAB software package (`https://github.com/auton1/interference`) previously described (Chapter 2, and Campbell *et al.* [19]).

3.3 Results

3.3.1 23andMe data with older parents

The full dataset consists of 19,099 meioses (9,551 female, 9,548 male), of which 797 are additional data from older parents (399 female, 398 male). The distribution of the parental ages is shown in Figure 3.1 for phase-known and phase-unknown meioses separately.

We estimated crossover interference parameters using the gamma-escape model of interference[18] to examine genetic distances between crossover events. In order to investigate age effects on crossover interference, we previously divided the data into 10 bins by quantile on the basis of age. Here, we added an additional bin containing all age-masked individuals. We observed a linear increase in the number of events escaping interference with increased maternal age, rising from 5.8% in mothers under 25 to 8.2% in mothers in the 35-40 age bin. Comfirming the expectation from the published analysis, estimates from mothers in the masked group showed a further increase in the proportion of escaping events, with 8.8% of events estimated to be interference escapers (Figure 3.2). Again confirming the expectation of the published data, no change with age was seen in strength parameter estimates for females, or for either parameter in males.

3.3.2 Crossover interference within individuals

We next estimated parameters of crossover interference using previously published data from human sperm and oocytes. In sperm, we use data from two different studies, one with 99 sperm cells obtained from an Asian male[11], and another using 91 sperm cells from a Caucasian male[10]. From a study in oocytes, we obtain data from 8 individual Asian females[12].

Using this data, we estimate interference parameters under the gamma-escape model on an individual basis, and compare to group-level estimates from the 23andMe dataset[19]. Using all available data from the oocyte study, we pool data from the FPN, PB1, PB2 for a single individual. All available sperm data is pooled for each individual. Parameter estimates for the gamma escape model are shown in Figure 3.3. Among female individuals, there is large amount of variation in the estimates for the interference strength parameters. Most of the single cell female samples overlap with the grouped estimates from the 23andMe study ($\nu = 7.13$, 95% confidence interval (CI) 6.95–7.33). Two samples

have estimates outside this range, with sample S02 lower ($\nu = 5.19$, 95% CI 4.40–6.40), and sample S03 higher ($\nu = 11.12$, 95% CI 9.27–14.87) that the 95% confidence intervals from 23andMe females. In the two male sperm samples, the strength estimates are again similar to that of the 23andMe male estimates. However both sperm estimates have larger confidence intervals, and slightly higher point estimates of interference strength.

When comparing the escape parameter, the female individuals had estimates that were notably lower than the 23andMe samples, in which the escape estimate, p, was 0.071, 95% CI 0.067–0.075. Two samples, S02 and S04, had escape estimates that were 0 or whose confidence intervals did not exclude 0, indicating a very low degree of escape in these individuals. Two individuals, S03 and S07, had estimates that overlapped with those from 23andMe females, with the remaining samples having estimates of 1–3%. In contrast, the male sperm samples both overlapped with the 23andMe male estimates, which have escape estimates of p= 0.059, 95% CI 0.054–0.064. However the confidence intervals were substantially wider than the group estimates, potentially owing to the smaller sample sizes.

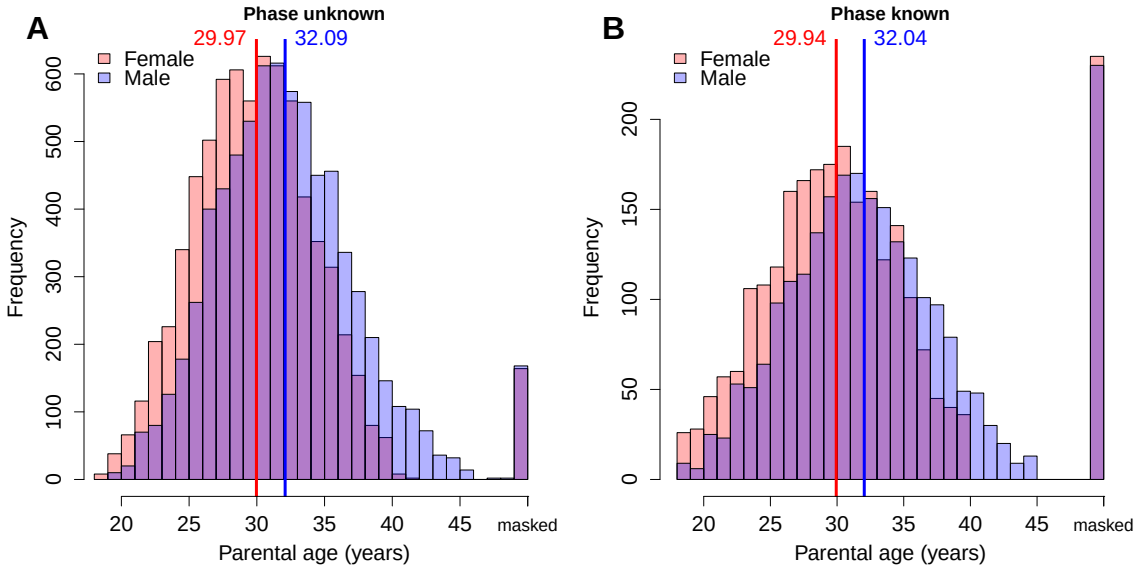

Figure 3.1: **Age distributions in the 23andMe dataset including older parents.** Phase-unknown parents are shown on the left panel (A), and phase-known parents, where ages were averaged across the children, are shown in the right panel (B). Additional data from older parents, in which the ages have been masked in females over 40 and males over 45 years old, are shown on the far-right of each distribution with the axis-label "masked." The vertical lines represent the mean of each distribution, excluding the older parents with masked ages.

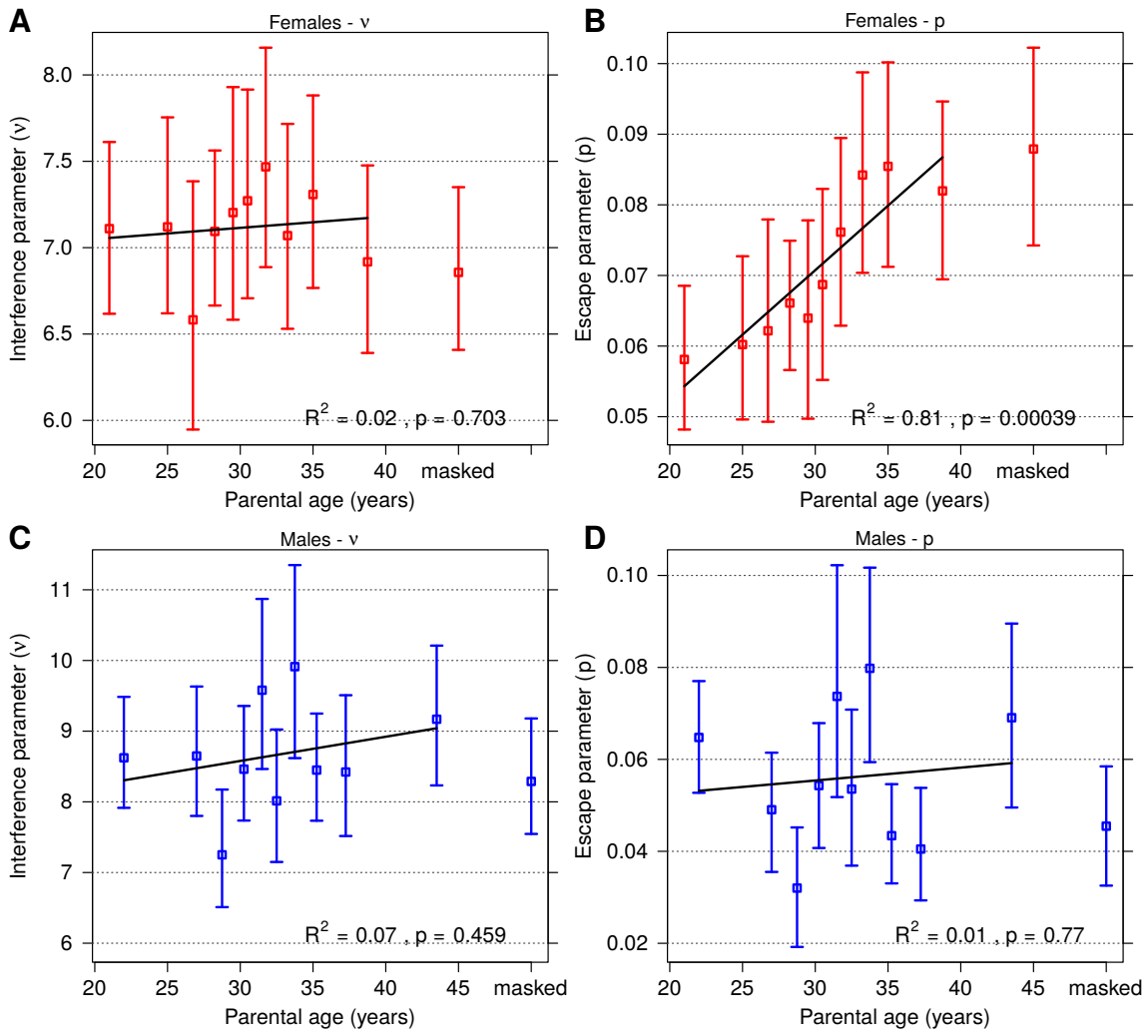

Figure 3.2: **Crossover interference parameters as a function of age.** Parameter estimates for females are shown in the top two panels (A and B), males in the bottom two panels (C and D). Interference strength estimates (ν) are shown in the left two panels (A and C), and estimates for the escape proportion are shown in the right panels (B and D). Data from older parents with masked ages are included in the plot with the axis label "masked." These older groups are excluded from the linear regression lines (shown in black).

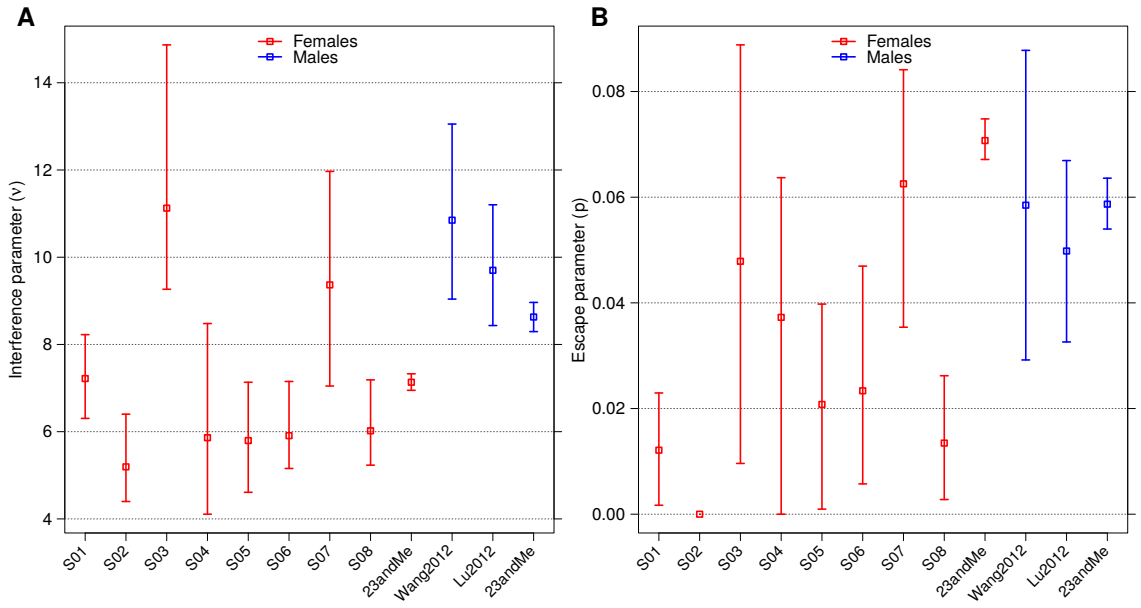

Figure 3.3: **Crossover interference parameters in single cell data.** Interference parameter estimates are shown for the gamma escape model. The strength parameter is shown in panel A, and the escape parameter in panel B. Data points for females (red) represent data data generated by Hou *et al.* [12], representing 8 individuals. Data for males (in blue) comes from Lu *et al.* [11] (n=99), Wang *et al.* [10] (n=99). 23andMe data [19] is provided for comparison and represents 2,184 females and 2,092 males).

3.4 Discussion

The questions of how recombination properties vary on an individual basis and with age have long been outstanding. I have presented here a series of analyses that attempt to shed light on these effects.

I have used an updated dataset which was previously presented without the age-masked individuals included here, both in this thesis (Chapter 2), and in a peer-reviewed journal article[19]. The original dataset was the first study of recombination in humans to find evidence for an increase in the proportion of events that escape crossover interference with increased age in females. The original data was suggestive of an increasing de-regulation of events with age in females. Considering the increased incidence of aneuploidy in older mothers[20], it was suggested that these observations could be related. Here, I have updated this finding to include data from older parents, which had been masked for privacy concerns. The updated dataset validates and extends the original finding that interference escape increases in older mothers, but not fathers. Although there is an overlap in the confidence intervals, the masked data in older mothers continues to increase from the previous bin and has no overlap with the youngest group, supporting the idea that the phenomenon is a linear increase.

In a separate analysis, I have taken previously published data from recombination studies that identified crossovers within individuals using multiple single cell assays. These data provide a picture of how much variation can be expected within a single individual, which appears to vary substantially. Although the number of individuals represented here is small, the techniques used to produce this data provide a promising avenue for future research as the cost and complexity to perform these analyses decreases. In the future, such analyses could be performed on a much larger scale, and provide valuable insight into the variation of recombination properties within individuals.

3.5 References

1. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310(5746):321–4 (2005). doi: 10.1126/science.1117196.

2. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61 (2007). doi:10.1038/nature06258.

3. Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327(5967):836–40 (2010). doi: 10.1126/science.1183439.

4. Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., *et al.* Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327(5967):876–9 (2010). doi:10.1126/science.1182363.

5. Parvanov, E. D., Petkov, P. M., and Paigen, K. Prdm9 controls activation of mammalian recombination hotspots. *Science* 327(5967):835 (2010). doi:10.1126/science.1181495.

6. Coop, G., Wen, X., Ober, C., Pritchard, J. K., and Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319(5868):1395–8 (2008). doi:10.1126/science.1151851.

7. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., *et al.* The landscape of recombination in African Americans. *Nature* 476(7359):170–5 (2011). doi:10.1038/nature10336.

8. Lynn, A., Ashley, T., and Hassold, T. Variation in human meiotic recombination. *Annual Review of Genomics and Human Genetics* 5:317–49 (2004). doi:10.1146/annurev.genom.4.070802.110217.

9. Jeffreys, A. J. and May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics* 36(2):151–6 (2004). doi:10.1038/ng1287.

10. Wang, J., Fan, H. C., Behr, B., and Quake, S. R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 150(2):402–12 (2012). doi: 10.1016/j.cell.2012.06.030.

11. Lu, S., Zong, C., Fan, W., Yang, M., Li, J., *et al.* Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338(6114):1627–30 (2012). doi: 10.1126/science.1229112.

12. Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., *et al.* Genome analyses of single human oocytes. *Cell* 155(7):1492–506 (2013). doi:10.1016/j.cell.2013.11.040.

13. Kong, A., Barnard, J., Gudbjartsson, D. F., Thorleifsson, G., Jonsdottir, G., *et al.* Recombination rate and reproductive success in humans. *Nature Genetics* 36(11):1203–6 (2004). doi:10.1038/ng1445.

14. Martin, H. C., Christ, R., Hussin, J. G., O'Connell, J., Gordon, S., *et al.* Multicohort analysis of the maternal age effect on recombination. *Nature communications* 6:7846 (2015). doi:10.1038/

ncomms8846.

15. Bleazard, T., Ju, Y. S., Sung, J., and Seo, J.-S. Fine-scale mapping of meiotic recombination in Asians. *BMC genetics* 14(1):19 (2013). doi:10.1186/1471-2156-14-19.

16. Hussin, J., Roy-Gagnon, M.-H., Gendron, R., Andelfinger, G., and Awadalla, P. Age-dependent recombination rates in human pedigrees. *PLoS Genetics* 7(9):e1002251 (2011). doi:10.1371/journal.pgen.1002251.

17. Broman, K. W. and Weber, J. L. Characterization of human crossover interference. *American Journal of Human Genetics* 66(6):1911–26 (2000). doi:10.1086/302923.

18. Housworth, E. and Stahl, F. Crossover Interference in Humans. *American Journal of Human Genetics* 73(1):188–197 (2003). doi:10.1086/376610.

19. Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D., and Auton, A. Escape from crossover interference increases with maternal age. *Nature Communications* 6:6260 (2015). doi:10.1038/ncomms7260.

20. Hassold, T. and Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics* 2(4):280–91 (2001). doi:10.1038/35066065.

Chapter 4

A pedigree-based map of recombination in the domestic dog genome

Christopher L Campbell[1], Claude Bhérer[1*,2], Bernice E Morrow[1], Adam R Boyko[3], and Adam Auton[1*]

This manuscript is currently under review:

[1] Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, New York 10461, USA.
[2] New York Genome Center, New York, New York 10013, USA
[3] Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA
[*] Former affiliation.

Correspondence and requests for materials should be addressed to
A.A. (email: adam.auton@gmail.com).

Abstract

Meiotic recombination in mammals has been shown to largely cluster into hotspots, which are targeted by the chromatin modifier PRDM9. The canid family, including wolves and dogs, has undergone a series of disrupting mutations in this gene, rendering *PRDM9* inactive. Given the importance of *PRDM9* it is of great interest to learn how its absence in the dog genome affects patterns of recombination placement. We have used genotypes from domestic dog pedigrees to generate sex specific genetic maps of recombination in this species. On a broad scale, we find that placement of recombination events in dogs is consistent with that in mice and apes, in that the majority of recombination occurs toward the telomeres in males, while female crossing over is more frequent and evenly spread along chromosomes. It has been previously suggested that dog recombination is more uniform in distribution than that of humans, however, we found that recombination in dogs is less uniform than humans. We examined the distribution of recombination within the genome, and find that recombination is elevated immediately upstream of the transcription start site, and around CpG islands, in agreement with previous studies, but find that this effect is stronger in male dogs. We also find evidence for positive crossover interference influencing the spacing between recombination events in dogs, as has been observed in other species including humans and mice. Overall our data suggests that dogs have similar broad scale properties of recombination to humans, while fine-scale recombination is similar to other species lacking *PRDM9*.

4.1 Introduction

The placement of recombination events within the genome is not random but instead is concentrated into regions known as hotspots. Recent work has identified the protein PRDM9 as responsible for targeting recombination events to hotspots[1–3]. PRDM9 is active early in meiotic prophase[4] and contains a zinc finger (ZF) array that binds to specific DNA motifs located at hotspot centers. Upon DNA binding, PRDM9 trimethylates lysine 4 on histone H3, and is presumed to recruit cellular machinery to initiate recombination through an unknown mechanism.

Recombination hotspots are a common feature of eukaryote genomes but are not typically conserved between species. For example, humans and chimpanzees have a complete absence of hotspot sharing, despite a high degree of overall DNA sequence identity[5–7]. This change in hotspot location appears to be driven by the rapid evolution of the ZF domain of *PRDM9*, which is subject to

strong selection in primates and rodents as well as a variety of ancient metazoans[8]. Alterations to the ZF domain modify DNA motif recognition and binding specificity[8] and hence contribute to a shifting landscape of active hotspots in the genome.

Evidence suggests that PRDM9 is required for the proper completion of meiosis. Loss of *PRDM9* causes sterility in male mice due to impairment of the progression of early meiotic prophase[4]. These mice, despite being sterile, still initiate double strand breaks (DSBs), and these breaks cluster into hotspots. However, there is almost no overlap with hotspots that occur in mice with functional PRDM9, and these DSBs occur preferentially at promoters and CpG rich regions of the genome[9]. This pattern is similar to that in other species lacking *PRDM9*, including birds[10], and yeast[11].

The canid orthologue of *PRDM9* has been inferred to have undergone multiple truncating mutations in the last exon, encoding the ZF array, and become a nonfunctional pseudogene[12]. These mutations are shared within the Canidae family that includes dogs, coyotes, wolves, and foxes and must have accumulated after their divergence from pandas, who do not share the mutations, approximately 49 Mya[8,12–14]. Nonetheless, these species are all able to complete meiosis and reproduce, which implies either that the function of PRDM9 in dogs is replaced by another gene or that recombination is able to complete successfully in its absence. A rare homozygous loss of function mutation in *PRDM9* has been recently reported in humans, in which a healthy mother was found to have mutations predicted to abolish both methlytransferase and DNA binding activity[15], leading to reduced crossover activity at PRDM9-dependent hotspots. Transmission of the mutation was detected in one of her three healthy children, suggesting that humans may be able to successfully complete meiosis and remain fertile without functional PRDM9.

Despite the lack of *PRDM9*, hotspot-like regions of recombination have been inferred in dogs from patterns of linkage disequilibrium (LD). These hotspots differ qualitatively from those found in humans, appearing to have a lower intensity of recombination rate, and covering a wider genomic interval ($\sim$4-18 kb compared with $\sim$2 kb in humans)[13,14]. However, direct comparisons are complicated by differences in the general LD properties of the species arising from, for example, population demography[14]. Most striking is the observation that recombination is preferentially targeted towards CpG rich regions, such as those found in gene promoter regions, which is similar to recombination patterns found in other species without *PRDM9*.

Over the last three decades, the study of recombination in dogs has progressed from initial low-

coverage linkage maps[16,17], bolstered by the assembly of a draft sequence of the dog genome[18], to higher-coverage pedigree maps[19], and high-resolution LD based maps from single nucleotide polymorphism (SNP) array and whole genome sequence data[13,14]. Here, we present a pedigree analysis of recombination in the domestic dog, *Canis lupus familiaris*, using high-density SNP microarray data, which allows investigation of the sex-specific distribution of recombination in the dog genome. Given the open questions regarding the role of PRDM9 in recombination, we compared the sex-specific landscape of recombination in the dog genome to that inferred from human pedigrees, and used this comparison to gain further insight into the effects of the presence or absence of PRDM9 on the mammalian recombination landscape.

4.2 Methods

Genotyping. The full dataset is derived from genomic analysis of 237 DNA samples, with 25 founder individuals (15 male, 10 female, pedigree structure shown in Figure 4.S1) from a colony of Labrador Retriever and Greyhound crosses maintained at Cornell University for over 30 years[20–22]. Genotyping was performed using genomic DNA as described in Hayward *et al.*[23] using Illumina CanineHD BeadChips that include more than 170,000 SNPs. All positions reported are given in canFam3.1 coordinates.

Filtering SNP data. To avoid spurious recombination calls due to genotyping error, we applied a set of filters on the variant data (outlined in Figure 4.S2). First, 586 SNPs were removed because they had missing genotypes in greater than 5% of the samples. We then used the PLINK[24] software (v1.07) to identify and remove a further 1,245 SNPs showing Mendelian errors in transmission (option `--mendel`). The error detection feature in the Merlin[25] software (v1.1.2, option `--error`) was used to identify and remove SNPs with genotypes that conflicted with pedigree structure and are likely to be genotyping errors. Three iterations of Merlin error detection were performed, removing 1,363, 66, and 7 SNPs, respectively. In total 2,771 unique SNPs were filtered out in the first round, leaving 163,400 SNPs for further analysis.

Calling crossover events. Autosomal recombination events were inferred using a combination of software tools. First, the dog genomes were phased without using pedigree information using

SHAPEIT2[26] (v2.r790). In order to avoid bias in our inference of recombination, we use a map file for phasing that has a constant rate of recombination (1 cM/Mb) between physical markers. Following phasing, we used the duoHMM[27] software (v0.1.4) to call recombination events using a hidden Markov model approach. The first duoHMM pass integrated pedigree structure information to correct phasing errors. Then, duoHMM was used to call crossovers in each parent-child duo of the pedigree, of which only high-confidence events were retained (with probability >0.5). The duoHMM method was also used to identify SNPs that have a high probability of genotyping error (which we removed if a SNP had a probability of error of >0.9). This method has been demonstrated to have a high sensitivity and low false discovery rate when compared to a standard Lander-Green[28] approach, such as that implemented in Merlin[25].

Filtering crossovers. Tightly clustered crossovers within individuals may occur naturally as a result of gene conversion, however another possibility is that genotype errors have caused false crossover calls. In many cases in our data we found double crossovers that belonged to a shared parent, and are clustered within the genome. Furthermore, these crossovers often used the same SNP for the interval boundaries. This strongly suggests genotyping error as a likely cause. We therefore removed any double crossovers if that cluster within 1 Mb, and if more than one meiosis transmitted from the same parent also has clustered crossovers with shared interval boundaries.

We further remove all crossovers attributed to meioses that have biologically abnormal crossover counts. We define thresholds separately for males and females, with a distribution centered around the median crossover count, and defining the boundaries as 4 standard deviations from the sex specific median crossover count, with the standard deviation estimated via the robust estimator of 1.4826 MAD (median absolute deviation).

Construction of the genetic map. Due to the high level of inbreeding and homozygosity in domestic dogs, it is often not possible to detect recombination events over a significant fraction of the genome within a given pedigree. For example, if a breeding pair have few heterozygous variants towards the telomeric ends of a given chromosome, then events occurring within such regions will be largely invisible as they will not be flanked by informative markers. Due to the high level of inbreeding within many dog samples, failure to account for this issue would result in an underestimate of the total

map length. To correct for this issue, we considered the location of informative markers within each pedigree, and scaled the genetic map accordingly.

In order for duoHMM to correctly identify a recombination event, it must be flanked by at least one heterozygous variant in the parent on each side. For each parent-child duo for which we are able to make crossover calls, we identified the positions of the first and last heterozygous variant on each chromosome, which represent the genomic range in which we are able to observe a crossover. Then, across all duos in our sample, we estimated the effective total number of meioses at each position along each chromosome (Figure 4.S3). The effective number of meioses was used in place of a fixed number of meioses when calculating the recombination fraction at each genomic interval.

Recombination fractions were converted to genetic distances using Haldane's map function. Comparing maps generated using the effective number of meioses to those using a fixed number of meioses, we observed an increase in autosomal map length for both females (59.7 cM) and males (49.2 cM), and an increase in the sex averaged map length of 48.6 cM (Figure 4.S4).

Estimation of crossover interference parameters. Crossover interference influences the spacing of crossover events when two or more occur on the same chromosome in the same meiosis. We modeled the distance between these crossovers using two models. The gamma model[29] assumes that the inter-crossover distances follow a simple gamma distribution with shape ν and rate 2ν, where ν is a unitless measure of the strength of crossover interference, with $\nu = 1$ representing no crossover interference, $\nu < 1$ representing negative interference (spaced closer than expected by chance), and $\nu > 1$ indicating positive crossover interference (spaced further apart than expected). The gamma-escape model, originally proposed by Housworth and Stahl[30] provides an extension to the simple gamma model. Here, crossovers that are governed by the interference effect ($\nu > 1$) are modeled to coexist alongside a subset of crossovers that escape interference ($\nu = 1$). A second parameter, p, is included to allow the second class of "escaping" crossovers to exist in a mixture with the interfering crossovers, and represents the proportion of events that escape interference. To measure crossover interference, we estimated the parameters ν and p using a MATLAB software package (`https://github.com/auton1/interference`) previously developed to analyze interference in humans[31].

In order to compare each of the fitted models, we used the Bayesian Information Criterion (BIC), which is given by: $BIC = -2\ln(L) + k\ln(n)$, where L is the maximum likelihood estimation from

the model fit, k is the number of free parameters in the model, and n is the number of observations. The model with the smallest BIC is preferred.

Gene annotations. Gene annotations for canFam3.1 were downloaded from Ensembl (build 81). We considered only protein coding genes located on the autosomes, and kept the longest isoform for each gene.

Thinning the human map. In order to make a valid comparison between dogs and humans with respect to the proportion of recombination occupying a given amount of sequence, we took a number of steps to ensure the datasets are as similar as possible. To enable sex-specific comparisons between the species, we used a human pedigree dataset (Campbell *et al.* [31]) rather than an LD-based map. As the human dataset is considerably larger, we randomly sampled 408 phase-known meioses (204 each from males and females) to match the size of the dog dataset. We then reduced the SNP density of the human data. The human dataset was genotyped on a microarray of higher density than we have available for dogs, which allows recombination events to be better resolved than what was possible for dogs. We therefore thinned the human dataset in an attempt to match the recombination event resolution in dogs using an *ad hoc* iterative process as follows: 1) Determine the inter-SNP distances for the human and dog datasets, 2) find SNPs that cluster more tightly in humans 3) remove a random subset of SNPs within each of these clusters, 4) iterate until the inter-SNP distributions and overall medians are similar between the two species. This iterative process yields a thinned framework of SNPs in the human dataset such that the new inter-SNP distances closely resemble those in dogs (Figure 4.S5). Following this thinning, we examined each individual crossover in the human data, and expanded the interval boundaries, if necessary, to the next available SNP in the newly selected framework. New sex specific genetic maps were then generated from this thinned data and used for the comparison to dogs.

Data availability. Supplementary data for this study, including sex-specific genetic maps, filtered crossover calls, and genotype data are available at
`https://github.com/clcampbell/dog_recombination`.

4.3 Results

Building the genetic map. We used SNP array genotype data from 237 domestic dogs to map recombination in the canine genome. After applying a series of filtering steps on the SNP data (Figure 4.S2), we identified crossovers using duoHMM, a tool previously developed to identify recombination events in human data[27]. Upon examination of the initial genetic maps, we identified several regions that exhibited biologically unrealistic recombination rates within concentrated physical regions and could be errors. Several of these regions overlap known segmental duplications and copy number variants[32,33], which could account for the observed clustering of crossovers. Another possible explanation is that the reference genome contains misplaced or inverted contigs within these regions, which would lead to calling of false crossover events on either side of the out of place region. We removed all 435 variants within four such regions, totaling 5.6 Mb of sequence (Table 4.S1), and a further 1,344 markers identified with the error detection feature of duoHMM.

After re-calling crossovers on this filtered data, we found 8,312 autosomal recombination events. We then identified and removed a set of clustered double crossovers, likely to be false calls, consisting of 90 events. Finally, we excluded meioses that have a biologically abnormal number of crossovers, excluding 3 female meioses (with 55, 41, and 119 crossovers), and 3 male meioses (with 5, 44, and 109 crossovers). In total we excluded 463 out of 8,312 crossovers (5.6%).

The filtered dataset consisted of 408 informative meioses, including 204 from females, and 204 from males. There are 7,849 well supported crossover events that could be localized to a median size of 102.1 kb (Figure 4.S6). The sex specific genetic maps had a mean resolution of less than 0.4 cM.

Comparison to previous studies. To assess the accuracy and validity of our results, we compared our maps to those from previous studies. We used sex specific maps from a previous pedigree analysis[19] as well as a sex-averaged LD-based map generated from whole genome sequencing data of 51 village dogs[14]. At the broad scale there is close agreement to our sex averaged map from the LD map (Pearson $r = 0.86$ at 5 Mb resolution), and from the pedigree sex averaged map ($r = 0.75$, Figure 4.S7A). The male map has a higher agreement with previous studies (LD $r = 0.80$, pedigree $r = 0.76$, Figure 4.S7B) than does the female map (LD $r = 0.67$, pedigree $r = 0.73$, Figure 4.S7C).

Study	Year	Female (cM)	Male (cM)	Ratio	Sex avg. (cM)
Mellersh et al.[16]	1997	1039	766	1.36	902.5
Neff et al.[17]	1999	1820	1290	1.41	1555
Wong et al.[19]	2010	2276	1909	1.19	2092.5
Axelsson et al.[13]	2012				3005
Auton et al.[14]	2013				2430
This study	2016	2162	1816	1.19	1978

Table 4.1: **Autosomal map length estimates.** Total map lengths are given in centimorgans, while the ratio represents the female to male map lengths. Sex specific map lengths are not available for the LD based maps.

Consistent with previous studies in dogs[16,17,19], and other mammals, including humans[31,34], females have a longer map length (2162 cM) than males (1816 cM, Tables 4.1 and 4.S2). We observed similar total genetic map lengths when compared to the Wong *et al.* [19] pedigree study, although our maps are slightly shorter (by 114 cM in females, 93 cM in males, Table 4.1). Map length is strongly correlated to physical length in both sexes (male r^2=0.82, female r^2=0.83), and in the sex averaged maps (r^2=0.88; Figure 4.S8). The ratio of female to male autosomal map length is 1.19, equivalent to Wong *et al.* [19], but notably lower than that of humans at 1.56[31].

Distribution of recombination. Previous studies showed that the recombination rate is elevated in telomeric regions and lower near the centromere, both in dogs[13,14,19], and other species[35]. We observed the same phenomenon, and this telomeric effect is largely driven by male recombination in both dogs and humans (Figures 4.1A, 4.S9). The opposite pattern is seen in two chromosomes, 27 and 32, supporting previous evidence[19] suggesting that the orientation of these chromosomes in the reference genome is likely reversed. Based on this, we reverse the physical coordinates for these two chromosomes for all further analyses.

We quantified the amount of recombination occurring at the telomeric ends of each chromosome arm by estimating the proportion of total recombination occurring in a physical window located at the telomeric end (Figure 4.1B). We found that 38.2% of recombination occurred within 5 Mb of the telomere in males, compared with 9.7% in females. Human males have a roughly equivalent proportion within this same region (30.2% within 5 Mb), while human females have a similar amount of recombination compared to female dogs (7.4% within 5 Mb). We conclude that, at a broad scale at least, the telomeric enrichment of recombination observed in male dogs is similar to that of humans.

Previous observations using LD recombination maps have raised the possibility that dog recombination may be more uniform in distribution throughout the genome than in humans due to the loss of *PRDM9*[13,14]. However, estimates from LD can be confounded by differences in the effective population size, which complicate such comparisons. Pedigree-based studies should not be subject to the same confounding issues, and to investigate this further, we examined the concentration of recombination rates across the genome using our pedigree maps, and compared this to human pedigree data[31]. This analysis is sensitive to the marker coverage and crossover resolution, so the human genetic maps have been reduced in resolution to match that of the dog data (see Methods, Figures

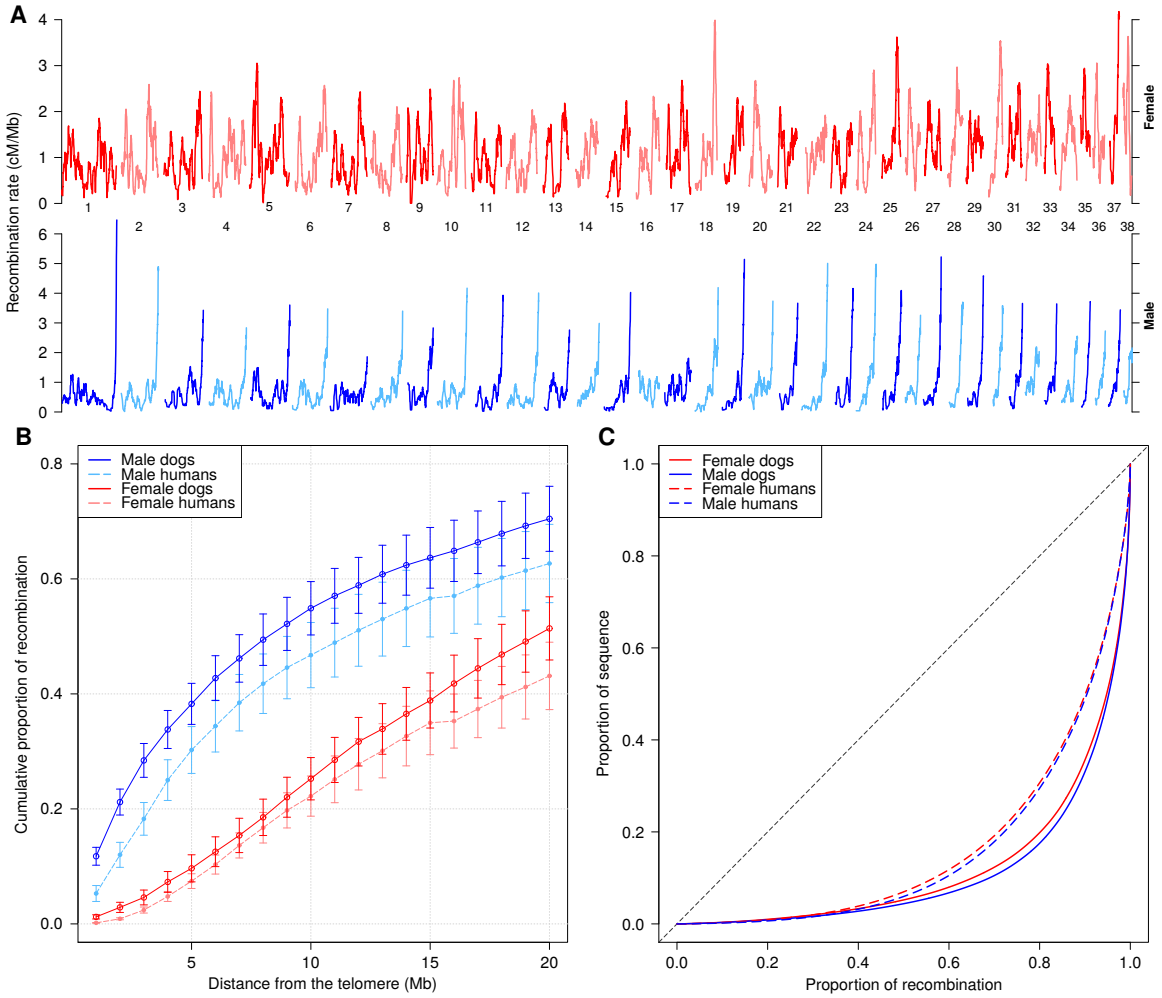

Figure 4.1: **The distribution of recombination across the genome.** (A) Broad scale recombination rates differ between males and females. Rates were smoothed at the 5 Mb scale. Chromosomes 27 and 32 are likely reversed in the canFam3.1 genome build, and are shown here with their physical coordinates reversed. (B) The proportion of recombination as a function of the distance from the telomeric end of each chromosome arm. Error bars represent a 95% confidence interval. (C) Proportion of recombination occupying various fractions of the sequence. The human data were thinned to match the SNP density and meiosis count of the dog dataset. For all panels, males and females are shown in shades of blue and red, respectively. Human data in panels B and C is shown in dashed lines.

4.S10A and C). We found that 80% of recombination occurred in a smaller proportion of the dog genome (17.5% male, 19.8% female, Figure 4.1C) than previously reported in LD-based estimates. In contrast to the LD-based findings, dog recombination is actually less uniform than in the human thinned data, in which males have 29.5%, and females 30.7% of their sequence containing the majority (80%) of recombination. In addition, males of both species have more focused recombination when compared to females. To address the possibility that differences in genome architecture or recombination rate distribution could account for these observations, we excluded telomeric regions from the analysis, and matched dog chromosomes with similarly sized human chromosome arms (Figures 4.S10B and D), with similar results. Therefore, it appears that crossovers are more concentrated within a smaller proportion of the dog genome than in humans, and that this effect is more pronounced in males of both species.

Recombination around genomic features. Starting from the observation that recombination is targeted to CpG islands concentrated at gene promoter regions[14], we looked for these effects in our data, and to what extent they are sex specific. We found that recombination rates were elevated around the TSS, both in the sex averaged and male maps, but no peak was discernible in females (Table 4.S3, Figure 4.S11A). Additionally, male recombination rate in the surrounding regions was higher than females, despite a lower genome wide recombination rate. This observation that can be partially explained by a modest enrichment in the number of genes (31%) occurring in the telomeric 25% of each chromosome where male recombination is more frequent. However, while the male background rate is higher in telomeric regions, males exhibited a peak at the TSS even in non-telomeric regions (Figure 4.S12A and B).

Both male and female recombination estimates showed elevated recombination surrounding CpG islands (Table 4.S3, Figure 4.S11B). The peak in male dogs was higher than females by 0.98 cM/Mb, with a high background rate in the surrounding sequence, which could be explained by clustering of CpGs, as well as an enrichment of CpG islands in telomeric, male driven recombination regions (42% in 25% of sequence, Figure 4.S12C and D). After thinning CpG islands to a uniform density throughout the genome, the male and female background rates were more comparable, but the male peak remains higher, suggesting that recombination around CpG islands is dominated by males (Figure 4.S13). We also examined recombination rates around H3K4 trimethylation marks found via ChIPseq

on dog spermatocytes[14]. As previously reported, the presence of these marks associated with elevated recombination rates, however this association can be explained by the proximity of CpG islands to H3K4me3 marks (Figure 4.S14), and we saw no differences between males and females.

Crossover interference. Crossover interference, a phenomenon that affects the physical spacing between pairs of crossover events occurring during the same meiosis, acts in various species, including humans[29–31], mice[36], and cattle[37]. To learn more about interference in dogs, we examined the distribution of inter-crossover distances in our dataset. We fit two models of crossover interference, the gamma model[29], and the gamma-escape model[30] (also known as the Housworth-Stahl model). The gamma-escape model is a mixture model that builds upon the gamma model, adding a subset of events that escape interference.

The no-interference model ($\nu = 1$), had a poor fit, with a lack of double crossovers in close proximity, indicating that positive crossover interference must be acting to some degree in dogs. When fitting the simple gamma model, estimates of interference strength in male and female dogs overlapped with each other. These estimates are comparable to a cytological study in dogs measuring the distance between MLH1 foci, which mark crossovers (ν=6.1)[38]. Comparing to humans, female dogs have a stronger strength of interference than human females, while the estimates for males of both species overlap (Figure 4.2A, Table 4.2).

In the interference-escape model, estimates of the strength of interference across the dog genome are higher in males than in females ($\nu_{female} = 14.05$, $\nu_{male} = 30.64$). This trend is similar to that seen in humans, with stronger crossover interference in males, however the parameter estimates are higher by a factor of 2 in females, and more than 3 in males ($\nu_{female} = 7.19$, $\nu_{male} = 8.93$ in humans, Figure 4.2B). In contrast, male dogs have a similar proportion of escaping events to humans (5.5% vs 5.9%), while female dogs have fewer events (3.5%) escaping than human females (7.1%, Figure 4.2C). We found support for both models of interference in the dog dataset, however we used BIC to make a formal comparison of the goodness of fit for each model. We found that in both sexes the gamma-escape model is preferred over the simple gamma model (Table 4.2), in agreement with previous findings supporting a two-pathway model of crossover interference in humans[30,31].

To test if this difference reflects a change in interference parameters with parental age, as is observed in humans[31], we divided our dataset by age into 7 approximately equal sized bins. No

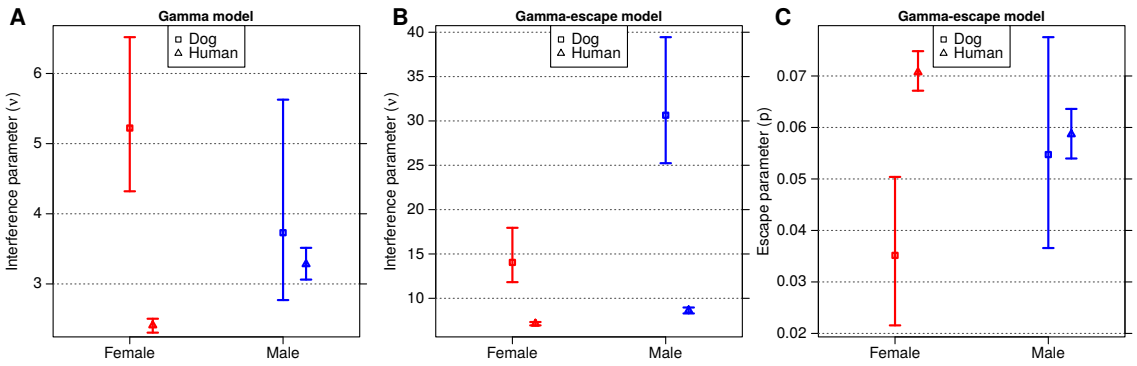

Figure 4.2: **Estimates of crossover interference parameters in the dog genome** using the simple gamma model (A) and Housworth-Stahl gamma-escape model (B and C). Panels A and B show the interference strength parameter, ν, for each model, while the right panel (C) shows the escape parameter, p, the proportion of events that escape interference. Males are shown in blue and females in red, while estimates for dogs are shown in boxes, and humans in triangles. The error bars represent a 95% confidence interval estimated from 1000 bootstrap iterations.

	Gamma model		Gamma-escape model		
	ν (95% CI)	BIC	ν (95% CI)	p (95% CI)	BIC
Male dogs	3.73 (2.77-5.63)	6617.1	14.05 (11.82-17.95)	0.055 (0.037-0.078)	6174.6
Female dogs	5.22 (4.32-6.52)	7580.9	30.64 (25.23-39.44)	0.035 (0.022-0.050)	7274.0
Male humans	3.28 (3.06-3.51)	99043.7	8.63 (8.29-8.96)	0.059 (0.054-0.064)	93205.3
Female humans	2.41 (2.31-2.50)	168775.7	7.13 (6.95-7.33)	0.071 (0.067-0.075)	156793.4

Table 4.2: **Autosomal crossover interference.** Parameter estimates are shown for both gamma and gamma-escape models for combined autosomes in dogs and humans. Numbers in parentheses represent 95% confidence intervals. BIC, Bayesian Information Criterion.

differences were observed for crossover parameters for either model in any of the age groups (Figure 4.S15). However, we should note that bootstrapped estimates produce large error bars and there is likely insufficient power to adequately detect any age differences in this dataset.

When reducing the resolution of the human dataset (see Methods), we found that the parameter estimates were largely unchanged compared to those of the full resolution data, although with wider confidence intervals (Figure 4.S16). This provides confidence that parameter estimation using these models is robust to crossover interval size resolution and dataset power, and that the parameters estimated from our dog data are likely to accurately reflect those of this dataset.

4.4 Discussion

Since the discovery of PRDM9 and its importance to recombination, questions about its full role have persisted. While *PRDM9* is under selection across a variety of species[8], a notable subset are missing a functional version of this protein, raising questions regarding the landscape of recombination in these species. Our pedigree study adds to existing work and provides insight into recombination in the absence of *PRDM9*. On a broad scale, our results are in agreement with those from previous studies, demonstrating that dog recombination is similar to other mammals, indicating that the presence or lack of *PRDM9* does not change broad scale patterns of crossover placement. In particular, a majority of crossing over occurs in telomeric regions in males, while female crossing over is both more frequent and more spread out.

We report a ratio of female to male map lengths of 1.19, equivalent to previous estimates from Wong *et al.*[19]. This groups domestic dogs with a large collection of species that exhibit sexual dimorphism in recombination in which the female has a higher rate of crossing over, including humans[31] and mice[39]. In contrast, cattle are one of the few species that have the opposite trend, with a recent study in domestic cattle estimating the male map length to be 10% longer than females[40]. An interesting suggestion from this study was that the overall recombination rate in males may have been affected by artificial selection pressure. Because artificial selection is more frequently focused on males, this can result in an increase in recombination if selection acts positively on recombination rate. If true, it is not implausible that this selective pressure could have altered recombination in dogs during their domestication as well, something that could be revealed through a comparison to wolves,

the closest ancestor to the modern domestic dog.

Initial estimates using LD maps in dogs indicated that 80% of all recombination falls into a fairly large (30-46%) amount of sequence[13,14], markedly more spread out than the <20% figure seen in human LD maps[41]. This supports the idea that PRDM9 acts to funnel recombination into hotspots in humans, and supports the hypothesis that dog recombination, lacking this hotspot specifying protein, is more uniform across the genome. While further investigation is necessary, our findings here, using pedigree data, suggest that dog recombination may actually be less uniform than humans. Furthermore, in both species, males appear slightly more focused than females. This effect in humans could potentially be explained by a higher male hotspot usage[31]. In dogs, this could be due to higher male rates around gene promoter regions and CpG islands that are concentrated towards the telomeres. This concentration of recombination at functional genomic elements is not unique to dogs, but appears to be shared among a number of species lacking *PRDM9*, including *PRDM9* knockout mice[9], Arabidopsis, yeast[11], and birds[10].

The concentration of recombination at these functional elements supports a working model for *PRDM9*-absent species, in which recombination occurs preferentially in regions of open chromatin. Another implication is that recombination hotspots in dogs and other *PRDM9*-absent species may be stable in evolutionary time, in contrast to current evidence against hotspot sharing in *PRDM9* dependent species. Since dog hotspots lack a strong motif that is likely to be targeted by a trans acting factor such as PRDM9[13,14], they are not likely to be subject to the hotspot paradox that acts to continually erode the binding capacity of hotspots, even as they are actively being used for recombination[2]. Evidence for hotspot stability has been found in two finch species, which share recombination hotspots that appear to be separated by tens of millions of years[10], as well as four yeast species sharing hotspots over 15 million years of evolution[11].

The distribution of inter-crossover distances in dogs supports the existence of positive crossover interference in the dog genome. Estimates of interference strength in the simple gamma model are roughly in line with those in humans. However, our results favor the gamma-escape model, supporting the idea that two separate pathways contribute to recombination in dogs. In this model, dog interference appears to be 2-3 times stronger than in humans, with a similar proportion of escaping events. Interestingly, while an increase in interference escape with age has been observed in human females[31], no such age effect was observed in dogs. Accepting that our canine sample size would

limit our ability to detect such effects, another potential explanation is that, in contrast to humans, the timing of meiotic events in dogs is substantially different. Recombination in human females begins and enters a potentially lengthy meiotic arrest prenatally, resuming just prior to ovulation. In contrast, meiosis in female dogs begins later, in the neonatal period. While recombination is complete prior to ovulation in humans, dogs ovulate immature oocytes, after which meiosis must complete before the oocyte becomes fertile, around 48 hours after ovulation[42,43].

Overall, these results add to a growing body of research in non-human recombination genetics, and provide a step towards answering many open questions in canine recombination. Further work is needed on larger and more diverse pedigrees, both in domestic dogs and other members of the Canidae family, including wolves, in order to form a more complete picture of recombination in this family.

4.5 Acknowledgments

We thank Yu Kong and Anthony Marcketta for their helpful discussions and assistance in the analysis of this data. We thank Liz Corey and the Cornell Veterinary Biobank for their assistance in pedigree analysis of the colony. C.L.C. was supported by the Training Program in Cellular and Molecular Biology and Genetics, T32 GM007491. C.B. was supported by a postdoctoral fellowship from the Fonds de la recherche en santé du Québec (FRQS). Data in this paper are from a thesis to be submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Graduate Division of Medical Sciences, Albert Einstein College of Medicine, Yeshiva University.

4.6 References

1. Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327(5967):836–40 (2010). doi: 10.1126/science.1183439.

2. Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., *et al.* Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327(5967):876–9 (2010). doi:10.1126/science.1182363.

3. Parvanov, E. D., Petkov, P. M., and Paigen, K. Prdm9 controls activation of mammalian recombination hotspots. *Science* 327(5967):835 (2010). doi:10.1126/science.1181495.

4. Hayashi, K., Yoshida, K., and Matsui, Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature* 438(7066):374–8 (2005). doi:10.1038/nature04112.

5. Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., *et al.* Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics* 37(4):429–34 (2005). doi:10.1038/ng1529.

6. Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308(5718):107–11 (2005). doi:10.1126/science.1105322.

7. Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., *et al.* A fine-scale chimpanzee genetic map from population sequencing. *Science* 336(6078):193–8 (2012). doi:10.1126/science.1216872.

8. Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., *et al.* Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genetics* 5(12):e1000753 (2009). doi:10.1371/journal.pgen.1000753.

9. Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. Genetic recombination is directed away from functional genomic elements in mice. *Nature* 485(7400):642–5 (2012). doi:10.1038/nature11089.

10. Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., *et al.* Stable recombination hotspots in birds. *Science* 350(6263):928–932 (2015). doi:10.1126/science.aad0843.

11. Lam, I. and Keeney, S. Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* 350(6263):932–937 (2015). doi:10.1126/science.aad0814.

12. Muñoz-Fuentes, V., Di Rienzo, A., and Vilà, C. Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. *PloS One* 6(11):e25498 (2011). doi:10.1371/journal.pone.0025498.

13. Axelsson, E., Webster, M. T., Ratnakumar, A., Ponting, C. P., and Lindblad-Toh, K. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Research* 22(1):51–63 (2012). doi:10.1101/gr.124123.111.

14. Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., *et al.* Genetic Recombination Is Targeted

towards Gene Promoter Regions in Dogs. *PLoS Genetics* 9(12):e1003984 (2013). doi:10.1371/journal.pgen.1003984.

15. Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* page aac8624 (2016). doi:10.1126/science.aac8624.

16. Mellersh, C. S., Langston, A. A., Acland, G. M., Fleming, M. A., Ray, K., *et al.* A linkage map of the canine genome. *Genomics* 46(3):326–36 (1997). doi:10.1006/geno.1997.5098.

17. Neff, M. W., Broman, K. W., Mellersh, C. S., Ray, K., Acland, G. M., *et al.* A second-generation genetic linkage map of the domestic dog, Canis familiaris. *Genetics* 151(2):803–20 (1999).

18. Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069):803–19 (2005). doi:10.1038/nature04338.

19. Wong, A. K., Ruhe, A. L., Dumont, B. L., Robertson, K. R., Guerrero, G., *et al.* A comprehensive linkage map of the dog genome. *Genetics* 184(2):595–605 (2010). doi:10.1534/genetics.109.106831.

20. Todhunter, R. J., Casella, G., Bliss, S. P., Lust, G., Williams, A. J., *et al.* Power of a Labrador Retriever-Greyhound pedigree for linkage analysis of hip dysplasia and osteoarthritis. *American Journal of Veterinary Research* 64(4):418–424 (2003). doi:10.2460/ajvr.2003.64.418.

21. Mateescu, R. G., Burton-Wurster, N. I., Tsai, K., Phavaphutanon, J., Zhang, Z., *et al.* Identification of quantitative trait loci for osteoarthritis of hip joints in dogs. *American Journal of Veterinary Research* 69(10):1294–300 (2008). doi:10.2460/ajvr.69.10.1294.

22. Phavaphutanon, J., Mateescu, R. G., Tsai, K. L., Schweitzer, P. A., Corey, E. E., *et al.* Evaluation of quantitative trait loci for hip dysplasia in Labrador Retrievers. *American Journal of Veterinary Research* 70(9):1094–101 (2009). doi:10.2460/ajvr.70.9.1094.

23. Hayward, J. J., Castelhano, M. G., Oliveira, K. C., Corey, E., Balkman, C., *et al.* Complex disease and phenotype mapping in the domestic dog. *Nature Communications* 7:10460 (2016). doi:10.1038/ncomms10460.

24. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81(3):559–75 (2007). doi:10.1086/519795.

25. Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 30(1):97–101 (2002). doi:10.1038/ng786.

26. Delaneau, O., Zagury, J.-F., and Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* 10(1):5–6 (2013). doi:10.1038/nmeth.2307.

27. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* 10(4):e1004234 (2014). doi:10.1371/journal.pgen.1004234.

28. Lander, E. S. and Green, P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* 84(8):2363–7 (1987).

29. Broman, K. W. and Weber, J. L. Characterization of human crossover interference. *American Journal of Human Genetics* 66(6):1911–26 (2000). doi:10.1086/302923.

30. Housworth, E. A. and Stahl, F. W. Crossover interference in humans. *American Journal of Human Genetics* 73(1):188–97 (2003). doi:10.1086/376610.

31. Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D., and Auton, A. Escape from crossover interference increases with maternal age. *Nature Communications* 6:6260 (2015). doi:10.1038/ncomms7260.

32. Nicholas, T. J., Cheng, Z., Ventura, M., Mealey, K., Eichler, E. E., *et al.* The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Research* 19(3):491–9 (2009). doi:10.1101/gr.084715.108.

33. Chen, W.-K., Swartz, J. D., Rush, L. J., and Alvarez, C. E. Mapping DNA structural variation in dogs. *Genome Research* 19(3):500–9 (2009). doi:10.1101/gr.083741.108.

34. Coop, G., Wen, X., Ober, C., Pritchard, J. K., and Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319(5868):1395–8 (2008). doi:10.1126/science.1151851.

35. de Massy, B. Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. *Annual Review of Genetics* 47:563–99 (2013). doi:10.1146/annurev-genet-110711-155423.

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

37. Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., *et al.* Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genetics* 8(7):e1002854 (2012). doi:10.1371/journal.pgen.1002854.

38. Basheva, E. A., Bidau, C. J., and Borodin, P. M. General pattern of meiotic recombination in male dogs estimated by MLH1 and RAD51 immunolocalization. *Chromosome Research* 16(5):709–19 (2008). doi:10.1007/s10577-008-1221-y.

39. Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., *et al.* A new standard genetic map for the laboratory mouse. *Genetics* 182(4):1335–44 (2009). doi:10.1534/genetics.109.105486.

40. Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., *et al.* Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLoS Genetics* 11(11):e1005387 (2015). doi:10.1371/journal.pgen.1005387.

41. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61 (2007). doi:10.1038/nature06258.

42. Freixa, L., García, M., and Egozcue, J. The timing of first meiotic prophase in oocytes from female domestic dogs (Canis familiaris). *Genome* 29(1):208–210 (1987). doi:10.1139/g87-036.

43. Chastant-Maillard, S., Viaris de Lesegno, C., Chebrout, M., Thoumire, S., Meylheuc, T., *et al.* The canine oocyte: uncommon features of in vivo and in vitro maturation. *Reproduction, Fertility, and Development* 23(3):391–402 (2011). doi:10.1071/RD10064.

4.7 Supplementary Information

Chromosome	Start	End	Size	# variants
6	44745965	47085070	2339105	182
16	53878711	56703603	2824892	221
19	20011075	20320803	309728	24
32	38654394	38810281	155887	8
		Total	**5629612**	**435**

Table 4.S1: **Regions removed from the dataset.**

CFA	Physical (bp)	First position (bp)	Last position (bp)	Female (cM)	Mean female rate (cM/Mb)	Male (cM)	Mean male rate (cM/Mb)	Sex avg. (cM)	Sex avg. rate (cM/Mb)	No. markers
1	122,678,785	4,283,592	122,309,715	95.54	0.78	85.17	0.69	90.12	0.73	8284
2	85,426,708	3,621,442	85,062,551	78.36	0.92	64.77	0.76	71.66	0.84	5550
3	91,889,043	5,604,604	91,556,345	75.63	0.82	62.64	0.68	68.77	0.75	6681
4	88,276,631	5,840,941	87,934,673	76.46	0.87	55.85	0.63	66.10	0.75	6231
5	88,915,250	1,243,143	88,673,195	95.67	1.08	68.00	0.76	81.70	0.92	6432
6	77,573,801	455,434	77,489,595	67.30	0.87	59.17	0.76	63.11	0.81	5320
7	80,974,532	180,153	80,809,723	70.31	0.87	47.59	0.59	58.81	0.73	5778
8	74,330,416	2,763,496	72,510,424	60.10	0.81	57.07	0.77	57.83	0.78	4846
9	61,074,082	876,259	60,812,630	62.66	1.03	50.65	0.83	55.67	0.91	4293
10	69,331,447	2,125,046	69,293,175	66.85	0.96	55.26	0.80	61.17	0.88	4573
11	74,389,097	4,087,888	74,253,347	58.71	0.79	48.36	0.65	53.60	0.72	4557
12	72,498,081	82,400	72,115,946	61.01	0.84	52.90	0.73	55.91	0.77	5364
13	63,241,923	4,067,434	62,932,928	54.05	0.85	45.71	0.72	49.52	0.78	4749
14	60,966,679	7,309,849	60,600,364	51.49	0.84	45.22	0.74	48.11	0.79	4153
15	64,190,966	4,913,124	64,007,939	48.69	0.76	44.57	0.69	46.43	0.72	4372
16	59,632,846	6,692,748	58,967,916	52.23	0.88	37.86	0.63	45.03	0.76	3829
17	64,289,059	5,285,642	63,501,532	63.44	0.99	50.07	0.78	56.58	0.88	4731
18	55,844,845	3,203,856	55,355,125	55.33	0.99	52.01	0.93	53.71	0.96	4014
19	53,741,614	3,189,264	53,349,320	53.22	0.99	49.99	0.93	52.29	0.97	3733
20	58,134,056	4,356,904	58,000,062	52.64	0.91	55.38	0.95	53.59	0.92	4140
21	50,858,623	4,450,666	50,719,350	51.66	1.02	44.16	0.87	47.84	0.94	3706
22	61,439,934	2,513,263	61,217,407	52.21	0.85	49.67	0.81	50.57	0.82	4422
23	52,294,480	1,203,392	52,291,577	49.31	0.94	44.47	0.85	46.63	0.89	3928
24	47,698,779	1,029,209	47,233,919	49.71	1.04	53.74	1.13	51.50	1.08	3625
25	51,628,933	6,038,820	51,469,123	56.49	1.09	48.57	0.94	52.44	1.02	3922
26	38,964,690	2,407,850	38,657,286	46.66	1.20	41.16	1.06	43.81	1.12	2853
27	45,876,710	444,525	42,191,669	48.31	1.05	48.70	1.06	47.01	1.02	3354
28	41,182,112	4,526,049	40,963,512	51.11	1.24	39.62	0.96	45.41	1.10	3074
29	41,845,238	913,671	41,543,120	49.96	1.19	42.06	1.01	44.75	1.07	3017
30	40,214,260	5,433,983	39,826,282	43.89	1.09	37.30	0.93	40.20	1.00	2942
31	39,895,921	580,882	39,466,279	51.82	1.30	38.07	0.95	44.32	1.11	2855
32	38,810,281	114,049	37,999,255	46.13	1.19	41.68	1.07	43.78	1.13	2746
33	31,377,067	447,033	30,994,965	45.46	1.45	36.01	1.15	40.75	1.30	2320
34	42,124,431	624,977	41,979,553	48.69	1.16	38.71	0.92	43.36	1.03	3239
35	26,524,999	1,164,664	26,257,078	41.87	1.58	30.59	1.15	35.88	1.35	2171
36	30,810,995	251,708	30,523,428	41.77	1.36	29.61	0.96	35.95	1.17	2286
37	30,902,991	1,364,851	30,583,437	45.59	1.48	36.18	1.17	40.42	1.31	2333
38	23,914,537	246,006	23,695,770	41.67	1.74	27.38	1.15	33.68	1.41	1895
			Total	2161.99	0.98	1815.91	0.82	1978.01	0.90	156,318

Table 4.S2: **Physical and genetic chromosome lengths.**

	Size (kb)	Male rate	Female rate	Difference (male - female)
TSS upstream	400	1.01	0.92	0.09
TSS	50	1.12	0.92	0.20
TSS downstream	400	1.03	0.94	0.09
CpG upstream	400	1.84	0.95	0.89
CpG island	50	2.09	1.11	0.98
CpG downstream	400	1.75	0.97	0.78

Table 4.S3: **Recombination rate around TSS and CpG islands.** Recombination rates are given in cM/Mb and were estimated in 10kb bins, and averaged over the indicated window. Rates surrounding CpG islands represent a 50kb window centered around the CpG island, while rates for the TSS are given as a 50kb bin ending at the TSS, capturing the elevated rate immediately upstream.

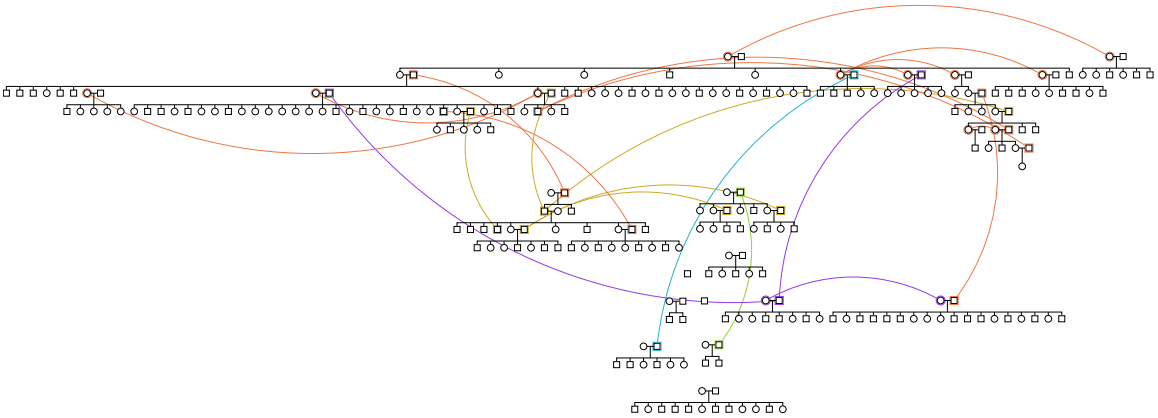

Figure 4.S1: **Structure of the dog pedigree.** Males are represented by squares, females by circles. Colored lines indicate individuals repeated on the plot, that are involved in more than one mating pair.

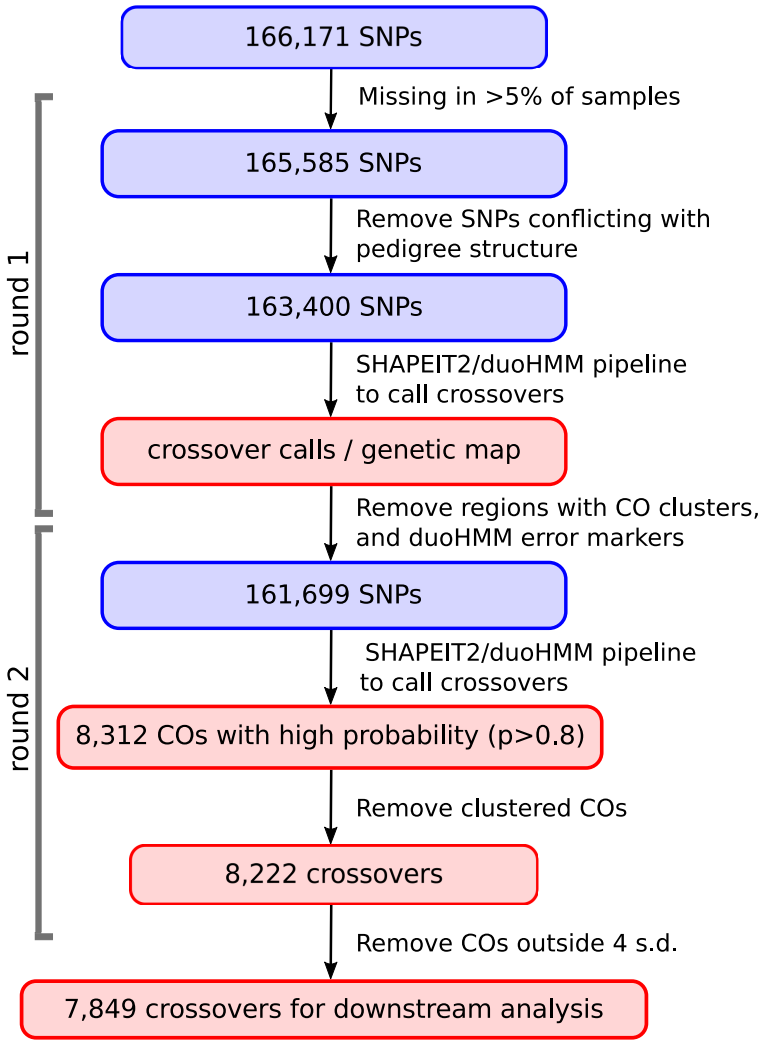

Figure 4.S2: **Overview of the analysis pipeline.** CO, crossovers; s.d., standard deviation.

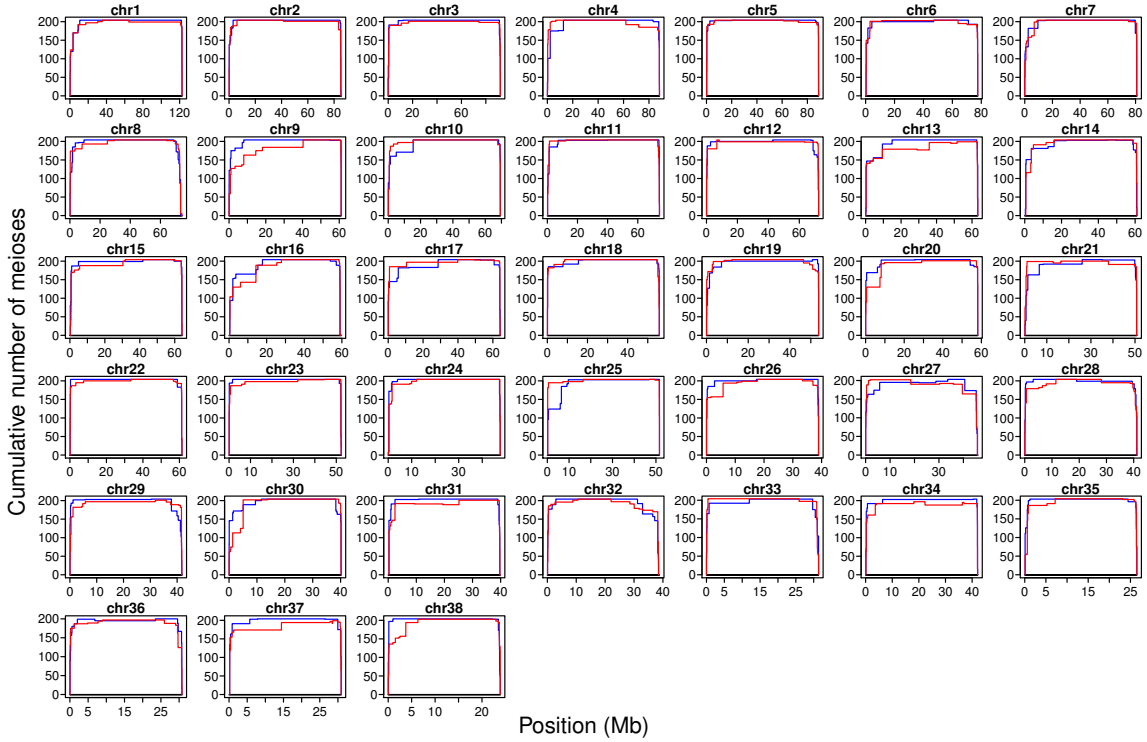

Figure 4.S3: **The effective number of meioses as a function of physical position is shown along each chromosome.** Red curves represent females (n=204), blue curves represent males (n=204).

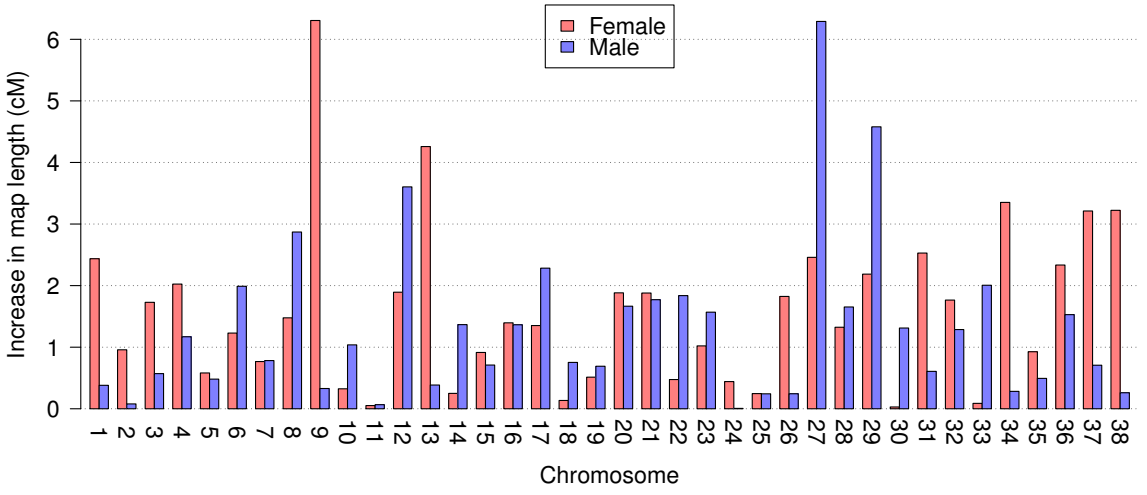

Figure 4.S4: **Increase in map length in each chromosome after accounting for the effective number of meioses.** Each bar represents the difference in map length after taking into account a reduced number of observable meioses towards chromosome ends compared to the map length calculated using a fixed number of meioses (n=204 for females in red, n=204 for males, blue).

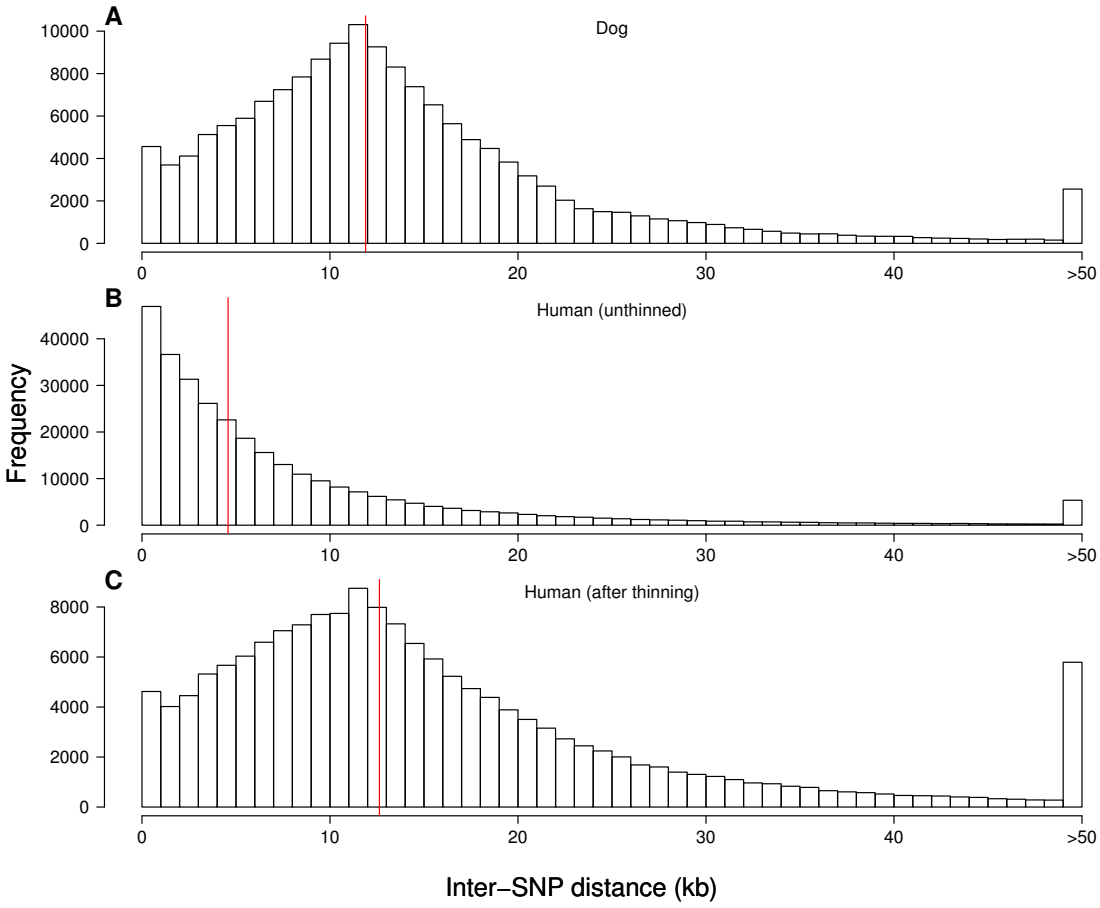

Figure 4.S5: **Distribution of inter-SNP distances in the dog data** (A), the human data prior to thinning (B), and the human data after the thinning procedure (C). The red line represents the median inter-SNP distance for each distribution.

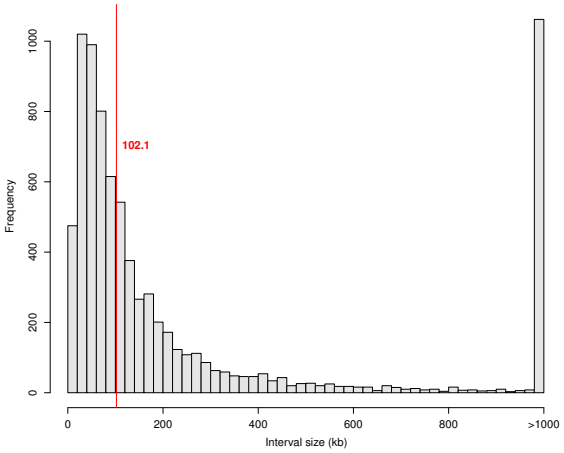

Figure 4.S6: **Distribution of crossover interval size.** The vertical red line represents the median interval size of 102.1 kb.

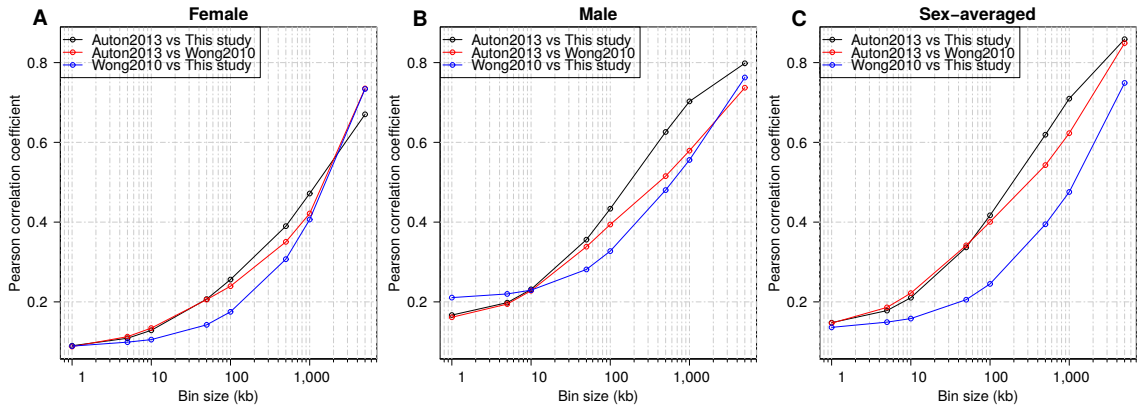

Figure 4.S7: **Pearson correlation between recombination rates** estimated from the Auton *et al.* [14] LD map, the pedigree maps from Wong *et al.* [19], and this study as a function of scale.

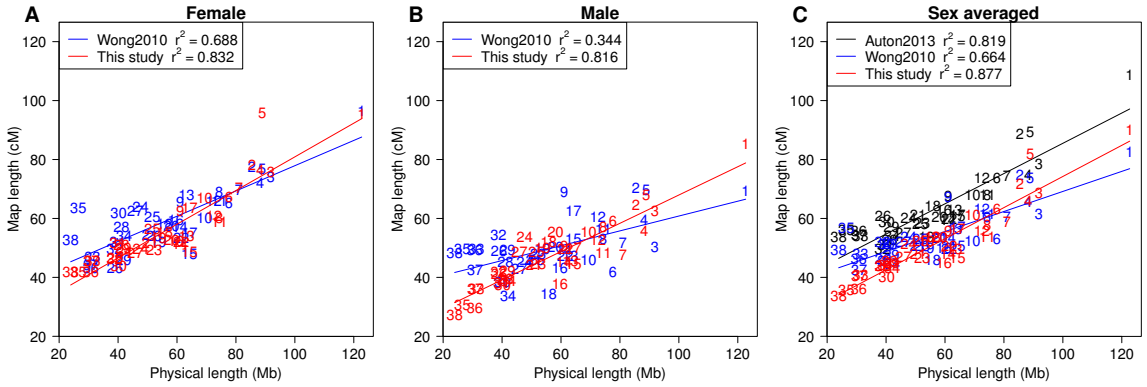

Figure 4.S8: **Map length as a function of physical length for each chromosome** for female (A), male (B), and sex-averaged (C) maps. Numbers refer to chromosomes with a linear regression line included. The sex-specific maps are compared to the Wong *et al.* [19] pedigree study, the sex-averaged map is additionally compared to the LD map from Auton *et al.* [14] (C, in black).

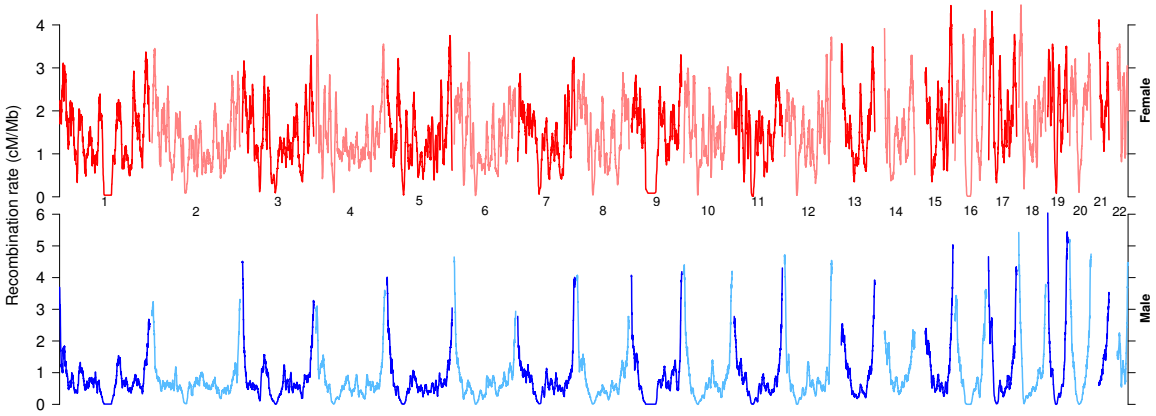

Figure 4.S9: **Recombination rate across the human genome** using the 23andMe genetic maps[31]. Female rates are shown in shades of red, male rates in blue shades. Recombination rates were smoothed at the 5 Mb scale.

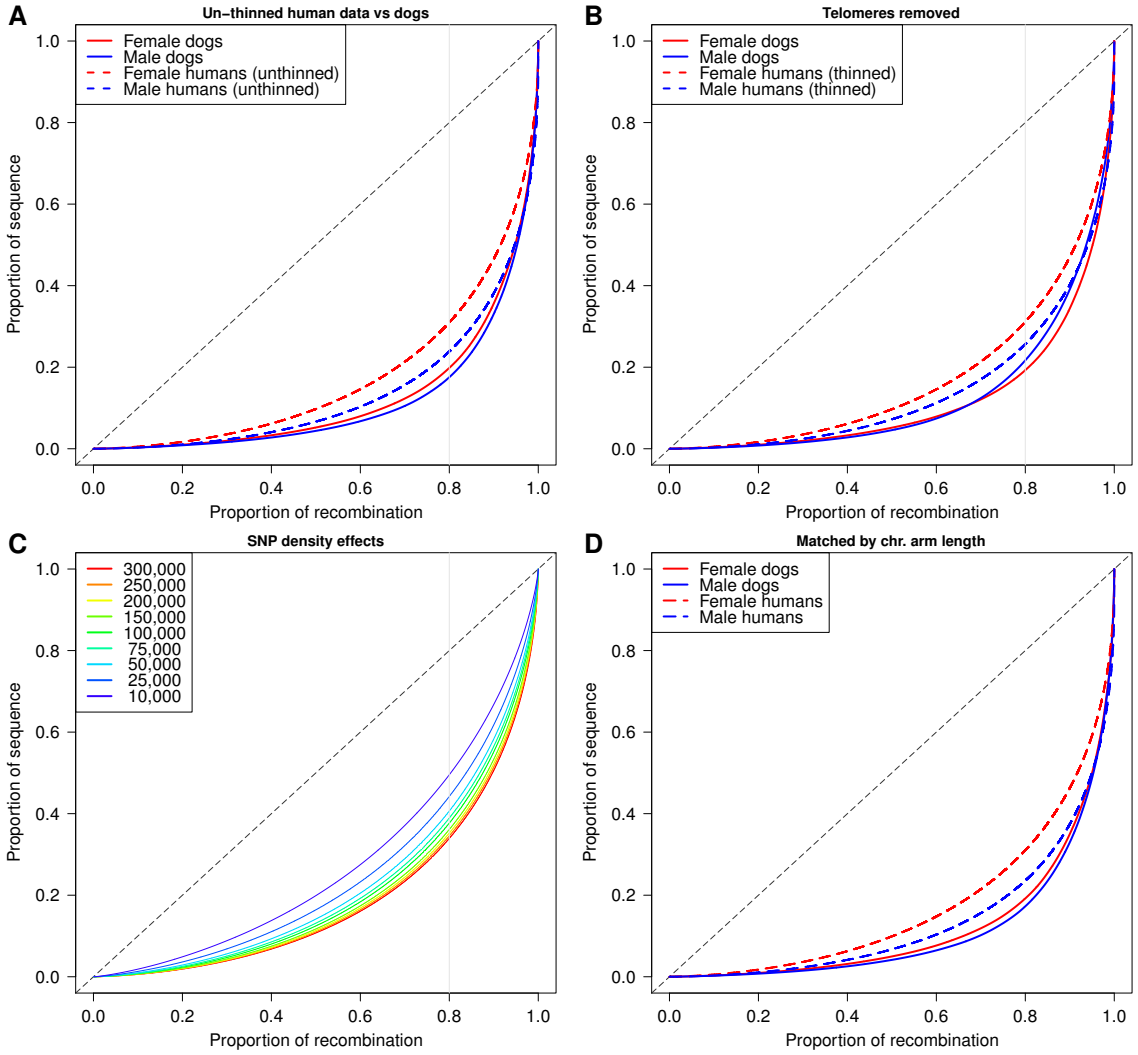

Figure 4.S10: **SNP density affects the proportion of recombination occupying various proportions of the sequence.** (A) Human data is shown prior to thinning, alongside data from dogs. Even without thinning in humans, dogs have a more concentrated distribution of recombination. (B) Dog data is shown with the thinned human data. Here, the most telomeric 15% (by physical distance) of each chromosome arm has been excluded for both species. Dog recombination remains more concentrated than humans, although the female and male dog curves have flipped at the 80% recombination mark, likely as a result of the removal of large amounts of telomeric recombination in males. (C) The effects of thinning the SNP framework used to create the genetic maps for the human data. Each curve represents a different marker density, from 300,000 SNPs (red line), to 10,000 (blue line). Reducing the SNP density moves the curve closer to unity and causes recombination to appear to be more spread out throughout the genome. (D) A reduced set of human and dog data are shown with the chromosome sizes approximately matching. Each dog chromosome was paired with a corresponding human chromosome arm of a similar size (within 30 Mb). This plot includes dog chromosomes 1 through 28; the remainder were too small to have potential matching human chromosome arms.

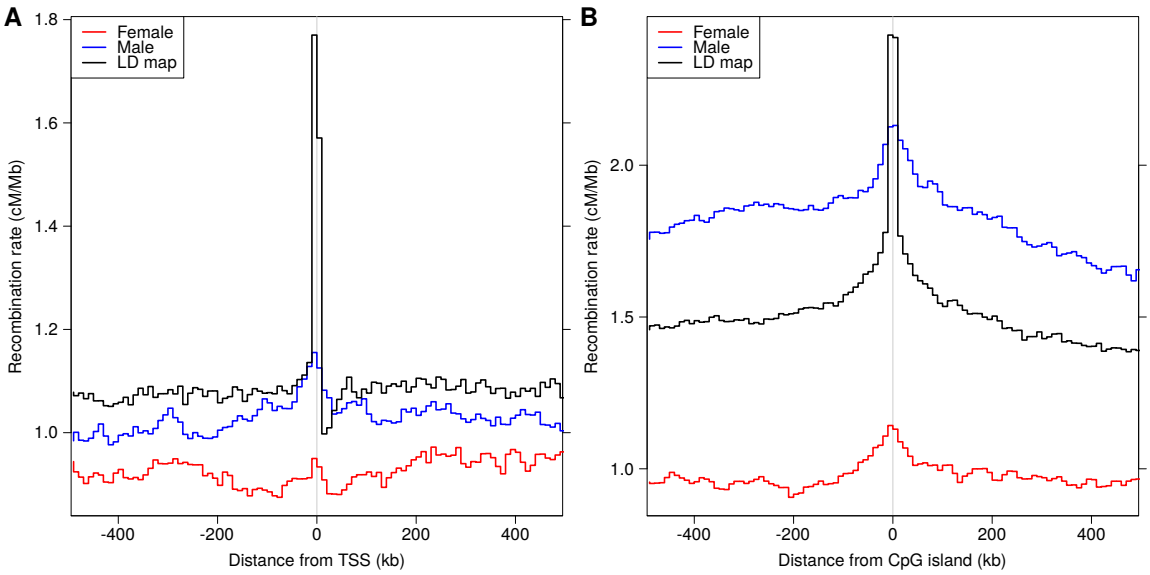

Figure 4.S11: **Sex differences in recombination** around the TSS (A) and CpG islands (B). Female rates are shown red, male in blue. LD-based estimates[14] are shown in black. Recombination rates were estimated in 10 kb windows.

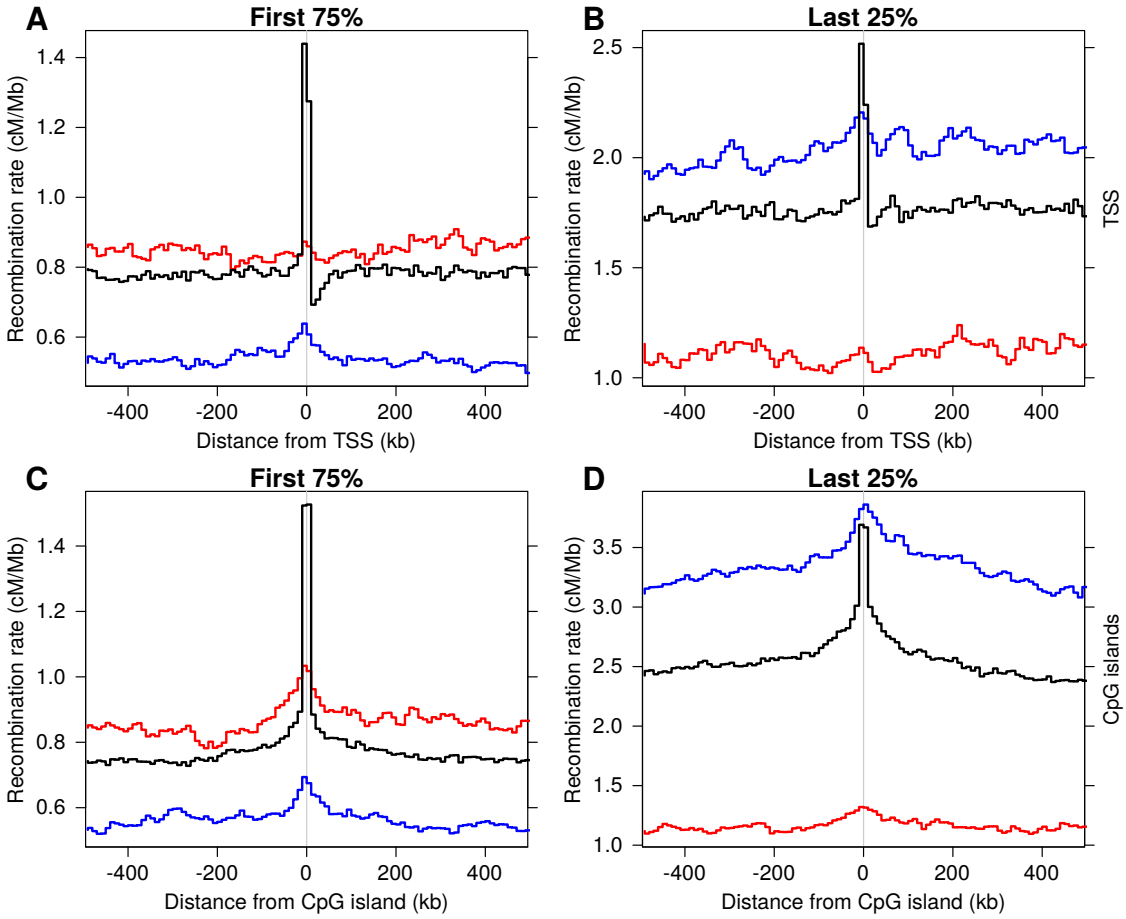

Figure 4.S12: **Recombination around TSS and CpG islands partitioned by chromosome position.** Male rates are in blue, female in red, rates from the LD map in black. Rates were estimated in 10 kb bins. Rates were estimated for each feature by taking the centromeric 75% (A and C) and telomeric 25% (B and D) of each chromosome separately.

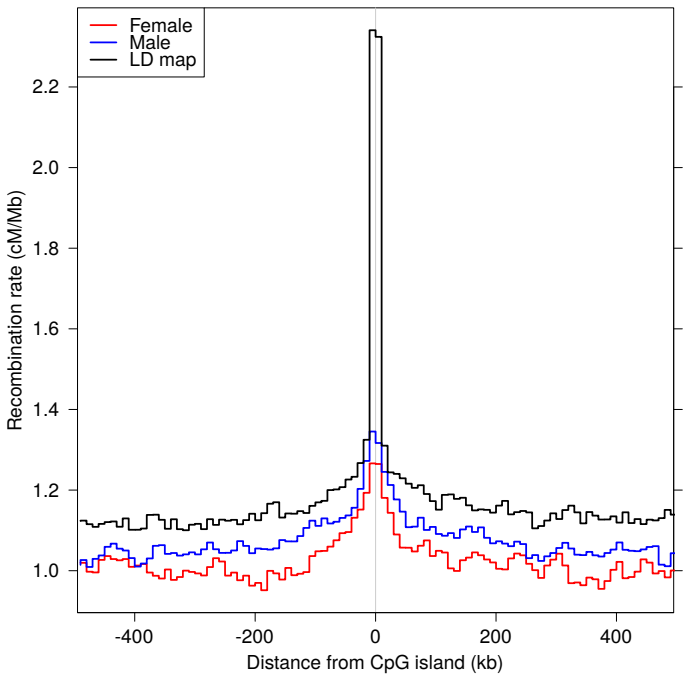

Figure 4.S13: **Recombination around a thinned subset of CpG islands.** Male rates are in blue, female in red, rates from the LD map in black. Rates were estimated in 10 kb bins. CpG islands were thinned to a uniform distribution by keeping a maximum of 5 per non-overlapping 500 kb window.

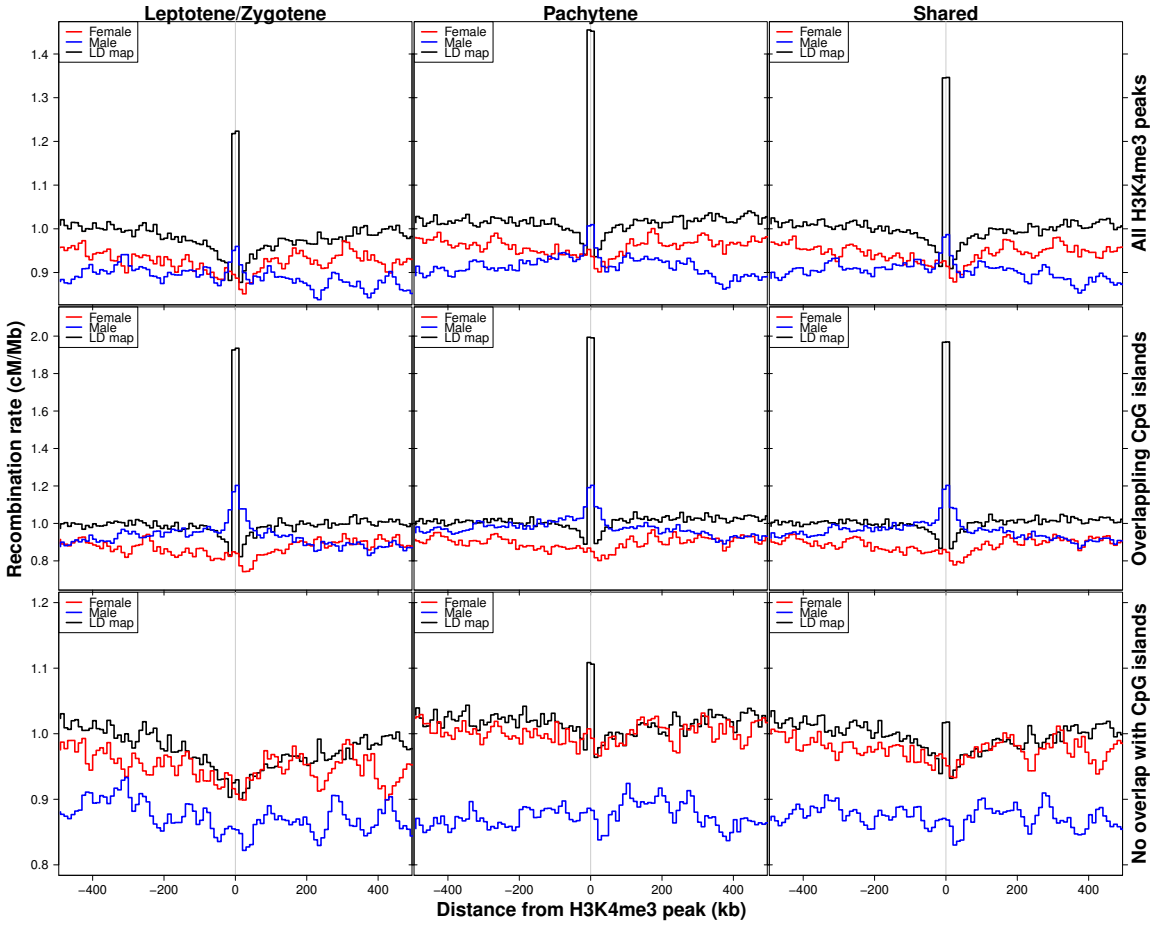

Figure 4.S14: **Recombination rate around H3K4 trimethylation marks found in dog spermatocytes of varying stages** Male rates are shown in blue, female in red, and rates from the LD-based map in black. The left column of plots shows peaks in early stages of meiosis (leptotene/zygotene), the middle column pachytene, and the right column the union of all peaks. Similarly, the top row includes all peaks, while the middle and bottom rows show peaks with and without overlap with CpG islands, respectively. Rates were estimated in 10 kb bins for all plots.

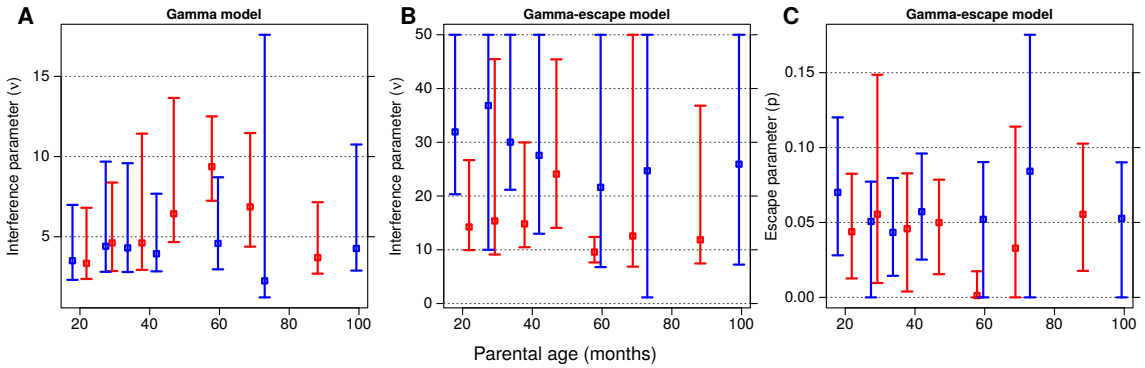

Figure 4.S15: **Estimates of crossover interference parameters in the dog genome as a function of age.** Dog meioses were partitioned into 7 approximately equal sized bins on the basis of parental age at birth. Interference strength for the simple gamma model is shown in A. The parameters for the Housworth-Stahl gamma-escape model are shown in B (interference strength) and C (escape). Males are shown in blue and females in red. The error bars represent a 95% confidence interval estimated from 100 bootstrap iterations.

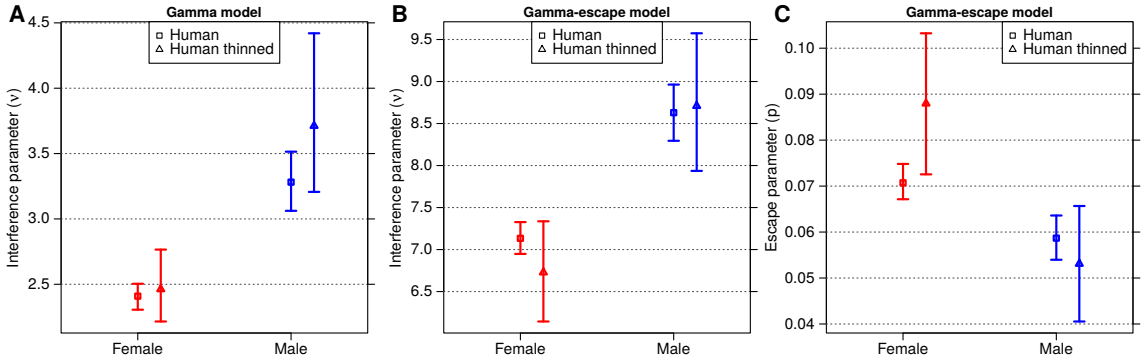

Figure 4.S16: **Estimates of crossover interference parameters in the human genome.** Here, the human data is thinned according to the procedure outlined in the Methods section, reducing the meiosis count and number of SNPs to more closely resemble what is found in dogs. Interference strength for the simple gamma model is shown in A. The parameters for the Housworth-Stahl gamma-escape model are shown in B (interference strength) and C (escape). Males are shown in blue and females in red. Estimates for the full resolution human data are shown in boxes, and the thinned human data in triangles. The error bars represent a 95% confidence interval estimated from 1000 bootstrap iterations.

Chapter 5

Detection of gene conversion in human admixed population genetic data

Christopher L. Campbell[1] and Adam Auton[1*]

This chapter contains unpublished data.

[1] Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, New York 10461, USA.
[*] Former affiliation.

5.1 Introduction

Recombination is a fundamental component of meiotic cell division, and the reshuffling of genetic variation has important evolutionary implications, enabling the action of natural selection. DNA double strand breaks (DSBs) resolve to one of two possible outcomes: crossover (CO) and non-crossover, or gene conversion (GC). Most research to date has been focused on the large-scale chromosomal exchanges that accompany CO, which are easily detected with a variety of direct and indirect methods. In contrast, gene conversion (GC), is a non-reciprocal process in which short segments of DNA are transferred from one parental chromosome to another.

Molecular studies support a model in which DNA DSBs occur at multiple points along a chromosome, and each of these DSBs undergoes a repair procedure that results in either CO or GC[1]. Sperm typing studies have estimated that GC occurs approximately ten times more frequently than crossover[1–3]. Gene conversion events are small, typically under 1000 bp, but could range from 50-2000 bp[2]. This small size therefore makes GC events difficult to detect using genome-wide inference methods such as pedigree studies or even linkage disequilibrium (LD) analysis.

Despite these difficulties, gene conversion has been successfully studied using pedigree approaches. A recent study used SNP array data from multiple three-generation pedigrees to identify approximately 100 GC events in humans[4]. Supporting results from molecular studies, gene conversions were found to occupy 100-1000 bp tracts. However, GC events were found to cluster unexpectedly, raising questions as to potential differences in the repair mechanism.

Gene conversion has been also shown to be biased in the exchange of alleles[5]. When occurring around a heterozygous SNP, GC can result in the non-Mendelian transfer of alleles, with the donor allele copied in a 3:1 ratio over the allele on the recipient chromosome. Furthermore, the copied allele is not chosen completely at random, and there is a dependency in which certain alleles are favored over others. Weak alleles (A and T) tend to be replaced with strong alleles (G and C). This is known as GC biased gene conversion (gcBGC), the over-transmission of G and C alleles during recombination. Biased gene conversion has evolutionary implications, and has contributed to an ongoing change on the base content of the genome[6].

Several statistical approaches have been proposed to create a model with which to detect gene conversion events from population genetic data. Gay *et al.*[7] created a hidden Markov model (HMM)

that models gene conversion together with crossover. This model was based upon a previous method for detecting recombination using patterns of LD, which modeled haplotype segments as a copied mosaic of previously observed haplotypes[8]. Another model, although it does not explicitly model gene conversion, is HAPMIX[9], which adapts the Li and Stephens model to perform ancestry deconvolution on admixed genomes. In this model, two divergent populations of haplotypes serve as a reference, and the admixed haplotype can copy from either population, with cross-population switches reflecting a change in haplotype ancestry. HAPMIX has been demonstrated to have a high sensitivity for partitioning admixed genomes[9].

The HAPMIX approach highlights a relatively new method of studying recombination in human populations having recent admixture. In this method, an admixed genome is modeled as a mosaic of haplotypes from two ancestral reference populations that have divergent patterns of allele frequencies (Figure 5.1). This technique was recently used to study recombination in African American populations[10]. This study found further evidence of population-specific hotspots, finding 2,500 hotspots unique to West Africans, and finding that these are tied to a novel PRDM9 binding motif.

Here, we will combine key features from each of these two models to develop a new method to detect gene conversion. Our method aims to detect gene conversion events using admixed population genetic data, where a gene conversion from on population will be detectable with increased contrast against the background of the other divergent population. From HAPMIX, we use the convention of modeling recombination separately both prior to, and after a point admixture event. From the Gay *et al.*[7] model, we model crossover and gene conversion with independent Markov chains. Using a combination of these two methods, our model provides a novel method to detect gene conversion events in humans using admixed genomes.

5.2 Methods

We provide here details of the implementation on the hidden Markov model used. We build upon the hidden Markov model framework put forth by Li and Stephens[8], in which an unknown haplotype is modeled as an imperfect mosaic of previously observed haplotypes. We adopt features from several extensions of this framework that have been used to address population genetics problems in recent years. The HAPMIX model[9] has been used successfully in the deconvolution of admixed genomes.

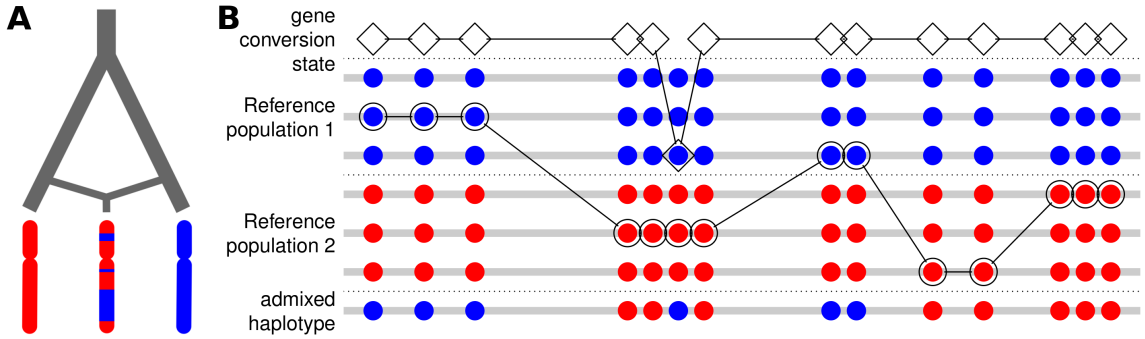

Figure 5.1: **Admixture approach to gene conversion detection.** (**A**) An admixed genome is shown (center) as a mixture of genomic segments from two divergent ancestral populations. (**B**) Diagram showing haplotypes in our HMM. Each horizontal line is a haplotype and each column of circles represents the alleles of a SNP, with blue and red colors designating alleles that are representative of two divergent ancestral reference populations. The haplotypes for the reference populations are used as templates for the admixed haplotype (bottom). The state path of haplotype transitions (X) is shown by open circles. The gene conversion state path (G) is shown in open diamonds. In this example there is one gene conversion event at SNP 6 in the admixed haplotype, which is copied from a haplotype in population 1. In all other SNPs there is no gene conversion ($G = 0$).

In HAPMIX, the Li and Stephens model is modified to include two separate groups of reference populations, corresponding to ancestral populations, from which segments of an unknown admixed haplotype can be assigned. The Li and Stephens model has also been modified for the detection of gene conversion events, which can be achieved by modeling gene conversion and crossover events simultaneously[7].

Here, we take key elements from each of these models: from HAPMIX the use of two divergent reference populations to increase contrast in ancestry, and from the Gay *et al.*[7] model the addition of a second Markov chain to model gene conversion events. Using the ancestry deconvolution approach we can assign blocks of sequence to one ancestral population or the other. In a block of sequence assigned to one ancestral population, gene conversion events from the other population will stand out against this background, and enable more accurate identification of these events.

5.2.1 Model details

We model gene conversion events along with haplotype transitions (crossovers) as independent Markov processes with transitions possible between each. In our model an admixed haplotype is composed of an imperfect mosaic of haplotypes from two ancestral populations, labeled P_1 and P_2, containing n_1 and n_2 haplotypes, respectively. We assume that the unknown, admixed haplotype underwent a point admixture event T generations in the past, in which a proportion, μ_1, of that haplotype's ancestry comes from P_1, with the remainder, $\mu_2 = 1 - \mu_1$, contributed from P_2.

The crossover chain is affected by both ancient and recent crossover events. Ancient events are controlled by the population crossover rate, $\rho = 4N_e r$, where N_e is the effective population size and r is the per-generation recombination rate. From HAPMIX, recent recombination events are modeled by considering the per generation recombination rate, r, and the estimated time since the admixture event, T. This T parameter allows the per generation recombination rate to be scaled to account only for recent events, and model specifically switches in ancestry. The ρ parameter is allowed to vary across regions tested in our model, and we use a genetic map to obtain the recombination rate r.

From the Gay *et al.*[7] model, the gene conversion chain is modeled similarly, with the frequency of gene conversion affected by the population gene conversion rate $\gamma = 4N_e g$, where g is the per-generation gene conversion rate. We the HAPMIX convention of modeling recent events using the

per-generation gene conversion rate, g and the time since admixture, T. At the same time, we model ancient gene conversion events, prior to admixture, using the γ parameter.

Parameter settings. We use two divergent reference populations, arbitrarily labeled. The first, population P_1, will be used to represent Europeans, while the second, P_2, will represent a population of African origin. We take into account the differences in effective population size, N_e, by setting $N_{e1} = 10{,}000$ and $N_{e2} = 18{,}000$. The population sizes translate to differences in the population recombination and gene conversion parameters for each population, giving ρ_1, ρ_2, γ_1, and γ_2. We use the HapMap genetic map[11] to obtain the recombination rate between pairs of sites. We set T to be 7 generations, and assume the ancestry contribution from Europeans, μ_1, is 0.2. We set the ratio between gene conversion and crossover rate, $f = g/r = 10$, based on estimates previously made using sperm typing[2] and population genetic[7] data. The expected length of a gene conversion tract, $1/\lambda$, is fixed at 500bp.

5.2.2 Transition probabilities

Following Gay *et al.*[7], the transition probabilities for crossover (X) and gene conversion (G) chains occur simultaneously. The crossover chain is independent and depends only upon the previous state within its own chain. However, it was necessary to consider gene conversion in the context of the current population of the crossover chain because our approach relies on the detection of gene conversion events that are copied from a different population from the haplotype as a whole. Therefore, we make a modification to the transition probabilities of the gene conversion chain:

$$\Pr(X_{j+1}, G_{j+1} | X_j, G_j) = \Pr(X_{j+1} | X_j) \Pr(G_{j+1} | G_j, X_j) \ . \tag{5.1}$$

For each site we consider the transition from one haplotype to another, taking into account population specific parameters for ρ, γ, μ, and n. If the destination haplotype belongs to P_1 we use ρ_1, γ_1, μ_1 and n_1; for haplotypes transitioning into P_2, ρ_2, γ_2, $\mu_2 = 1 - \mu_1$, and n_2. The population containing the haplotype at site j is labeled p_k, and the population that contains the destination haplotype, at site $j+1$, is labeled p_l.

Starting probabilities

There is an equal probability of starting the crossover chain in any given haplotype, scaled by the estimated ancestry contribution (μ) from each population to the admixed haplotype, where $x \in p_l$:

$$\Pr(X_1 = x) = \mu_l / n_l \ . \tag{5.2}$$

The gene conversion chain can start either within or outside of a gene conversion state. For the null state we consider the possibility of having ended a gene conversion tract from any of the haplotypes, and the estimated rate of gene conversion since admixture. The gene conversion chain can also start within a haplotype in each of our two reference populations, dependent on the expected rate of gene conversion since admixture and weighting the haplotypes of each population by the ancestry contribution, μ, to the admixed haplotype. Since we have no information outside this first site, we set the value of $g = fr$, where we use the estimated genome-wide recombination rate of 1.1 cM/Mb for r.

$$\Pr(G_1 = g) = \begin{cases} \dfrac{\lambda(n_1 + n_2)}{\lambda(n_1 + n_2) + gT} & \text{if } g = 0 \\ \dfrac{gT}{\lambda(n_1 + n_2) + gT} \dfrac{\mu_l}{n_l} & \text{if } g \neq 0 \text{ and } g \in p_l \end{cases} \tag{5.3}$$

Crossover transition probabilities

The probability of transitioning from a hidden state at site j to a hidden state at site $j + 1$ depends on several parameters, including the physical distance d_j (in base pairs) and recombination rate, r_j (e.g., cM/Mb). We adopt the approach used by HAPMIX to capture recent crossovers since admixture as a product of the per-generation genetic distance between markers, $r_j d_j$, and the number of generations since admixture, T. Ancient recombination is modeled using the population scaled recombination rate, ρ, scaled by the number of haplotypes in the population. Each population has its

own ρ parameter, which depends on the effective population size.

$$\Pr(X_{j+1} = x' | X_j = x) =$$
$$\begin{cases} (1 - e^{-r_j d_j T}) \frac{\mu_l}{n_l} & \text{if } x \neq x' \text{ and } p_k \neq p_l \\ e^{-r_j d_j T}(1 - e^{-\rho_{l,j} d_j / n_l}) \frac{1}{n_l} + (1 - e^{-r_j d_j T}) \frac{\mu_l}{n_l} & \text{if } x \neq x' \text{ and } p_k = p_l \\ e^{-r_j d_j T} e^{-\rho_{l,j} d_j / n_l} + e^{-r_j d_j T}(1 - e^{-\rho_{l,j} d_j / n_l}) \frac{1}{n_l} + (1 - e^{-r_j d_j T}) \frac{\mu_l}{n_l} & \text{if } x = x' \text{ and } p_k = p_l \end{cases} \tag{5.4}$$

Gene conversion transition probabilities

Here we follow the Gay *et al.*[7] model with modifications to account for the two ancestral populations and switches between them, focusing on identifying the recent events since admixture. In the case where X_j and G_j occur within the same population, we must account for gene conversions occurring both prior to, and since the admixture event. The population scaled rate, γ, is used to model gene conversion events occurring prior to admixture. For gene conversions occurring after the admixture event, we use the HAPMIX convention, as in the crossover transitions, of modeling recent gene conversion events by the per generation gene conversion rate, g, and the number of generations since admixture, T. We are specifically interested in capturing the situation in which X_j and G_j occur across different populations. In this case, we do not model ancient events, but focus on capturing only recent gene conversions.

Generally, the probability of both starting a gene conversion and ending one within the interval is taken into account to determine $\Pr(G_{j+1}|G_j, X_j)$. The rate of ending a gene conversion tract, λ, is modeled geometrically as a function of physical distance. This allows allowing termination to occur at any point within the interval and for the tract to "reset" at any time regardless of the current state. The probability of starting a gene conversion event is given by the product of g and the number of generations since admixture, T.

The gene conversion null state. In the first transition, we move from a null gene conversion state ($G_j = 0$) to another null gene conversion state ($G_{j+1} = 0$). This is given by the probability of having

no reset event and no gene conversion event in the interval, either prior to, or after admixture.

$$\Pr(G_{j+1} = 0 | G_j = 0, X_j = x) = e^{-\lambda d_j} e^{-g_j d_j T} e^{-d_j \gamma_l / n_l} + \int_0^{d_j} \lambda e^{-\lambda x} e^{-g_j x T} e^{-x \gamma_l / n_l} \, \mathrm{d}x \quad \text{if } x \in p_l \tag{5.5}$$

The integral represents the possibility that there was a reset event, but no gene conversion event.

Entering a gene conversion event. The second transition describes the probability of entering a gene conversion state from a null state. When g and x are within the same population, we must account for gene conversions events occurring prior to admixture, as well as those occurring after. We use the following constant, Z, to scale the gene conversion rate. The left term accounts for the post-admixture gene conversion rate, $g_j T$, scaled by u_l / n_l for the target population. The right term captures events occurring prior to admixture, using the population scaled gene conversion rate γ_l / n_l.

$$Z = \left[\frac{g_j T}{g_j T + \gamma_l / n_l} \mu_l + \frac{\gamma_l / n_l}{g_j T + \gamma_l / n_l} \right] \frac{1}{n_l} \tag{5.6}$$

For the transition, we consider case in which there has been no reset, but a gene conversion event has taken place prior to or after T. We also account for the case (within the integral) in which there was a reset event, with a gene conversion event taking place afterwards. The population from which the gene conversion chain now copies is taken into account within the gene conversion term using μ_l for the ancestry proportion for p_l, where $g \in p_l$:

$$\Pr(G_{j+1} = g | G_j = 0, X_j = x) = \begin{cases} e^{-\lambda d_j} (1 - e^{-g_j d_j T - d_j \gamma_l / n_l}) Z + \\ \int_0^{d_j} \lambda e^{-\lambda x} (1 - e^{-g_j x T - x \gamma_l / n_l}) Z \, \mathrm{d}x & \text{if } g \in p_l \text{ and } x \in p_l \\ e^{-\lambda d_j} (1 - e^{-g_j d_j T}) \frac{\mu_l}{n_l} + \int_0^{d_j} \lambda e^{-\lambda x} (1 - e^{-g_j x T}) \frac{\mu_l}{n_l} \, \mathrm{d}x & \text{if } g \in p_l \text{ and } x \notin p_l \end{cases} \tag{5.7}$$

Ending a gene conversion event. In the third case, we describe the probability of ending a gene conversion within the interval. As in (5.5), we are transitioning into a null gene conversion state and the population information is taken from the X chain:

$$\Pr(G_{j+1} = 0 | G_j = g, X_j = x) = \int_0^{d_j} \lambda e^{-\lambda x} e^{-g_j x T} e^{-x \gamma_l / n_l} \, \mathrm{d}x \quad \text{if } x \in p_l \tag{5.8}$$

Continuing a gene conversion event. Finally, we consider the case where we transition from a gene conversion state to another gene conversion state. In the situation where $g = g'$ we are simply continuing to copy from the same haplotype in the same gene conversion event. This is given by the probability of having no reset event, or having a reset, and then a gene conversion back to the same haplotype. When $g \neq g'$ we are transitioning from a gene conversion state in one haplotype to a different gene conversion in a different haplotype (an overlapping gene conversion). Within this interval we have a reset event followed a gene conversion event to a different haplotype, potentially in a different parental population:

$$\Pr(G_{j+1} = g' | G_j = g, X_j = x) = \begin{cases} e^{-\lambda d_j} + \displaystyle\int_0^{d_j} \lambda e^{-\lambda x}(1 - e^{-g_j x T - x\gamma_l/n_l}) Z \, \mathrm{d}x & \text{if } g = g' \text{ and } g' \in p_l \text{ and } x \in p_l \\ e^{-\lambda d_j} + \displaystyle\int_0^{d_j} \lambda e^{-\lambda x}(1 - e^{-g_j x T}) \frac{\mu_l}{n_l} \, \mathrm{d}x & \text{if } g = g' \text{ and } g' \in p_l \text{ and } x \notin p_l \\ \displaystyle\int_0^{d_j} \lambda e^{-\lambda x}(1 - e^{-g_j x T - x\gamma_l/n_l}) Z \, \mathrm{d}x & \text{if } g \neq g' \text{ and } g' \in p_l \text{ and } x \in p_l \\ \displaystyle\int_0^{d_j} \lambda e^{-\lambda x}(1 - e^{-g_j x T}) \frac{\mu_l}{n_l} \, \mathrm{d}x & \text{if } g \neq g' \text{ and } g' \in p_l \text{ and } x \notin p_l \end{cases} \tag{5.9}$$

5.2.3 Emission probabilities.

We allow for errors in the copying chain with a mutation rate defined using Watterson's estimator[8,12]. We use a separate θ for each population:

$$\theta_l = \left(\sum_{m=1}^{n_l - 1} \frac{1}{m} \right)^{-1} . \tag{5.10}$$

We then follow the Li and Stephens model to calculate the emission probability at site j for the unknown admixed haplotype, a, conditional on the underlying hidden state (X_j, G_j). We compare the haplotype from which we are copying, c, where $c = X_j$ if $G_j = 0$, else $c = G_j$:

$$e_a(j | X_j, G_j) = \begin{cases} \dfrac{\theta_l}{2(n_l + \theta_l)} & \text{if } h_{a,j} \neq h_{c,j} \text{ and } h_{c,j} \in p_l \\ \dfrac{2n_l + \theta_l}{2(n_l + \theta_l)} & \text{if } h_{a,j} = h_{c,j} \text{ and } h_{c,j} \in p_l \end{cases} . \tag{5.11}$$

5.2.4 Computational Efficiency

HMM state space. Given the large number of states in our model ($(n_1 + n_2)$ for the crossover chain, plus $(n_1 + n_2)^2$ for the gene conversion chain), increasing the number of haplotypes in the reference

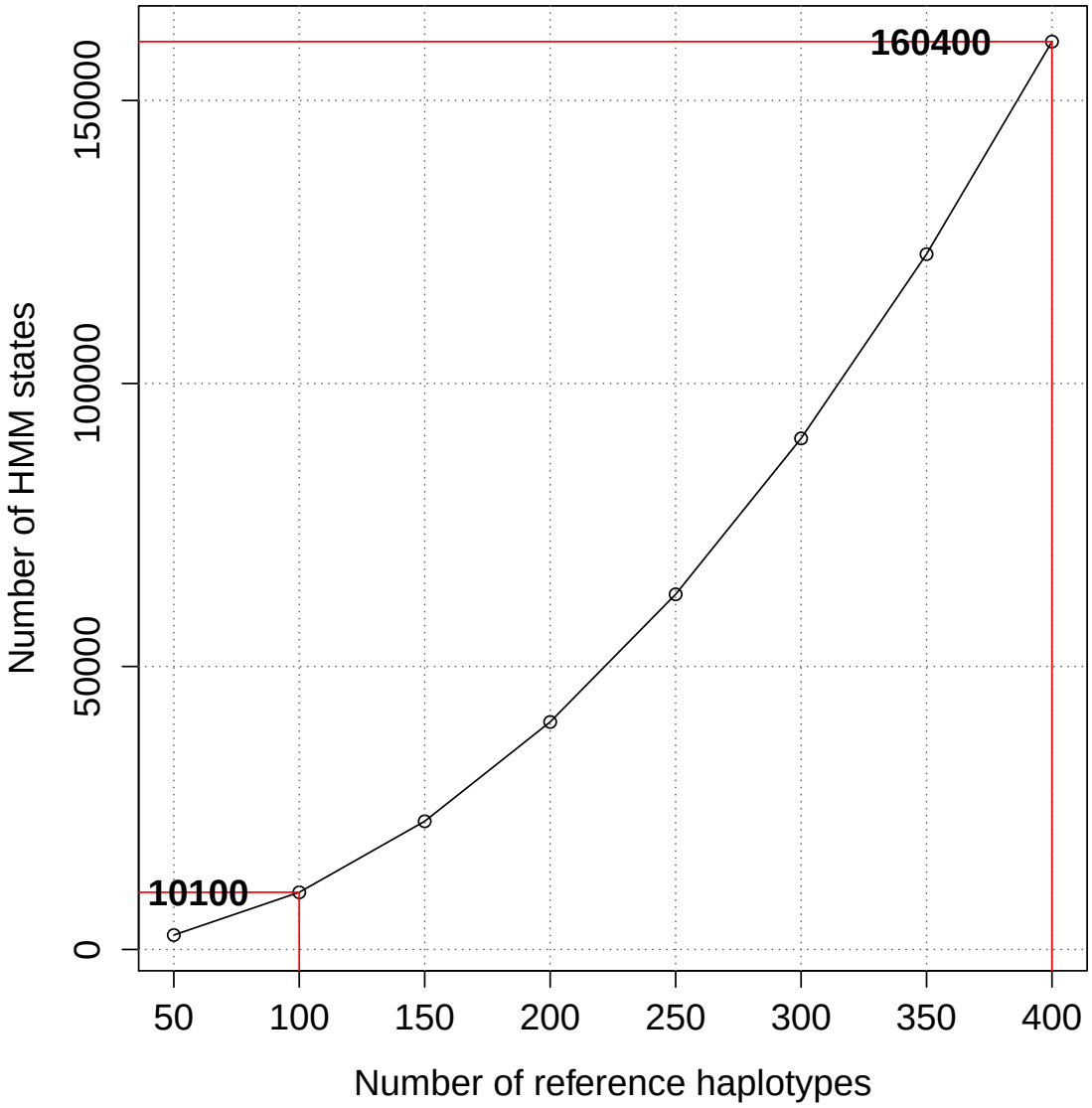

Figure 5.2: **Complexity of the gene conversion model.** The number of states is shown as a function of the total number of reference haplotypes. For a small number of haplotypes (50 in each reference population), there are 10,100 states. With 200 haplotypes in each population this increases to 160,400 states.

populations quickly produces an excess of states at each site, most with very low probability. Using just 100 reference haplotypes requires 10,100 states, while increasing to 400 haplotypes produces 160,400 states for each site (Figure 5.2). Computation using the forward-backward algorithm on even the highest performance modern processors is infeasible for even a modest number of sites. It is important to have a large enough population of reference haplotypes, in order to avoid incomplete capturing of diversity in the reference population, and provide a close enough match to each segment in the unknown admixed sample. Considering the availability of public data from the HapMap[11] and 1000 Genomes projects[13] we set a target of approximately 200 phased haplotypes in each reference population.

Two-pass approach. To reduce the computation time for this model, we apply a two-pass approach. In the first pass, we use only the crossover chain, and omit all gene conversion states from the model, reducing the number of states to $(n_1 + n_2)$. The mutation parameters for both populations, which allow for a degree of copying error in the Markov chain, are set to approximately 0 (equivalent to machine epsilon). This configuration forces the most likely state path (calculated using the Viterbi algorithm) to make a switch at each point a mismatch is detected between the copying haplotype and the unknown admixed haplotype (Figure 5.3A).

Using this no-error Viterbi path, we identify long stretches (greater than 14 sites) where the admixed haplotype matches exactly to a reference haplotype. These long stretches are likely to represent evolutionarily conserved shared haplotype segments. Since our goal is to detect gene conversion events from one ancestral population against the background of the other population, these long stretches are unlikely to contain any information of interest (and this is supported by a lack of gene conversion events in our simulation). Instead, we look for regions where the Viterbi path makes short jumps between haplotypes of different reference populations, indicating a perfectly matching stretch of haplotype was not present in the reference. These clusters of haplotype switches represent potential regions in which cross-population gene conversion events (or crossover) have disrupted linkage between neighboring sites. We select haplotype stretches in the no-error Viterbi path that are short (≤ 14 sites), and make a cross-population switch in haplotypes. For each of these stretches identified we select an additional 3 sites on each side of the stretch to capture events that may have occurred at the beginning or end.

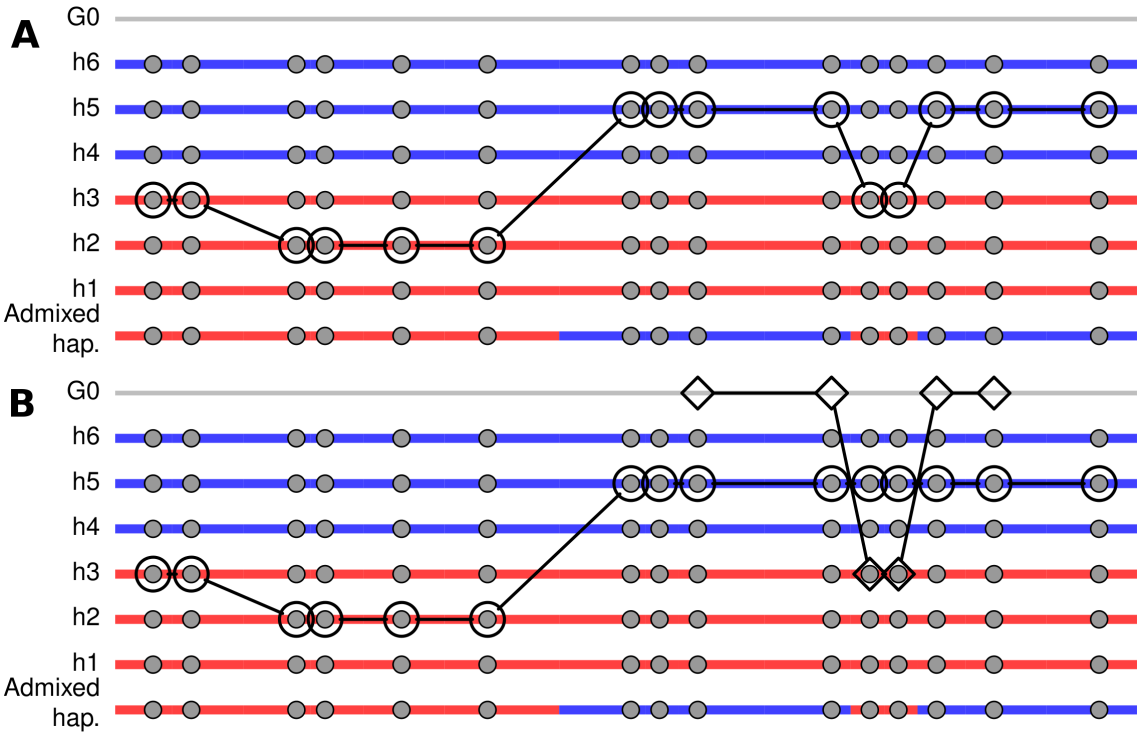

Figure 5.3: **Schematic of the two-pass HMM approach.** Each horizontal line is a haplotype, with blue and red colors designating segments that are representative of two divergent ancestral reference populations. Each column of circles represents the alleles of a SNP. The state path of haplotype transitions (X) is shown by open circles while the gene conversion state path (G) is shown in open diamonds. In **A** the gene conversion chain is not used and the crossover chain is forced to jump to $h3$ to accommodate the gene conversion. In **B** the gene conversion chain is not used except for a few sites on either side of the putative gene conversion event.

We then make a second pass through the data, making a deeper inspection of the stretches of contiguous sites marked in the first pass though the data (Figure 5.3B). Here the HMM is run using only the crossover chain (as in the first pass), except at sites marked for deeper inspection. At these sites the gene conversion chain is turned on and we allow all possible transitions under the full HMM. This technique allows a drastic reduction in the model complexity at sites that are unlikely to contain a gene conversion event (or at least those that are able to be detected by our model).

Further reduction in computation time is achieved by running the second pass on smaller subsamples taken from the total body of reference haplotypes. For example, with a group of 200 haplotypes in each reference population, we draw 10 random samples of 50 haplotypes from each. We then run the reduced complexity model on the 10 subsets. Finally, the results from each of the subsets are combined by taking the mean posterior probability of a cross-population gene conversion event. This allows a drastic reduction in the number of total states, and thus a much faster run time.

5.2.5 Simulation

In order to test the ability of the model to identify gene conversion events, we first evaluate its performance using simulated data where the true GC events are known. Data from the 1000 Genomes Project (Phase III[13]) serves as source data from which we draw haplotypes for our parental populations.

Since our model attempts to infer recombination rates, the use of data 1000 Genomes data that was assembled and phased using a genetic map (the HapMap phase II map[11]) would introduce a potential bias. In order to avoid this bias, we first re-phase the data. We use SHAPEIT[14] to construct phased haplotypes from genotype data. SHAPEIT requires a genetic map as a prior input for phasing, so we create a map with the same intervals as the standard HapMap genetic map, but with the recombination rate at each interval set to 1.1 cM/Mb. In addition, our method uses a genetic map, with estimates of the recombination rate at each site, to provide values for parameters for the per-generation crossover rate, r, and by extension, the gene conversion rate, g (which is equal to fr). As this also has the potential to introduce bias, we again use a genetic map with a fixed recombination rate.

To create simulated admixed haplotypes, we use a 1 Mb region on chr22, taking all phased haplotypes from two groups of reference populations. For European haplotypes we select CEU, GWR,

and TSI populations (n=399), and we select YRI and LWK for African populations (n=302). We draw 100 haplotypes from each of two reference populations. SNP density is thinned to keep 300 sites, and singleton variants are removed from the analysis.

An initial point admixture is then simulated for 50 haplotypes by drawing one haplotype from each reference population and creating a recombinant "offspring" haplotype. The number of crossovers is determined by sampling from a Poisson distribution ($r = 1/\text{Mb}$). Gene conversions are likewise placed under a Poisson process ($g = 10/1\text{Mb}$; $1/\lambda = 500\text{bp}$). Following the initial admixture event, six subsequent generations were simulated using intra-population mating, with CO and GC events recorded each time they occurred. Our HMM is then run using the haplotypes from the simulated admixed population using the two-pass approach.

5.2.6 Code availability.

The C++ code for this model is available at `https://github.com/clcampbell/CGmix`.

5.3 Results

We test the performance of our model using simulated genomes on real human data from the 1000 Genomes Project[13]. Using a small sub-section of chromosome 22, we select parent haplotypes from CEU and YRI populations, representing individuals of European and African descent, respectively. Recombination crossovers are placed with a rate of 1 / Mb, and gene conversions at 10 times this rate (10 / Mb). We use a mixing proportion, μ = 0.5, instead of the 0.8 value estimated from previous methods[9]. The location and type of each gene conversion is recorded for evaluation purposes.

Gene conversions can be classified into one of three types, based upon how our model will be able to detect them. We are unable to detect GCs that do not overlap a SNP, and label these "invisible" GC events. GCs that do overlap a heterozygous SNP, and do not produce any change in the alleles are labelled "silent" GC events. Finally, GC events that overlap a SNP and produce a detectable change in the allele are termed "non-silent" events, and are the only ones that are detectable by our model. Therefore, our results only account for non-silent gene conversion events, and we do not consider the other cases, even though they may occur as a result of the simulation.

For all GC events detected, we record the total posterior probability of finding a GC event at that site, and whether this is a real (non-silent) GC event. Using a threshold of 0.5 for the posterior probability, we estimate the true positive rate at 8.0%, while the false discovery rate is 27.3%.

5.4 Discussion

The detection of recombination within population genetic data using LD based approaches has been used in a number of studies in humans, yielding valuable results[11,15,16]. A number of methods have been based on a model described by Li and Stephens[8] in 2003, in which haplotypes are modeled as a mosaic of segments from other haplotypes. Several extensions from this method have produced valuable models to detect gene conversion[7], and accurately infer admixture breakpoints[9,10]. The admixture approach to recombination is a promising one, especially considering the expanding availability of data on a number of human populations of diverse origins, many of them admixed.

The detection of gene conversion events has long been a difficult problem, owing to their small size, and the confounding effects of genotyping error. Most studies focus on crossover and ignore the effects of gene conversion, which provides an incomplete picture of the recombination process as a whole. In our model, the use of admixed genomes provides a unique leverage, in which we detect gene conversion events from one ancestral population against a contrasting background from the other. In our model, we adopt the Li and Stephens model to include gene conversion, as in Gay *et al.*[7], and two reference populations, as in HAPMIX[9]. This resulting model attempts to identify cross-population gene conversion events with a higher contrast.

The confounding effects of genotyping error and the small size of gene conversion events therefore constrain the sensitivity of the model substantially, which we report here to be 8%. This low sensitivity is expected, due to the nature of the problem, and this model can still be considered to be successful even when missing the vast majority of true events. The detection of even a few hundred gene conversions in human data would represent a substantial advance in our knowledge. The FDR of the model is far more important, and lowering this rate to under 10% would be a key advance that would be necessary before the model can be run on admixed samples with any reliability. Future work could focus on simplifying the model, and reducing the FDR.

While the complexity of this model is quite high, we have made several advances that allow

the computation time to be brought down. First, we have reduced the state space by dividing the reference populations into ten random subsets. The model was run on these subsets, and the results combined to generate an approximation to the model under the full set of reference haplotypes. In addition, we built in the ability to run our model in two parts. The first pass runs quickly, using only the crossover chain, and screens for potential breakpoints. In the second pass, the full model with both CO and GC chains is used, but only on a subset of the sites that had been previously marked for deeper inspection. These advances can potentially be generalized to other HMMs of similar structure in the future. Despite these advances, the complexity is such that it is impractical to run this data on full genomes without significant advances in computing efficiency. In addition inference using this model, while theoretically possible, is not practical. Other models use maximum likelihood methods to estimate optimal parameter settings. For example, in the Li and Stephens model, this technique was used to estimate $\hat{\rho}$ from human data, providing approximations to the recombination rate across the genome.

This model provides a method to detect recombination and gene conversion in admixed population genetic data. While the complexity of this model is high, it has the potential to be used in human data, especially as the availability of admixed data increases. Combined with an increase in computing efficiency, or a redesign of the model, this admixture approach has promise for future work to further characterize gene conversion in the human genome.

5.5 References

1. Baudat, F. and de Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Research* 15(5):565–77 (2007). doi: 10.1007/s10577-007-1140-3.

2. Jeffreys, A. J. and May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics* 36(2):151–6 (2004). doi:10.1038/ng1287.

3. Cole, F., Keeney, S., and Jasin, M. Preaching about the converted: how meiotic gene conversion influences genomic diversity. *Annals of the New York Academy of Sciences* 1267:95–102 (2012). doi:10.1111/j.1749-6632.2012.06595.x.

4. Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* 4:e04637 (2015). doi:10.7554/eLife.04637.

5. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics* 8(10):762–75 (2007). doi:10.1038/nrg2193.

6. Bhérer, C. and Auton, A. Biased Gene Conversion and Its Impact on Genome Evolution (2014). doi:doi:10.1002/9780470015902.a0020834.pub2.

7. Gay, J., Myers, S., and McVean, G. Estimating meiotic gene conversion rates from population genetic data. *Genetics* 177(2):881–94 (2007). doi:10.1534/genetics.107.078907.

8. Li, N. and Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165(4):2213–33 (2003).

9. Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* 5(6):e1000519 (2009). doi:10.1371/journal.pgen.1000519.

10. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., *et al.* The landscape of recombination in African Americans. *Nature* 476(7359):170–5 (2011). doi:10.1038/nature10336.

11. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61 (2007). doi:10.1038/nature06258.

12. Watterson, G. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7(2):256–276 (1975). doi:10.1016/0040-5809(75)90020-9.

13. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526(7571):68–74 (2015). doi:10.1038/nature15393.

14. Delaneau, O., Zagury, J.-F., and Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* 10(1):5–6 (2013). doi:10.1038/nmeth.2307.

15. McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* 304(5670):581–4 (2004).

doi:10.1126/science.1092500.

16. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310(5746):321–4 (2005). doi: 10.1126/science.1117196.

Chapter 6

Discussion

In my thesis, I have presented pedigree analyses of meiotic recombination in both humans and dogs. The 23andMe analysis in humans represents a comprehensive study with a large number of meioses, adding to existing research to further characterize recombination. This study examines sexual dimorphism in recombination, specifically evaluating overall rate distribution, as well as regulatory mechanisms of hotspot usage, and crossover interference. A major advance from this study was to further characterize maternal age effects on recombination. These age effects were found to manifest across several aspects of recombination placement, contributing to an overall pattern of deregulation with increased age.

Recombination is not nearly as well characterized in dogs. However, dogs present a unique species in which to study recombination, largely due to the multiple truncating mutations that have rendered PRDM9 inactive millions of years ago. The loss of this key recombination regulatory mechanism raises important questions on the role of recombination and specifically how the dog genome has been shaped by recombination in the absence of PRDM9. The pedigree analysis presented in Chapter 4 represents an important advance towards this goal. Here, I have shown that dog recombination is similar to humans on a broad scale, with high rates at the telomeric ends being male-driven. At the fine scale, the lack of PRDM9 is evident in the concentration of recombination at gene promoters, with sex differences that differ from those of humans in some aspects.

In addition, I have developed a statistical model for the detection of gene conversion in human admixed population genetic data. This model integrates key features of two previous models. First it leverages the divergence between two distinct reference populations to pick out gene conversion events in admixed population genetic data. Second, it models gene conversion and crossover simultaneously using two different Markov chains. Although it is computationally intensive, this model has the potential to advance our understanding of gene conversion and its role within the recombination process.

6.1 Sex dimorphism in recombination

6.1.1 Heterochiasmy

Heterochiasmy, the difference in recombination rate between the sexes, has been observed in a wide variety of species over the course of decades of studies. In most studied species, the recombination

rate is higher in females, and a number of possible explanations exist to explain this. Many studies suggest that heterochiasmy is tied to the biological differences in meiosis between the sexes. Theoretical explanations suggest that natural selection could act to modify recombination rate in either the haploid or diploid life stages. Since diploid selection is presumably stronger in males, this would contribute to a reduced recombination rate to keep favorable haplotypes in successful males[1]. Selection at the haploid stage is presumed to be restricted to males, since fertilization marks the completion of female meiosis, and there is essentially no haploid phase[2]. Another possibility is that meiotic drive plays a role in driving evolution of the female recombination rate. That is, alleles that have an increased probability to be transmitted in oogenesis are more likely to exert selective pressures that modify the female recombination rate[3].

While unable to address these hypotheses directly, the pedigree data presented in this thesis adds a further data point to the ratio of female to male recombination in humans and dogs; two species separated by millions of years of evolutionary divergence. In humans, the consensus among many studies over the past decades is that this ratio is approximately 1.6 within the autosomes. This puts humans among the most heterochiasmate species currently known, although there are a number of studies in amphibians and fish that report much higher ratios (Table 1.2).

In dogs, I found the ratio to be considerably lower, at 1.2 (Chapter 4), which matches the ratio reported in a previous dog pedigree study[4], both of which studied recombination in purebred dogs. This indicates that the level of heterochiasmy is not as strong in dogs as it is in humans. Inbred dogs have been subject to strong artificial selection since the creation of modern breeds, and artificial selection has been suggested to have a strong effect on recombination rates. This suggestion came from a study in domestic cattle, in which a decline in recombination rate was seen over time in response to artificial selection[5]. Considering dogs have been subject to extreme artificial selection as well, selection may have played a role in the modification of dog recombination.

6.1.2 Hotspot overlap.

PRDM9 has been shown to undergo rapid evolution within its DNA-binding zinc finger array[6,7], and this evolution underlies the lack of sharing between human populations[8], and between species[9]. Previous work in humans found no difference in hotspot usage between male and females[10], with the caveat that a relatively small number of meioses was used. In this thesis, I report for the first time

that males have a higher hotspot usage than females, with a difference of 4.6%. This difference is not due to the position of hotspots within the genome, nor to recombination rate differences between males and females.

One possible explanation for this is that the set of reference hotspots used was generated from LD studies of recombination[11,12]. They are therefore a sex-averaged collection, and may be biased in representation of hotspots from males and females. Given that females have a greater number of dimorphic regions of recombination than males (Appendix A, Bhérer *et al.* [13]), it is plausible that there are a greater number of female hotspots, resulting in a more diffuse hotspot usage per individual. While it is currently not possible to answer this question, future expansions to pedigree studies will provide the possibility of generating a collection of unbiased hotspots independently for males and females.

6.1.3 Recombination around the TSS

In humans, it was previously estimated that the recombination rate increases near the transcription start site (TSS), drops sharply within gene regions, and increases again after the transcription termination site[11,12,14–16]. In dogs, the recombination rate exhibits a sharp peak just upstream of the TSS, in gene promoter regions[17]. Data presented in this thesis has the potential to expand these observations. There appear to be sex differences in the recombination rate around gene regions, particularly the TSS, in both humans and dogs.

In humans, female recombination has a sharp peak centered around the TSS, that extends approximately 10 kb on either side, while the male rate appears flat throughout the region (Figure A.2A). When removing genes that have PRDM9 binding motifs within 5 kb of the TSS, the female peak is no longer seen (Figure A.2B), suggesting that this rate increase is somehow driven by PRDM9 binding. Although the dog data is of lower resolution, it appears to show the opposite effect. Here, the male rate appears higher, both in the small peak just upstream of the TSS, and in the surrounding region (Figure 4.S11A).

When removing regions that are PRDM9 influenced from the human data, the recombination rate surrounding the TSS looks similar to that of dogs. The reason for these differences is not known, but it is possible that human crossovers that are not governed by PRDM9 act similarly to those of

PRDM9-absent species. Recombination in the absence of PRDM9 has been suggested to locate preferentially to regions of open chromatin, which includes gene promoter regions[17–19].

6.2 Crossover interference

6.2.1 Interference on an individual basis.

Most studies in humans so far have looked at interference on a group basis. This is a necessity of the methods used to study interference, which require that a model be fit to a distribution of many inter-crossover distance measurements. The level of variability in interference within single individuals is currently unknown.

Using publicly available data, from sperm[20,21], and oocytes[22], I have taken the first steps toward addressing this question (Chapter 3). These data indicate that the level of interference varies widely both between, and within individuals. While these results represent a small number of samples (2 males, 8 females), the increasing availability of genetic data from single-cells means that it will soon become feasible to study interference on an individual basis.

6.2.2 Interference parameters across the human genome

Whether the strength of crossover interference varies with chromosome size has been an open question, with conflicting reports on this relationship. In a cytological study in human males, Lian *et al.*[23] suggested that interference strength is higher in smaller chromosomes. This study used the simple gamma distribution, assuming one class of crossovers which are all interfering. However, a reanalysis of this data was performed using the two pathway gamma-escape model[24]. Here, the researchers argued that the gamma model was inappropriate for fitting to immunofluorescence data since the original study considered only cases in which there are two or more MLH1 foci per stained chromosome. The results of the reanalysis under the two pathway model indicated that interference strength remained constant and independent of chromosome size[24].

Using the data from the 23andMe cohort, I have shown that interference strength does depend on chromosome size under the two pathway model in both males and females (Figure 2.S8). Here, smaller chromosomes (measured using genetic map length) had a greater strength of interference

when compared to larger chromosomes. This supports the findings of Lian *et al.*[23], and makes intuitive sense, since two crossovers placed on a large chromosome will naturally be further apart that on a smaller chromosome.

When considering the proportion of events that escape interference, there was a high level of variability among the chromosomes, and no relationship with chromosome size. We are unable to make any conclusions along these lines in dogs, due to a reduced dataset size, and lower power to detect interference for smaller chromosomes.

6.2.3 Implications of the two-pathway model in humans and dogs

The two pathway model categorizes crossovers into two classes: those that are subject to strong positive interference, and those with no interference. A suggestion from Housworth and Stahl[25] was that these two pathways could correspond to crossovers placed by different mechanisms that are temporally separated. Non-interfering events were predicted to occur early in meiosis, and aided in the pairing and synapsis of homologues into the synaptonemal complex (SC). Events that occurred later were part of a disjunction pathway, and subject to strong interference that spaced out the events along each chromosome. The wider spacing of crossovers likely assists in proper disjunction, reducing the risk of aneuploidy. This is suggestive of mechanistic differences that define the two pathways, in which different mechanisms control DSB initiation and placement, and DSB repair mechanisms influence the decision to repair a break as a crossover or gene conversion[26–28].

Substantial evidence has accumulated in support of the two pathway model for recombination in humans[25,29,30]. While there is strong evidence that crossovers are indeed comprised of these two classes, whether these classes correspond to early/pairing and late/disjunction is unknown. An additional confounding factor is PRDM9, which acts to initiate DSBs in the meiotic process. How or if PRDM9 fits into the framework of the two pathway model is not known. However, PRDM9 is known to act early, generating DSBs in the zygotene phase[31]. This early action of PRDM9, along with its localization to specific DNA sequence motifs, suggests that it could be part of the early pairing pathway, and would therefore produce crossovers that are independent of interference.

The dog pedigree study provides an excellent opportunity to examine crossover interference in the absence of PRDM9. I found evidence for strong positive crossover interference in dogs, supporting previous cytological data[32], and the two pathway model was favored over the simple gamma model.

This adds to existing findings from this study and others[4,17,33] that dog recombination is broadly similar to that of humans, apes, and mice. While the sample size of the dog study makes a firm conclusion difficult, the support for the two pathway model of interference in dogs would suggest that PRDM9 is likely not involved with interference. This supports the idea that PRDM9 is part of an early acting, non-interfering recombination initiation pathway.

6.3 Age effects on recombination

There have been conflicting reports of age effects on human recombination in the past decades. A number of studies reported that the maternal recombination rate increases with age[10,34], while others have reported the opposite[35,36]. In most cases, the reported effect is subtle, with only a few extra crossovers found over a 10 year difference in age. In all studies, the effect was limited to females, with no change in rate found in males.

In the 23andMe study (Chapter 2), I present evidence that the maternal recombination rate increases with age, with a sharp increase in the recombination rate found in mothers over 39 years of age. Since the publication of our findings from the 23andMe data more evidence has surfaced that reinforces this. A multi-cohort analysis using over 6,000 meioses found a small and significant positive association with recombination rate and age[37]. In this study, six cohorts were examined using multiple statistical methods, lending convincing support to the existence of this effect.

There are additional age effects on recombination, and I have shown, for the first time, that crossover interference parameters change with age. While there is no change in interference strength, the proportion of events that escape regulation by interference rises sharply in older mothers. This effect is robust to varying divisions of the data, and further data from an additional older age group shows that the escape proportion continues to increase in a linear fashion (Chapter 3). In addition, there is a weak effect on hotspot usage with age in females, with hotspot usage decreasing slightly.

Several interpretations for the rate increase in older mothers have been put forth. The finding of these age related effects, which include an increase in crossover count and an increase in escape, along with higher rates of aneuploidy in older mothers[38], suggests that all of these observations may be related to similar underlying mechanisms. When taking interference into account alongside the apparent increase in recombination rate, there are several possibilities. First, it is possible that

mothers with a baseline higher recombination rate are able to continue to have healthy children until a later age[34]. However, the increased clustering of crossovers that are seen as a result of interference escape is not clearly explained by this model.

Second is the possibility that oocytes enter and exit meiosis in a specific order, and that oocytes that exit early have stronger and more frequent recombination than those that exit later. However, a recent study presented evidence against the existence of a production line in humans, finding no differences in recombination rate among fetal oocytes[39]. Whether interference could affect oocytes differently across a production line while there are no differences in rate is not known.

It is also possible that some form of deregulation is occurring during the meiotic arrest period, which can last for decades. The arrest occurs after the full assembly of the synaptonemal complex, while the chromosomes are frozen mid-crossover. Over time, the connections tethering the complex together have been shown to degrade[40], which can in turn lead to aneuploidy during disjunction later in meiosis. However, in this model, interference has presumably already exerted its effect, since crossovers are placed prior to arrest.

The concept of deregulation is an attractive one, since it is apparent that the positioning of crossovers within the genome is carefully regulated by multiple mechanisms. The preferential positioning of crossovers within hotspots by PRDM9, the spacing of crossovers by interference, broad-scale positioning, and crossover assurance and homeostasis can all be viewed in the context of regulatory constraints on recombination. Taken together, these regulatory mechanisms exhibit careful control over broad and fine scale crossover placement, while still retaining some element of randomness. Differences in these elements are apparent between males and females, with males having a higher hotspot usage, and a stronger interference strength, despite a lower number of crossovers overall. With respect to age, males show no change in the distribution of their events, while data for females suggests changes in interference, recombination rate, and potentially hotspot usage as a function of age. This evidence suggests that female recombination is subject to a deregulation phenomenon that increases in strength with age and it is possible that this is caused by an interaction or conflict between crossover regulating mechanisms.

6.4 Proposed model for recombination initiation and resolution

A growing body of work has led to the characterization of recombination in an expanding number of diverse species, including many without PRDM9. From a LD based study in dogs, it was found that recombination preferentially locates to gene promoter regions, located just upstream of the TSS[17]. A similar finding was made in mice in which PRDM9 had been inactivated[41]. Here, crossovers did not properly resolve, but DSBs instead located preferentially to H3K4 marks located upstream of the TSS. Work in other species lacking PRDM9 has found similar trends in birds[18], and yeast[19,42]. This body of work suggests a pattern in which, without PRDM9, recombination is directed by default into regions of open chromatin, which frequently encompass gene promoter regions. These regions are easily accessible by recombination initiation machinery that may provide an easy target for DSBs.

In species that depend on it, PRDM9 is known to act on a fine scale to direct recombination into narrow hotspots. It is possible that there are a proportion of crossovers in humans that may be independent of PRDM9, and this is supported by the finding of a healthy mother without PRDM9[43]. In addition, hotspot overlap varies considerably on an individual basis[10,30]. Therefore it is possible that crossovers are placed with both PRDM9 dependent and independent mechanisms, but it is not known what proportion belong to each category. An attractive explanation is to merge these crossover categories with those in the two pathway interference model. The support for the two pathway model of crossover interference in dogs (Chapter 4) in the absence of PRDM9 suggests that PRDM9 is not involved in interference. This therefore suggests that PRDM9 is part of an early crossover pathway that is independent of interference. A second category of crossovers would be placed later, and subject to strong interference. The results showing that human recombination appears to become deregulated with age in females supports the idea of two recombination pathways that are under differential control in males and females. These raises the possibility that one class of crossovers may act as a backup method to ensure disjunction, and that this class is responsible for disjunction errors in older mothers.

6.5 Strengths of this work

Pedigree studies are powerful methods for studying the crossover patterns in recombinant gametes across a single generation. They allow the assignment of events to a specific individual, and therefore allow us to study how recombination differs between males and females.

Considering the data presented within this thesis, one major strength is the sample size for the human recombination study in Chapter 2. Here, I used data obtained from a collaboration with 23andMe to analyze more 18,000 meioses in one of the largest pedigree studies conducted in humans[30]. This large sample size allowed a detailed analysis to be made of human recombination in males and females. Perhaps most importantly, the number of individuals was high enough to allow the division of the data into multiple age groups. After this division, each sub-group had a high enough number of samples to allow a clear trend to be discovered, both in the increase in recombination rate, and the amount of interference escape, both of which were found to increase with age in females.

The 23andMe data proved to be of very high quality and relatively minimal quality control steps were necessary to generate genetic maps that closely resembled previous high quality maps from the deCODE study[16], and the Hapmap project[12]. In addition, the reference assembly for the human genome is well characterized and includes very few gaps or misplaced contigs, which could result in false crossover calls and an inflation of the genetic map. Therefore, most of the quality control process focused on the pedigree data itself, and the removal of families or meioses with biologically implausible numbers of crossovers.

Recombination fraction correction for homozygosity. The recombination fraction between markers, θ, is represented as the ratio of the number of recombinants in a given interval to the number of meioses studied, and determines the recombination rate:

$$\theta = \frac{\text{\# of recombinants}}{\text{\# of meioses studied}}.$$

The number of meioses is usually a fixed number reflecting the composition of the dataset. This is sufficient for typical datasets, and has been used for decades in pedigree studies in humans and other species.

However, when working with inbred dogs, I discovered that the heterozygosity of the samples presented an issue. Crossover events are represented as a genomic interval that is flanked by het-

erozygous markers in the parent. In a number of the parents, I found a lack of heterozygous markers, especially towards the telomeric ends of the chromosomes. In some cases, the first heterozygous marker did not occur until many Megabases into the chromosome. Therefore, it became obvious that our study was missing crossovers occurring toward the telomeres, and this was reflected in the genetic maps. When comparing to the previous pedigree study from Wong *et al.* [4], our dog maps had a consistently shorter map length.

Therefore, some correction was necessary. I changed the denominator of the recombination fraction to reflect the effective number of recombinants within each specific interval. This change to the recombination fraction resulted in map lengths that were much closer to those of previous studies. This modification to genetic map constriction should prove useful in future pedigree studies, even those that do not study inbred populations.

6.6 Limitations of this work and alternative approaches

6.6.1 Limitations of pedigree studies

Pedigree studies, by their nature, are inferential studies of recombination and are limited to observe only the transmitted gametes that produce viable offspring. By contrast, a direct approach would be to observe a cell across the entire meiotic cycle, capturing events as they occur, and be able to analyze all four products of meiosis.

Data availability. At the same time, large sample sizes are required for a comprehensive analysis of the results from a pedigree analysis. Given that there are only 20-60 events in a given (human) meiosis [10,16,44], large numbers of meioses are necessary. Generating data of this size is difficult, both in terms of cost, and in sample availability, especially in non-humans. In addition, pedigree studies often rely on accurate genealogical and historical records, which are unique to humans.

Unfortunately, the requirement for large sample sizes is firm and the same type of data cannot be obtained through other methods. However, Kong *et al.* [16] took a novel approach to bypass the sample size requirement imposed by this method. Leveraging the high degree of relatedness within Icelanders, phase was inferred via computational methods, allowing crossover identification with only a subset of individuals fully genotyped. The effective study size was increased well beyond the num-

ber of fully genotyped individuals, but this increase in sample size came with a penalty. The methods used meant that the 5 Mb of sequence near the telomeric ends of the chromosomes were unable to be accurately characterized, where male recombination is typically higher.

In recent years, the cost of whole-genome genotyping methods, including SNP microarrays and even whole-genome sequencing, have continuously reduced in cost. With this reduction in cost, and increasing availability of data, pedigree studies will become more widespread. The data generated from these studies will reveal more about the recombination landscape in humans, and in presently uncharacterized species.

6.6.2 Cohort composition

Humans. The 23andMe cohort consisted of approximately 70% of samples that were of European descent, which limits the conclusions that could be drawn for populations outside of Europe. This limitation can be directly seen in the analysis of hotspot usage across populations in Figure 2.2B, in which the error bars for non-European populations are much wider that those for Europeans. Multiple studies have demonstrated differing properties of recombination in non-European populations[8,36,45], highlighting the importance of expanding these studies to other populations.

Dogs. Within this thesis, the study of dog recombination (Chapter 4) in particular was hampered by a low sample size. In addition, the dog genome is not as well characterized, and the reference assembly consists of larger, and more frequent gaps and misplaced contigs. This was made especially apparent when, having been first incredibly fortunate with the excellent 23andMe dataset, I transitioned to the pedigree analysis in inbred dogs. The genotype data itself was of high quality, however the inbred nature of the dogs used made the inference of crossover locations difficult, with a number of false positives. These factors made for an especially difficult quality control process, with many iterations necessary to remove poorly assembled regions that caused severe jumps in the genetic map.

In addition, the small sample size was limiting for many of the analyses attempted including a test for any age effect. The resolution of the genetic map was not high enough to examine fine scale differences between males and females in depth. Additional work with a larger cohort is necessary to sufficiently address these questions. This could be possible through an increased availability of

public genotyping data, for instance, or through collaboration with a research institution with existing data of sufficient size.

6.7 Future directions

The results presented within this thesis represent an advance in the characterization of recombination in both humans and dogs. At the same time, they raise a number of further research topics that need to be addressed.

First, it is necessary to expand further the datasets available to study recombination in humans, as well as other species. Successful pedigree studies depend on the availability of sufficient number of meioses. As the cost of genotyping large numbers of individuals and families further decreases, it should be possible to extend pedigree analysis further. An interesting example of this can be seen in this thesis, in which a private company, 23andMe, has been very successful at collecting large amounts of genotyping data. While the 23andMe data was not originally intended to be used for recombination studies, their Research Portal[46] allows investigators to apply to use the data for potentially unconventional purposes. Resources such as the 23andMe sample database will continue to grow, providing a reduction to some of the barriers in the collection of data. It will prove interesting to revisit the 23andMe dataset as it continues to expand into the future.

With the availability of increased amounts of data comes the opportunity for a more even sampling of the human population as a whole. Most recombination studies focus on, or are limited to, a single human population which are often genetic isolates such as Hutterites[10], Icelanders[16], or French-Canadians[35]. The 23andMe study itself has a substantial European bias. While these studies undoubtedly provide valuable data, it would be interesting to undertake a large-scale study comparing recombination across multiple worldwide populations.

It is also of great importance to further expand recombination studies to more non-humans species. Chimpanzees, our closest ancestor, provide an important potential venue to study recombination from an evolutionary perspective. Currently there is a LD map of recombination in chimpanzees[9], however no pedigree studies have yet been completed. The extent of sex differences within chimpanzees, and how they compare to those observed in humans, remains unknown and could provide important clues to recombination in humans and how it has evolved since the time of divergence.

Expanding pedigree studies to look for sex differences in PRDM9-absent species has the potential to reveal much about the regulatory mechanisms that govern crossover placement. The study of inbred dogs presented in Chapter 4 provides an important first step towards this, but a greater sample size and a greater breed diversity is necessary for further research. A comparison of domestic dogs to village dogs, as well as to wolves, their closest ancestor, could provide important clues in the evolution of recombination in the absence of PRDM9. Canids have lost PRDM9 through mutation, and we can assume therefore that it must have been previously functional. It would therefore be of great interest to compare canid recombination to that of the Giant Panda, their closest PRDM9-dependent relative. Such a comparison could reveal how the recombination processes in these two species have diverged over millions of years of evolutionary history.

Further investigation into the phenomenon of crossover interference has the potential to reveal much about the recombination process. The 23andMe study is the largest interference study so far in humans, but the conclusions on interference are made by considering all samples on a group basis. The continuing progress in single cell sequencing, both in spermatocytes and oocytes, provides more opportunity to study large numbers of recombination products from a single individual. I have demonstrated this on a small scale in Chapter 3, finding that there is substantial variability within and between individuals. In addition, using data from DSB initiation maps[47], interference can be studied on the DSB level. Future research at both the DSB and crossover level, using larger sample sizes, has the potential to reveal mechanisms of interference, and on which meiotic stage they act.

The human work presented here on the maternal age effect has potential clinical significance. The incidence of aneuploidy has a baseline of around 1-4% in males and 2-3% in females in their twenties, but this number rises sharply, to 30-40% in females in their forties[40,48]. Currently such aneuploidies are not preventable, although several clinical screening procedures are available, and more are continually evolving. Recently, the analysis of circulating fetal “cell-free DNA” in maternal blood has been used in the development of a non-invasive assay for the detection of fetal aneuploidy[49]. This technique has been used with great success to detect a number of trisomies including 21[50], 18[51], and 13[51]. However, even with the success of the cell-free DNA approach, positive results are often followed up with invasive secondary tests, and this is an area in which more work is necessary.

There is a clear connection between failures of recombination at meiosis I, and aneuploidies that result in an extra or missing chromosome[40]. The work presented in Chapters 2 and 3 extends these

findings and has the potential to improve understanding of the underlying mechanisms. The maternal age related findings of an increase in recombination rate and interference escape point to a possible decline in recombination regulation mechanisms. These mechanisms are presumably in place to prevent non-disjunction and ensure healthy gametes. This research has the potential to further our understanding of how the recombination process changes with age in females, which can in turn inform future diagnostic screening approaches.

In conclusion, the study of human recombination has advanced substantially in the years since the completion of the Human Genome Project. The data presented within this thesis suggests a complex interaction of regulatory mechanisms that appear to be subject to deregulation in older mothers, suggestive of a connection to aneuploid pregnancies. While the study of recombination in other species lag behind humans, this gap is being continually eroded, with the advancement of genotyping technology and the collection of new data both contributing to new findings. The further study of PRDM9-absent species, including dogs, has the opportunity to increase the level of understanding in these species as well as humans. At the same time, more research is necessary to further characterize gene conversion and provide a more complete picture of the recombination landscape as a whole.

6.8 References

1. Trivers, R. Sex differences in rates of recombination and sexual selection. In R. Michod and B. Levin, editors, *The evolution of sex*, pages 270–286. Sinauer Press, Sunderland, Massachusetts (1988).

2. Lenormand, T. and Dutheil, J. Recombination difference between sexes: a role for haploid selection. *PLoS biology* 3(3):e63 (2005). doi:10.1371/journal.pbio.0030063.

3. Brandvain, Y. and Coop, G. Scrambling eggs: meiotic drive and the evolution of female recombination rates. *Genetics* 190(2):709–23 (2012). doi:10.1534/genetics.111.136721.

4. Wong, A. K., Ruhe, A. L., Dumont, B. L., Robertson, K. R., Guerrero, G., *et al.* A comprehensive linkage map of the dog genome. *Genetics* 184(2):595–605 (2010). doi:10.1534/genetics.109.106831.

5. Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., *et al.* Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLoS Genetics* 11(11):e1005387 (2015). doi:10.1371/journal.pgen.1005387.

6. Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., *et al.* Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genetics* 5(12):e1000753 (2009). doi:10.1371/journal.pgen.1000753.

7. Ponting, C. P. What are the genomic drivers of the rapid evolution of PRDM9? *Trends in Genetics* 27(5):165–71 (2011). doi:10.1016/j.tig.2011.02.001.

8. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., *et al.* The landscape of recombination in African Americans. *Nature* 476(7359):170–5 (2011). doi:10.1038/nature10336.

9. Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., *et al.* A fine-scale chimpanzee genetic map from population sequencing. *Science* 336(6078):193–8 (2012). doi:10.1126/science.1216872.

10. Coop, G., Wen, X., Ober, C., Pritchard, J. K., and Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319(5868):1395–8 (2008). doi:10.1126/science.1151851.

11. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310(5746):321–4 (2005). doi:10.1126/science.1117196.

12. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61 (2007). doi:10.1038/nature06258.

13. Bhérer, C., Campbell, C. L., and Auton, A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Unpublished - under review* (2016).

14. McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* 304(5670):581–4 (2004). doi:10.1126/science.1092500.

15. Spencer, C. C. A., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., *et al.* The influence of recombination on human genetic diversity. *PLoS Genetics* 2(9):e148 (2006). doi:10.1371/journal.pgen.0020148.

16. Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099–103 (2010). doi:10.1038/nature09525.

17. Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., *et al.* Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genetics* 9(12):e1003984 (2013). doi:10.1371/journal.pgen.1003984.

18. Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., *et al.* Stable recombination hotspots in birds. *Science* 350(6263):928–932 (2015). doi:10.1126/science.aad0843.

19. Lam, I. and Keeney, S. Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* 350(6263):932–937 (2015). doi:10.1126/science.aad0814.

20. Wang, J., Fan, H. C., Behr, B., and Quake, S. R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 150(2):402–12 (2012). doi:10.1016/j.cell.2012.06.030.

21. Lu, S., Zong, C., Fan, W., Yang, M., Li, J., *et al.* Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338(6114):1627–30 (2012). doi:10.1126/science.1229112.

22. Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., *et al.* Genome analyses of single human oocytes. *Cell* 155(7):1492–506 (2013). doi:10.1016/j.cell.2013.11.040.

23. Lian, J., Yin, Y., Oliver-Bonet, M., Liehr, T., Ko, E., *et al.* Variation in crossover interference levels on individual chromosomes from human males. *Human molecular genetics* 17(17):2583–94 (2008). doi:10.1093/hmg/ddn158.

24. Housworth, E. A. and Stahl, F. W. Is there variation in crossover interference levels among chromosomes from human males? *Genetics* 183(1):403–5 (2009). doi:10.1534/genetics.109.103853.

25. Housworth, E. and Stahl, F. Crossover Interference in Humans. *American Journal of Human Genetics* 73(1):188–197 (2003). doi:10.1086/376610.

26. Baudat, F. and de Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Research* 15(5):565–77 (2007). doi:10.1007/s10577-007-1140-3.

27. Berchowitz, L. E. and Copenhaver, G. P. Genetic interference: don't stand so close to me. *Current genomics* 11(2):91–102 (2010). doi:10.2174/138920210790886835.

28. Stahl, F. W. and Foss, H. M. A two-pathway analysis of meiotic crossing over and gene conversion in Saccharomyces cerevisiae. *Genetics* 186(2):515–36 (2010). doi:10.1534/genetics.110.121194.

29. Fledel-Alon, A., Wilson, D. J., Broman, K., Wen, X., Ober, C., *et al.* Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genetics* 5(9):e1000658 (2009). doi: 10.1371/journal.pgen.1000658.

30. Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D., and Auton, A. Escape from crossover interference increases with maternal age. *Nature Communications* 6:6260 (2015). doi:10.1038/ncomms7260.

31. Hayashi, K., Yoshida, K., and Matsui, Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature* 438(7066):374–8 (2005). doi:10.1038/nature04112.

32. Basheva, E. A., Bidau, C. J., and Borodin, P. M. General pattern of meiotic recombination in male dogs estimated by MLH1 and RAD51 immunolocalization. *Chromosome Research* 16(5):709–19 (2008). doi:10.1007/s10577-008-1221-y.

33. Axelsson, E., Webster, M. T., Ratnakumar, A., Ponting, C. P., and Lindblad-Toh, K. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Research* 22(1):51–63 (2012). doi:10.1101/gr.124123.111.

34. Kong, A., Barnard, J., Gudbjartsson, D. F., Thorleifsson, G., Jonsdottir, G., *et al.* Recombination rate and reproductive success in humans. *Nature Genetics* 36(11):1203–6 (2004). doi:10.1038/ng1445.

35. Hussin, J., Roy-Gagnon, M.-H., Gendron, R., Andelfinger, G., and Awadalla, P. Age-dependent recombination rates in human pedigrees. *PLoS Genetics* 7(9):e1002251 (2011). doi:10.1371/journal.pgen.1002251.

36. Bleazard, T., Ju, Y. S., Sung, J., and Seo, J.-S. Fine-scale mapping of meiotic recombination in Asians. *BMC genetics* 14(1):19 (2013). doi:10.1186/1471-2156-14-19.

37. Martin, H. C., Christ, R., Hussin, J. G., O'Connell, J., Gordon, S., *et al.* Multicohort analysis of the maternal age effect on recombination. *Nature communications* 6:7846 (2015). doi:10.1038/ncomms8846.

38. Hassold, T. and Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics* 2(4):280–91 (2001). doi:10.1038/35066065.

39. Rowsey, R., Gruhn, J., Broman, K. W., Hunt, P. A., and Hassold, T. Examining variation in recombination levels in the human female: A test of the production-line hypothesis. *American Journal of Human Genetics* 95(1):108–112 (2014). doi:10.1016/j.ajhg.2014.06.008.

40. Nagaoka, S. I., Hassold, T. J., and Hunt, P. A. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nature Reviews Genetics* 13(7):493–504 (2012). doi:10.1038/nrg3245.

41. Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. Genetic recombination is directed away from functional genomic elements in mice. *Nature* 485(7400):642–5 (2012). doi:10.1038/nature11089.

42. Nicolas, A., Treco, D., Schultes, N. P., and Szostak, J. W. An initiation site for meiotic gene conversion in the yeast Saccharomyces cerevisiae. *Nature* 338(6210):35–9 (1989). doi:10.1038/338035a0.

43. Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* page aac8624 (2016). doi:10.1126/science.aac8624.

44. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American Journal of Human Genetics* 63(3):861–869 (1998). doi:10.1086/302011.

45. Berg, I. L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N. J., *et al.* 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):12378–83 (2011). doi:10.1073/pnas.1109531108.

46. 23andMe. 23andMe Research Portal. *http://www.23andme.com/researchportal/* (2013).

47. Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., *et al.* Recombination initiation maps of individual human genomes. *Science* 346(6211):1256442–1256442 (2014). doi:10.1126/science.1256442.

48. Hassold, T. and Hunt, P. Maternal age and chromosomally abnormal pregnancies: what we know and what we wish we knew. *Current opinion in pediatrics* 21(6):703–8 (2009). doi:10.1097/MOP.0b013e328332c6ab.

49. Lo, Y. M. D. and Chiu, R. W. K. Prenatal diagnosis: progress through plasma nucleic acids. *Nature reviews. Genetics* 8(1):71–7 (2007). doi:10.1038/nrg1982.

50. Papageorgiou, E. A., Karagrigoriou, A., Tsaliki, E., Velissariou, V., Carter, N. P., *et al.* Fetal-specific DNA methylation ratio permits noninvasive prenatal diagnosis of trisomy 21. *Nature medicine* 17(4):510–3 (2011). doi:10.1038/nm.2312.

51. Palomaki, G. E., Deciu, C., Kloza, E. M., Lambert-Messerlian, G. M., Haddow, J. E., *et al.* DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. *Genetics in medicine : official journal of the American College of Medical Genetics* 14(3):296–305 (2012). doi:10.1038/gim.2011.73.

Appendix A

Christopher L Campbell

Results presented within this appendix were extracted from a manuscript currently under review:

Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales.

Claude Bhérer[1*,2], Christopher L Campbell[1], and Adam Auton[1*].

[1] Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, New York 10461, USA.
[2] New York Genome Center, New York, New York 10013, USA
[3] Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA
[*] Former affiliation.
[#] Present address.

Corresponding authors:
Adam Auton (adam.auton@gmail.com)
Claude Bhérer (claudebherer@gmail.com)

An expanded view of sex dimorphism in recombination in humans and dogs

I have extracted here two figures from a manuscript currently in the review stage (Bhérer *et al.*[1], and attached to this appendix). These have relevance to the data presented in Chapter 4, in which recombination in dogs is compared to that of humans.

The concentration of recombination in the genome

The relationship of genetic to physical distance along a chromosome is not equal. While physical distance is a discrete measurement of the number and sequence of base pairs, recombination does not distribute equally along the chromosome. One way to quantify this is to measure the proportion of recombination occurring in various proportions of sequence. This analysis has been used in a number of studies in recombination in both human[2,3], and dogs[4].

The data presented within this manuscript was used to generate pedigree maps using 104,268 meioses in humans (57,930 female, 46,338 male), collecting data from multiple public sources. This provides the opportunity to study for the first time the concentration of recombination in the genome using sex specific pedigree data. Figure A.1 shows this data. A previously published LD recombination map from the HapMap[3] study is shown in the black line, and represents current understanding (80% of recombination occupying less than 20% of sequence).

In the new pedigree map built (dashed lines), recombination appears to be much less concentrated than in the HapMap study. In addition, the pedigree map underwent a refinement step, aimed at reducing the crossover interval size and producing sex-specific maps on a finer scale. Here, crossover resolution was improved using a method that narrows the intervals defining the crossover boundaries based on overlap with other events. Using the resulting maps (solid lines), recombination appears much more concentrated than both the unrefined pedigree map and the HapMap LD map.

The large differences in the unrefined and refined pedigree maps illustrates the sensitivity of this type of analysis, and emphasizes that care must be taken in interpretation of the results, especially in comparisons to other datasets with differing characteristics. In particular, it appears that the crossover interval size, which has been narrowed in the refined map, has a large effect on the estimated propor-

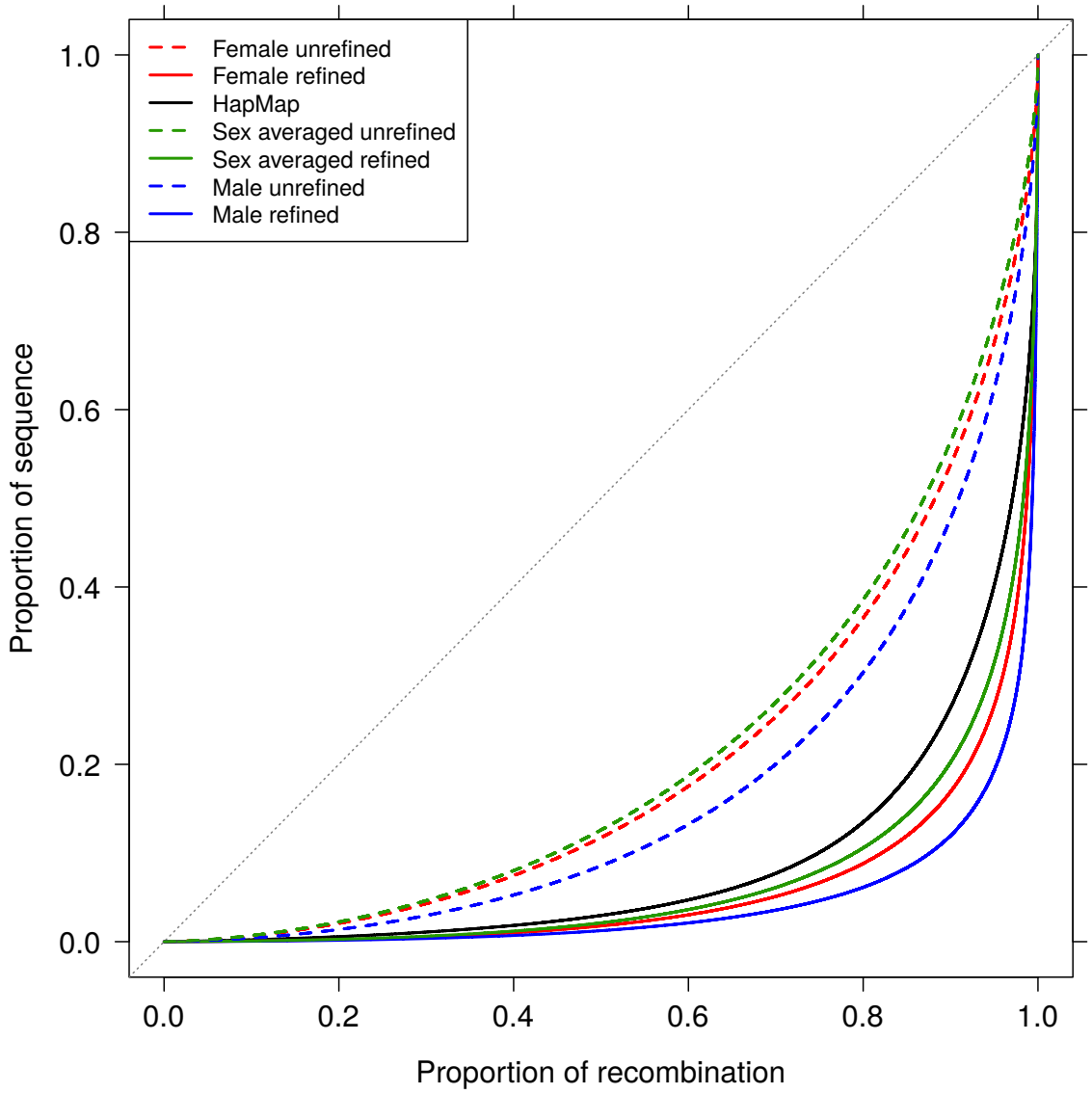

Figure A.1: **The proportion of recombination in various proportions of sequence.** Data is shown from multiple human sources. Dashed lines represent data from the unrefined maps, while data from the refined maps is shown with solid lines. For the refined and unrefined maps, females are shown in red, males in blue, and sex averaged maps in green. Data from the Hapmap LD-based map[3] is shown with a solid black line. This figure is from Bhérer *et al.*[1] (Figure 1E).

tions in the plot, causing recombination to appear more concentrated. These observations support the conclusions from the analysis of dog recombination (Chapter 4) in two ways. First, male recombination is more concentrated in both unrefined and refined maps, mirroring the finding from the thinned human pedigree maps used in Figure 4.1C. Second, the variation in the unrefined and refined curves presented here was shown to be dependent on the crossover interval resolution. This provides support for the validity of the comparison of the thinned human map to the dog data, provided that the human map was thinned properly (Chapter 4, Figures 4.1C, 4.S10). Thus, this supports the conclusion that dog recombination is more concentrated than that of humans.

Recombination around the transcription start site

Another finding is that dog recombination is preferentially located to gene promoter regions located just upstream of the transcription start site (TSS)[4]. This supports the idea that recombination without PRDM9 is directed to regions of open chromatin. Sex specific data in dogs indicated that this effect may be male driven (Chapter 4, Figures 4.S11 and 4.S12), however the resolution is substantially lower than the LD based dog map.

In this study, the pedigree maps were used to quantify recombination around the TSS in humans. Recombination in humans was previously shown to be elevated in the vicinity of the TSS and depressed in gene regions[2,3,5,6]. This study expands upon these findings, using the high resolution of the pedigree based map to show recombination rates for both males and females around the TSS.

Intriguingly, the elevation of recombination around the TSS seems to be entirely driven by females, with males showing no change in rate across the 100 kb window shown in Figure A.2A. This is in direct contrast to the result seen in dogs, where male rates appeared higher just upstream of the TSS. This effect was not due to position within the chromosome for either species, and the same patterns were observed when considering centromeric and telomeric regions separately.

Furthermore, it appears that this elevation in human females is related to PRDM9 binding motifs. The set of genes was partitioned into two groups: those that have PRDM9 binding motifs located with 5 kb, and those without the motif. When excluding genes with a nearby motif, the peak in recombination is eliminated (Figure A.2B). Again, no elevation was seen in males around genes with or without the motif.

This raises the possibility that PRDM9 may be responsible for the rate elevation in females. If true, this could mean that PRDM9 plays a role in sexual dimorphism in recombination. This would fit in with results seen in humans, in which males were found to have a higher hotspot usage than females, by 4.6% (Chapter 2). These results may serve to further explain sex differences in recombination properties. In dogs, which do not have PRDM9, recombination appears to locate instead to regions of open chromatin located at gene promoter regions.

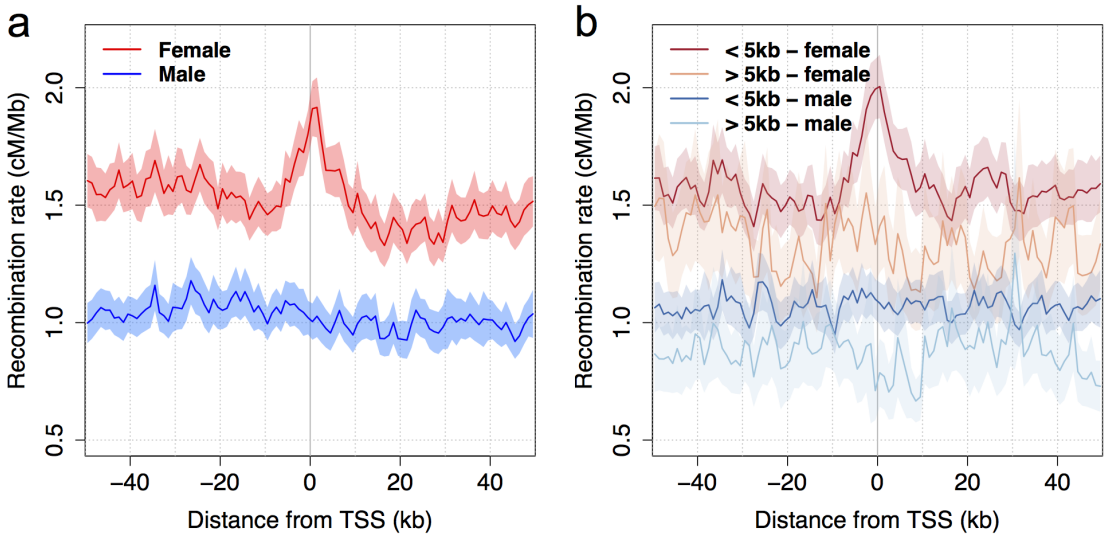

Figure A.2: **Male and female recombination rate around the TSS in humans.** The recombination rate (in cM/Mb) is estimated in 1 kb bins around the transcription start sites (TSS). From the autosomes, 15,239 genes are used, and are thinned so that no two genes in this set fall within 5 kb. Male rates are shown in shades of blue, while female rates are shown in red shades. Panel A represents the entire set of genes. In panel B, data from each sex has been split to include or exclude genes that have the PRDM9 13 bp motif within 5 kb of the TSS. This figure is from Bhérer *et al.* [1] (Figure 5) and was generated by Claude Bhérer.

References

1. Bhérer, C., Campbell, C. L., and Auton, A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Unpublished - under review* (2016).

2. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310(5746):321–4 (2005). doi: 10.1126/science.1117196.

3. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–61 (2007). doi:10.1038/nature06258.

4. Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., *et al.* Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genetics* 9(12):e1003984 (2013). doi:10.1371/journal.pgen.1003984.

5. McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* 304(5670):581–4 (2004). doi:10.1126/science.1092500.

6. Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099–103 (2010). doi:10.1038/nature09525.