STRUCTURAL VARIATION AND THE EVOLUTION OF THE MOUSE GENOME

Andrew Parker Morgan

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Genetics in the School of Medicine.

Chapel Hill
2017

Approved by:

Fernando Pardo-Manuel de Villena

Terry Magnuson

Leonard McMillan

Daniel Pomp

Thomas Petes

Gary Churchill

ABSTRACT

Andrew Parker Morgan: Structural variation and the evolution of the mouse genome
(Under the direction of Fernando Pardo-Manuel de Villena)

Genetic variation in populations is governed by four basic forces: mutation, recombination, natural selection and genetic drift. Mutation is the source of new alleles, which are assorted among chromosomes by recombination. Selection and drift dictate the magnitude and direction of changes in allele frequency over time. These forces are intimately linked to meiosis, and asymmetries in meiosis create the opportunity for intragenomic conflict: competition between selfish alleles at the same locus for transmission to progeny. Such conflicts manifest as selection at the population level but subvert the Darwinian concept of fitness.

The aim of this thesis is to characterize three of the four basic forces — recombination, mutation and intragenomic conflict — using the house mouse as a model system. I focus on the role of large segmental duplications, long tracts of repeated sequence that make up approximately 10% of mammalian genomes and are the site of the preponderance of structural variation between individuals. First I use two laboratory populations, the Collaborative Cross and the Diversity Outbred, to analyze the effects of sex and genetic background on the rate of recombination. I discover that (crossover) recombination is strongly suppressed in both sexes near large multiallelic copy-number variants. Second I reconstruct in detail the evolution of one such variant, *R2d2*. I show that *R2d2* represents an ancient duplication that has been amplified to more than 100 copies in some lineages of European mice. Alleles with high copy number ($R2d2^{HC}$) are associated with suppressed recombination but have an extremely high mutation rate. They are also selfish, having risen to high frequency in wild and laboratory populations by meiotic drive, in spite of their deleterious effect on reproductive fitness. Finally I perform a comprehensive survey of sequence and structural variation on the mouse Y chromosome. I show that patterns of nucleotide and structural diversity have been shaped by intragenomic conflict with the X chromosome.

ACKNOWLEDGEMENTS

The work presented in this thesis was possible only through the advice, support and material contribution of many faculty, trainees and staff at the University of North Carolina and elsewhere. Many individuals deserve my thanks; any omissions are by fault of memory.

I am foremost grateful to my adviser, Fernando Pardo-Manuel de Villena. His generosity, deep enthusiasm for science, high expectations and relentless eye for detail have shaped my development as a scientist and prepared me well for the future.

I was fortunate to work alongside many talented students, postdocs, researchers and staff in the Pardo-Manuel de Villena group and the Department of Genetics: David Aylor, Tim Bell, Ryan Buus, John Calaway, Mark Calaway, Sarah Cates, Amelia Clayshulte, Jim Crowley, John Didion, Lauren Donoghue, Marty Ferris, Justin Gooch, Pablo Hock, Leeanna Hyacinth, Samir Kelada, Yunjung Kim, Colton Linnertz, Rachel McMullan, Darla Miller, Randy Nonneman, Dan Oreper, Corey Quackenbush, Nikki Robinson, Paola Gusti-Rodriguez, Allison Ryan, Ginger Shaw, John Shorter, Pat Sullivan, Wei Sun, Will Valdar, Brit Wanstrath, Lucy Williams and Vasyl Zhabotynsky.

Regular interactions with the group of Leonard McMillan in the Department of Computer Science have been a valuable part of my training: Matt Holt, Seth Greenstein, Jeremy Wang, Shunping Huang, Chen-Ping Fu, Katy Kao, Catie Welsh, Sebastian Sigmon and Maya Najarian.

Through my adviser's collaboration with Gary Churchill I have enjoyed working with several researchers at the Jackson Laboratory: Dan Gatti, KB Choi and Chris Baker.

I am grateful for the guidance of my thesis committee: Terry Magnuson (chair), Leonard McMillan, Daniel Pomp, Thomas Petes and Gary Churchill.

My studies would not have been possible without funding from the Bioinformatics and Computational Biology Training Grant, the Medical Scientist Training Program, the Center for Genome Dynamics, the Mutant Mouse Resource and Research Center, the National Institute of Allergy and Infectious Disease and the National Institute of Mental Health. I thank the administrative staff in the Department of Genetics, John Cornett and Cara Marlow, and the School of Medicine, Alison

Regan and Carol Herion, for their patience and support.

I thank Todd Vision, Mark Heise and the late Gene Orringer for providing opportunities for me to get started in research as an undergraduate.

Finally, I am grateful for the continued love and support of my family, particularly my wife Katie, and to the camaraderie of my fellow MD/PhD students.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction I: The mammalian germline

The aim of this thesis is to characterize three basic forces governing the level and distribution of genetic variation in populations: recombination, mutation and intragenomic conflict. In particular I focus on the role of large structural variants in repetitive regions of the genome and the ways in which differences between males and females influence the accumulation of mutations and create sex-specific opportunities for the spread of selfish genetic elements.

♢ ♢ ♢

The entirety of individual's genetic information is transmitted to the next generation by a single cell, the gamete. In most animal species, a dedicated population of cells — the germline — is responsible for the production of gametes. Only mutations arising in the germline are heritable. Although mutations arising in all other (somatic) tissues may have important consequences for the organism — for instance, cancer — they are not transmitted, and neither contribute to the genetic diversity of the population nor are subject to natural selection.

Patterns of genetic diversity in populations are therefore intimately connected to processes that maintain the integrity of the genome during the series of cell divisions that ultimately lead to functional gametes. In this section I review key features of germline structure and function from an evolutionary perspective, with special attention to differences between males and females. Although I focus on the properties of the germline in mammals, many of these properties are deeply conserved across multicellular eukaryotes.

1.1 Meiosis: from diploid to haploid

Mammals, like most animals, are diploid and reproduce sexually, inheriting one copy of the genome from each parent. In order to reproduce the organism must reduce the diploid genome to a haploid state. This is achieved by the specialized form of cell division called *meiosis*. Meiosis is

ubiquitous among eukaryotic taxa and its key steps are conserved from higher organisms to the simplest unicellular species (**Figure 1.1**). As the channel through which all genetic information must flow from one generation to the next, the events of meiosis are of fundamental evolutionary importance. Errors in meiosis are also clinically relevant, as they frequently lead to infertility and developmental disorders.

Like mitosis, meiosis begins with a round of DNA replication in *primary gametocytes* which are diploid ($2N$) and have a double complement of DNA ($4C$). Here the program diverges from mitosis: replication is followed by a reductive division (meiosis I), in which the two members of each homologous chromosome pair are segregated to daughter cells. Segregation of maternal and paternal copies occurs independently and stochastically for each chromosome pair. The daughter cells of the reductive division, the *secondary gametocytes*, are haploid ($1N$) and no longer genetically equivalent, but still carry a double complement of DNA ($2C$). At the second, equational division (meiosis II), sister chromatids are segregated into gametes which are both haploid ($1N$) and carry a single genome-equivalent of DNA ($1C$).

The segregation of homologous chromosomes during meiosis I provides the mechanistic basis for Mendel's two rules of inheritance: the "law of segregation", which states that for each chromosome either the maternal or the paternal copy is transmitted, but not both; and the "law of independent assortment", which states that this process occurs independently for each chromosome pair. Mendel's laws were based on empirical observations in breeding experiments long before the details of meiosis were known, but they elegantly and accurately predicted patterns of inheritance and formed the basis of the chromosome theory of heredity.

With few exceptions in animals, meiosis I is preceded by an extended prophase during which some or all chromosomes undergo *recombination*, a tightly-regulated genetic exchange between homologs inherited from each parent[1]. Recombination begins with the programmed introduction of double-strand breaks (DSBs) at up to hundreds of sites per germ cell. One strand is enzymatically resected at the free ends, and the resulting naked strands scan the homologous chromosome for matching sequence. This homology search mediates chromosome pairing, a critical step of meiosis I. A small subset (on the order of 10%) of free ends form physical connections between homologous chromosomes called *chiasmata*. A chiasma holds members of the chromosome pair in tension as they attach to the spindle apparatus and are pulled towards opposite poles. In mammals, a

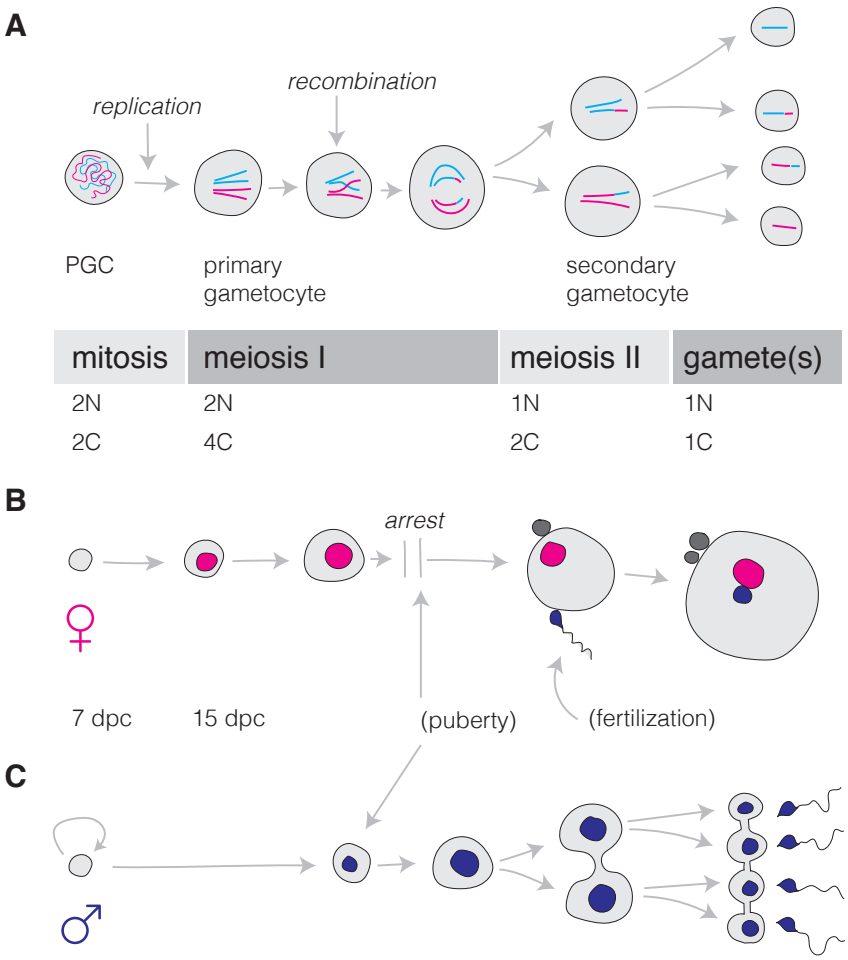

Figure 1.1: Overview of meiosis and gametogenesis. (**A**) General features of meiosis in eukaryotes. A round of DNA synthesis occurs in primary gametocytes, derived from the primordial germ cell (PGC) pool. Recombination occurs during the first meiotic prophase, followed by a reductive division and then an equational division to produce up to four gametes per PGC. The total chromosome number (N) and DNA complement (C) is shown for each stage. (**B**) Female meiosis and oogenesis. PGCs have begin to differentiate by 7 dpc and meiosis begins by 15 dpc. Primary oocytes arrest after forming chiasmata, and meiosis does not resume until after puberty (upon ovulation). The second meiotic division is triggered by fertilization. Each division produces one functional product and one polar body; the product of meiosis II is the mature oocyte. (**C**) Male meiosis and spermatogenesis. The PGC pool in males continues to divide throughout life. Meiosis is not initiated until puberty, but once it begins, it continues in orderly waves. Secondary spermatocytes and round spermatids may be connected as a syncytium.

minimum of one chiasma per chromosome arm is generally required to ensure proper disjunction . Aberrant disjunction leads to gametes with either too few or too many chromosomes (*aneuploidy*); many aneuploidies are incompatible with life, and those that are not are associated with profound developmental and reproductive defects.

Chiasmata are resolved as crossovers — reciprocal exchanges between the chromosome segements up- and downstream of the initiating DSB — during metaphase I. These crossovers enhance the combinatorial mixing between parental genomes already induced by independent assortment. The stochastic nature of the recombination process along chromosomes ensures that the haploid genome of each gamete is a unique mosaic of the maternal and paternal genomes. By generating new combinations of alleles, recombination facilitates natural selection[2]. Variation in the rate of recombination along the genome is an important determinant of haplotype structure in populations[3,4].

The process of recombination and its evolutionary significance is reviewed in more detail in **??**.

1.2 Gametogenesis in mammals

Meiosis is embedded in the process of gametogenesis — the production of sperm or oocytes — in higher organisms. Because gametogenesis differs between the sexes, the regulation and timing of meiosis also differs between males and females. To understand sex differences in gametogenesis therefore requires a brief discussion of sex determination in mammals.

1.2.1 Sex determination

Placental (*eutherian*) mammals — those mammals besides marsupials (such as kangaroos) and monotremes (platypus and echidna) — have a chromosomal system of sex determination. Individuals with two X chromosomes are genetically female, and those with an X and a Y are genetically male. The X and Y chromosomes are together referred to as the *sex chromosomes* (the remaining pairs being *autosomes*); they are the only chromosome pair in which the two homologs are different in content and structure[5]. However, the mammalian embyro begins its development with the potential to be anatomically and physiologically male or female: the precursors of both male and female gonads are both present in the urogenital ridge regardless of sex chromosome karyotype. In the absence of a Y chromosome, the gonad differentiates into an ovary. The presence of a single factor expressed from the Y chromosome — the product of the gene *Sry* (sex-determining region on the Y) — is sufficient to induce the development of the testis[6]. The process of *primary*

sex determination is complete upon specification of the gonad by 13.5 days post-conception (dpc) in mouse. Remaining internal and external anatomical structures characteristic of each sex then develop under the influence of hormones secreted by somatic cells in the gonad. Experiments using mice with abnormal sex chromosome karyotypes were instrumental in demonstrating the distinction between primary and secondary sex determination, and the role of the Y chromosome in the former[7].

The *primordial germ cells* (PGCs) that will give rise to the germline differentiate very early in embryonic development, around 8.5 dpc in mouse and 15 dpc in humans[8]. By 12 dpc (mouse) or 50 dpc (human) the PGCs have migrated into the developing gonad. Importantly, these steps precede the process of primary sex determination, and indeed the eventual fate of PGCs as oocytes or sperm depends on the sex of the somatic cells in the gonad and not the sex-chromosome karyotype of the PGCs themselves. Once PGCs have reached the gonad the germline developmental programs of females and males diverge.

1.2.2 Oogenesis

Female germ cells enter meiosis by 15 dpc[9], having undergone a total of approximately 25 cell divisions since fertilization[10]. The entire pool primary oocytes — numbering in the tens of thousands — undergoes recombination and the formation of chiasmata but arrest at the end of the first meiotic prophase, and remain arrested until puberty (**Figure 1.1B**). Thereafter small cohorts of oocytes are periodically released from meiotic arrest in synchrony with cyclic physiological changes in the uterus (the estrous or menstrual cycle) in anticipation of pregnancy. Completion of meiosis I is concurrent with ovulation. This division is asymmetric, producing one secondary oocyte and a non-functional polar body. The secondary oocyte arrests again; meiosis II is triggered by fertilization and is again asymmetric, yielding one mature oocyte and a polar body. Each primary oocyte therefore gives rise to a single functional gamete and three non-functional products.

Two properties of mammalian oogenesis — the extended period of arrest during prophase I and the asymmetry of both meiotic divisions — present challenges unique to females. The first is to maintain the integrity of attachment to and alignment on the spindle apparatus for a period of months to decades. Indeed, nearly all cases of aneuploidy in humans are of maternal origin, and the risk of improper disjunction increases dramatically with maternal age[11]. Chiasmata, which allow homologs to be held in tension perpendicular to the spindle, are thought to be critical for the stability

of chromosomes during meiosis I: nondisjunction is associated with decreased recombination in several human trisomies[12]. Strong selection against aneuploidy may be responsible for the apparent requirement of one crossover per chromosome arm in mammals[13].

The second challenge to be addressed during female meiosis is a more teleological one. The asymmetry of both meiotic divisions leaves open the possibility of *intragenomic conflict*: a selfish allele that can exploit the meiotic machinery to increase its probability of segregation to the oocyte rather than the polar bodies can increase in frequency independent of its effect on organismal fitness. This phenomenon of non-random segregation is called *meiotic drive*. True meiotic drive — as opposed to other systems influencing the transmission ratio at a locus, such as selection against a particular class of gametes — requires asymmetric meiotic division, an asymmetric meiotic spindle, and functional heterozygosity at a locus that mediates interaction with the spindle[14]. These requirements can be met in female but not male meiosis in mammals. The role of meiotic drive in the evolution of a specific locus in mouse, *R2d2*, is the subject of **Chapter 4** of this thesis.

From a practical point of view, the fact that they key events in female meiosis occur while the female herself is an early embryo make them quite difficult to study. The study of meiosis in general, and recombination in particular, has benefitted greatly from model systems in which all four products of a single meiosis can be observed simultaneously as a tetrad, as is the case in many fungi (*e.g. Saccharomyces cerevisiae*, *Neurospora crassa*) and flowering plants (*e.g. Arabidopsis thaliana*). The products of mammalian meiosis are not a tetrad, and heroic technical efforts are required to recover the closest equivalent structure — an oocyte with its corresponding polar bodies — in a mammal[15]. Knowledge of mammalian meiosis and recombination is therefore somewhat biased towards the male germline.

1.2.3 Spermatogenesis

Whereas meiosis begins in females concurrent with early embryonic and gonadal development, male germ cells do not enter meiosis until the onset of puberty. Instead, the PGCs differentiate into a population of self-renewing stem cells called *spermatogonia* that undergo a period of arrest from 16.5 dpc until the first week after birth. Thereafter they divide every 8 days for the remainder of life (**Figure 1.1C**). The spermatogonia reside along the *seminiferous epithelium* that lines the basement membrane of the tubules of the testis. They are nurtured by somatic cell lineages (Sertoli and Leydig cells) that lie between the tubules.

At puberty, spermatogonia enter meiosis in coordinated waves taking $30 - 33$ days each in mouse. The resulting primary and secondary spermatocytes migrate towards the lumen of the tubule in an orderly fashion with clearly-defined stages[16]. In fact, chains of spermatogonia (and later spermatocytes) often form physically-connected chains that share cytoplasm as a syncytium. The products of male meiosis are round spermatids, which further differentiate into mature spermatozoa. In contrast to female meiosis, both meiotic divisions are symmetric in males, and a single spermatogonium gives rise to four functional gametes.

The sex chromosomes present two challenges unique to male meiosis. First, proper segregation of the X and Y chromosomes requires that they pair and recombine despite their very different size and structure. In most mammals, pairing and recombination are restricted to a short region of residual homology at the tips of the sex chromosomes called the *pseudoautosomal regions* (PAR)[5]. (The human sex chromosomes are metacentric and have a PAR on each chromosome arm; the mouse sex chromosomes are acrocentric and have a single PAR at the end of their long arms.) The PAR thus becomes a site of intense crossover activity: the recombination rate per unit physical distance is $10 - 50$ times greater than on the autosomes.

The second challenge involves transcriptional control of the sex chromosomes during meiosis. Expression of some Y-linked genes (including *Zfy1* and *Zfy2*) is toxic during meiosis and leads to infertility[17]. As a result, transcription from the unpaired regions of the sex chromosomes is epigenetically suppressed from late in the first meiotic prophase until after meiosis is complete. This *meiotic sex chromosome inactivation* (MSCI) is absolutely required for fertility in mouse and is a special case of the more general process of *meiotic silencing of unsynapsed chromatin* (MSUC)[18,19]. MSUC, in turn, is thought to have evolved as a defense against transcription from selfish genetic elements in the germline. The unpaired regions of the X and Y chromosome co-localize in a structure referred to as the *sex body* which is spatially distinct from the autosomes during late prophase and mediates suppression of both transcription and recombination in these regions[20].

Because male meiosis is symmetric with respect to cell fate, true meiotic drive cannot occur in males. However, there are many cases of distorted transmission due to gametic selection in heterozygous males, the prototypical example of which in mammals is the mouse *t*-haplotype (reviewed in Lyon[21]). These are two-component systems involving a *responder* locus, at which a selfish allele is transmitted at higher than the expected Mendelian frequency of $1/2$; and a *distorter*

that is toxic in sperm carrying the wild-type allele[22]. Recombination between the responder and distorter ablates drive; the two components are often tightly linked on a single chromosome and in some cases held in complete linkage by an inversion[23].

When transmission distortion affects the sex chromosomes it is known as *sex-ratio drive* and has particular evolutionary importance. Unequal transmission of the sex chromosomes in the heterogametic sex (males, in mammals) may distort the sex ratio in the population, eventually reducing mean fitness, and fuel intragenomic conflict between the sex chromosomes[24]. Sex-ratio drive is thought to have played an major role in shaping the gene content of the mammalian sex chromosomes[25]. I return to these ideas in **Chapter 6**.

1.3 The nature of germline genetic variation

All heritable genetic variation arises by mutation in the germline. I briefly discuss the spectrum of germline mutations — in increasing order of size from single-base substitutions to large-scale rearrangements — and their relative rates, and how they differ between the male and female germline.

1.3.1 Small-scale sequence variation

The simplest class of mutations is single-base changes (*single-nucleotide variants*, SNVs; or point mutations). SNVs arise at a sex-averaged rate on the order of $10^{-9} - 10^{-8}$ per base per generation ($\text{bp}^{-1}\text{gen}^{-1}$) in mammalian genomes. The current estimate for humans is $1.2 \times 10^{-8}\,\text{bp}^{-1}\text{gen}^{-1}$ and for mouse $5 \times 10^{-9}\,\text{bp}^{-1}\text{gen}^{-1}$[26]. This equates to approximately 30 new mutations per haploid gamete. SNVs are therefore by far the most frequent type of new mutations and the most abundant class of variants between individuals in the population, accounting for $> 99.9\%$ of known variants in humans[27].

Point mutations are dependent on DNA replication: they arise either as simple copying errors, or via repair of damaged dinucleotide pairs. The rate of point mutation varies dramatically depending on local sequence context, largely as a result of the susceptibility of different dinucleotides to DNA damage (reviewed extensively in[26]). Transitions (swaps between nucleotides of the same chemical class) are approximately two-fold more frequent than transversions (swaps between nucleotides of different classes). Among transitions, mutations at CpG dinucleotides are approximately tenfold more abundant than expected based on the number of CpGs in mammalian genomes. Current evidence for context-dependence of SNVs is drawn mostly from cross-species comparisons, but

patterns of within-population variation from recent large-scale surveys are in good agreement[28].

Short insertion and deletion (*indel*) polymorphisms, operationally limited to polymorphisms < 10 bp in size, are between seven- and ten-fold less abundant then SNVs in both humans[27] and mice[29]. However, indels at *microsatellite* sequences — tandemly-arrayed repeats of $1 - 8$ bp — arise at a much higher rate. Mutation in microsatellites tends to occur in units of the underlying repeat via slippage of DNA polymerase against the template strand. Mutation rate at microsatellites varies over several orders of magnitude depending on the length of the tandem array and the size of the repeat unit, with longer tracts of smaller repeat units being less stable in general[30]. Direct estimates of microsatellite mutation rates in mouse[31] and human pedigrees[32] are similar, on the order 10^{-4}.

It should be noted that SNVs are much more readily detected than indels (especially at mirosatellite loci) using current technologies. Structural variants, discussed below, are more challenging still. The consequences of this ascertainment bias are discussed in more detail in **§ 1.5**. Existing catalogs of sequence variation in organisms with large and complex genomes like those of mammals thus offer an incomplete view of the true genetic diversity in populations[33].

1.3.2 Sub-chromosomal structural variation

Mutations that alter the copy number, order or organization of kilobase- to megabase-sized regions of the genome can be gathered under the umbrella of *structural variants* (SVs). These can be subdivided into *copy-number variants* (CNVs); rearrangements in the absence of changes in copy number; and insertions or excisions of *transposable elements* (TEs).

Copy-number variants (CNVs). CNVs are the best-characterized class of SVs, both in population frequency and in *de novo* rates. Although the mutation rate per generation for CNVs is estimated to be lower than for SNVs (1.2×10^{-2} per haploid gamete[34]), CNVs affect a much larger genomic territory. Indeed the total number of base pairs which differ between two humans is as much as 100-fold larger for CNVs than SNVs[35]. Because of the larger footprint of CNVs, an individual CNV is more likely to overlap functional elements of the genome including protein-coding genes than an individual SNV or small indel. Mutations incompatible with embryonic development and live birth will go unobserved in pedigree- or population-based studies. It is therefore important to note that it is more difficult to deconvolute the effects of mutation from purifying selection for CNVs than for smaller variants.

The mutation rate for *de novo* CNVs varies by orders of magnitude along the genome. The sequence feature most strongly associated with copy-number mutation is the presence of existing *segmental duplications* (SDs), operationally defined as duplications > 1 kb in size with $> 90\%$ pairwise sequence identity[34]. More than half of currently-known CNVs segregating in the human population are associated with SDs[36], and SD-associated CNVs are larger and more likely to have multiple alleles than otherwise[37,38,39]. The preponderance (5-10%) of SD content in mammalian genomes and their polymorphism in populations were among the earliest and most striking observations of the first era of genome-sequencing projects[40,41,42,43]. The CNV mutation rate at some SDs in humans is high enough that similar mutations reoccur with detectable frequency at the same loci. At least 50 such loci in the human genome are associated with developmental or psychiatric disorders — 7q11 (Williams-Beuren syndrome), 11p15.5 (β-thalassemia), 16p11.2 (autism), 17q11.2 (neurofibromatosis type 1), 17p12 (Charcot-Marie-Tooth syndrome type 1A) — collectively known as "genomic disorders"[40].

CNVs associated with SDs are thought to arise primarily via *non-allelic homologous recombination* (NAHR), that is, illegitimate recombination between duplicated sequences[44]. NAHR was first proposed as mechanism for changes in copy number by Sturtevant in 1925, studying the *bar* locus in *Drosophila melanogaster*[45] and was subsequently recognized as an important mechanism in the expansion of gene families[46]. Exchange may occur between duplicates on the same chromatid, resulting in deletion; or between sister chromatids or homologous chromosomes, yielding either deletion or duplication plus a reciprocal product (**Figure 1.2**). Direct estimates of the relative contribution of intra-chromatid versus inter-chromosomal exchange in sperm suggest that inter-chromosomal exchanges with the homologous chromosome predominate[47]. This indicates that, at least at the CNV hotspots studied, SD-associated mutations are predominantly meiotic rather than mitotic in origin. NAHR in mammals apparently requires tracts of uninterrupted sequence similarity on the order of 100 bp in length[48].

A subset of copy-number mutations are associated with SDs but are not "recurrent" in the sense that the same breakpoints are not reused by independent mutational events. The mechanism(s) underlying these mutations have more in common with those generating CNVs not associated with SDs[44]. These non-recurrent CNVs may arise by one of a family of DNA-replication-based mechanisms that involve synthesis of a new DNA strand in response to DNA damage[49,50]. These

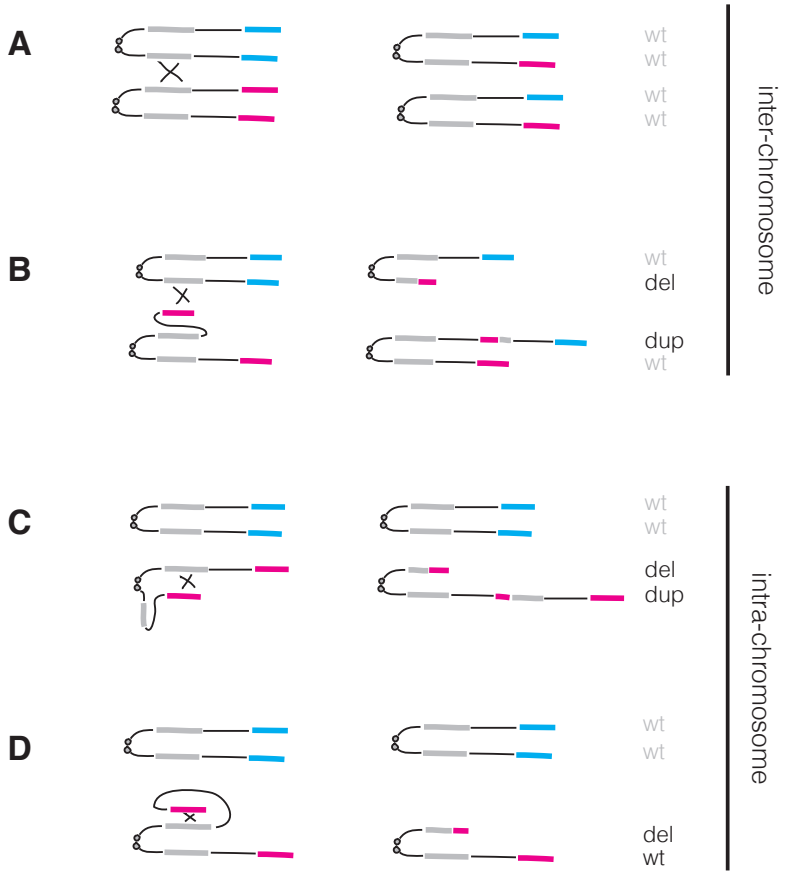

Figure 1.2: Formation of structural variants via non-allelic recombination between duplicate sequences. Duplicates are colored grey and either pink (maternal chromosomes) or blue (paternal chromosomes). All four chromatids are shown. (**A**) Normal recombination with crossing over between allelic duplicates. All four products are wild-type (wt). (**B**) Recombination with crossing over between non-allelic duplicates on different chromosomes. Two products are wild-type and the other two are a deletion (del) and its reciprocal duplication (dup). (**C**) Recombination between two maternal sister chromatids. Paternal chromosomes are both wild-type while the maternal are a deletion and reciprocal duplication. (**D**) Intra-chromatid recombination between non-allelic copies on one maternal chromatid. Paternal chromosomes and the unaffected maternal chromatid are all wild-type, while the affected maternal chromatid has a deletion. Unlike cases B and C, there is no reciprocal duplication product. Note that alternative outcomes are possible if duplicates have inverted rather than direct orientation.

mechanisms — and the resulting CNV alleles — may become quite complex, involving multiple double-strand breaks and switches in the repair template. Non-recurrent CNVs arising by aberrant replication are, by construction, mitotic rather than meiotic events.

Rearrangements. Changes in sequence organization in the absence of changes in copy number are among the most difficult class of mutations to detect. (In fact copy-number mutations, especially those associated with SDs, frequently also involve rearrangement[44].) From the point of view of this thesis, the most important class of rearrangement is *inversion*[51]. Crossing-over is supressed or completely absent across an inversion in heterozygotes because of spatial constraints on chromosome pairing; when an (odd number of) crossover does occur, the resulting products are one dicentric and one acentric chromosome, typically incompatible with stable segregation to gametes. Inversions are extremely important in the processes of karyotype evolution and speciation[52], and in particular in the evolution of sex chromosomes[53]. These ideas are revisited in **Chapter 4** and **Chapter 6**.

Transposable elements. Approximately half of both the human[54] and mouse[55] reference genomes are comprised of sequence corresponding to TEs. TEs are viruses and virally-derived fragments that were actively replicating in the host germline at some point during its evolutionary history and have integrated into the host genome. Although the biology of TEs is fascinating and appears to differ between mammalian lineages, a full review is beyond the scope of this thesis. It suffices to say that more than $100,000$ polymorphic TE insertions have been identified in inbred strains of mice and that most TE polymorphism — both insertion and excision — arises in the male germline[56]. Because TEs can be collapsed onto a few ancestral sequence families, they are also called *interspersed repeats*. Illegitimate recombination between TEs mediates some recurrent CNVs, although the contribution is smaller than for SDs[44].

Cytological markers. A final and class of genetic variation deserves mention: variation in cytologically-visible markers of gross chromosome structure including centromeric heterochromatin[57], "nucleolar organizer regions" (NORs; clusters of ribosomal RNA genes)[58] and "homogeneously-staining regions"[59]. Although variation in chromatin-staining patterns has been historically important in phylogeny and in the study of the mutagenic effects of environmental exposures[60],

the relationship of cytological markers to underlying sequence changes is largely unknown. The corresponding sequences are at best incompletely represented in reference genome assemblies of both mouse and human. Their repetitive nature makes them refractory to accurate assembly even with significant manual effort. Yet at least some of these genomic features are associated with important phenotypes: variation in centromeres, for instance, directly influences the fairness of chromosome segregation in meiosis[14,61].

1.3.3 Sex differences in mutation rates

The rate and spectrum of mutation differs between the males and females as a direct consequences of sex differences in germline physiology. The best-characterized patterns are the increased risk of aneuploidy in oocytes discussed earlier in this chapter and the so-called *paternal-age effect* on point mutations.

It has long been known that the recurrence risk of many Mendelian disorders with autosomal dominant inheritance pattern — including Apert syndrome, Waardenburg syndrome, osteogenesis imperfecta, neurofibromatosis, certain types of achondroplasia and (X-linked) haemophelia A — increases with paternal age[62,63]. The observation originated from two early heroes in genetics, Wilhelm Weinberg[64] and J.B.S. Haldane[65]. The correlation between parental ages induced by normal marriage patterns has historically made it difficult to independently estimate the contribution of paternal versus paternal age to recurrence risk[66]. Furthermore, some have argued that the paternal-age effect could be explained in part by spermatogonial selection[67,68]. This hypothesis is based on the observation that tumor suppressors, growth factors or their receptors (*e.g. FGFR3*, a fibroblast growth factor receptor, in achondroplasia) are overrepresented among the causative genes for Mendelian disorders with a paternal-age effect. If a *de novo* mutation confers a tendency for spermatogonia to divide more rapidly, the mutation is more likely to be transmitted and observed recurrence risk represents a conflation of mutation rate with spermatogonial growth rate. The spermatogonial selection hypothesis also predicts a faster-than-linear increase in mutational burden with age, a pattern which has been observed for a subset of disorders with a paternal-age effect[63].

Recent large-scale sequencing efforts in non-disease human pedigrees have demonstrated conclusively that the average mutation rate in the male germline is three- to four-fold higher than in the female germline[69,70,28]. This is consistent with estimates based on comparisons between humans and great apes[71]. Furthermore, the rate of point mutations does indeed depend on paternal

but not maternal age[70,28], with an extra one to two mutations trasmitted for each extra year of age. A similar relationship exists for microsatellites[32]. Age-associated mutations are not uniformly distributed across the genome but tend to occur in early-replicating regions and near genes[28] and are thus more likely to have functional consequences.

It is less clear whether the rate of *de novo* structural mutation differs between the sexes or is related to age. At least two large cohort studies of intellectual disability or developmental delay have reported an excess of *de novo* CNVs of paternal origin and an increased burden of *de novo* CNVs in offspring of older fathers[72,73]. However, a strong maternal bias has been reported for recurrent mutations at the autism-associated 16p11.2 locus[74]. Only one study (the Genome of the Netherlands Project[75]) has estimated the rate of *de novo* structural mutation in non-disease pedigrees, and it found a two-fold excess of mutations on paternal versus maternal haplotypes. The discrepancies between these results might be resolved by stratifying mutations according to likely meiotic versus mitotic origin. *De novo* CNVs arising in or near existing segmental duplications — perhaps by illegitimate recombination during meiosis — tend to show a more subtle paternal bias and no age effect[35]. This implies that spontaneous mutations in regions of the genome susceptible to recurrent CNVs, including loci associated with human disease, are less likely to be sex- or age-biased. By contrast, non-recurrent mutations — presumably arising by mechanisms involving errors of replication[44] — apparently occur more frequently in the male germline and are subject to an age effect[72]. However, the empirical evidence for these patterns remains thin.

Sex differences in the rate and spectrum of spontaneous mutations seem to be parsimoniously explained by the male and female germline discussed earlier in this chapter. Whereas all oocytes have undergone an approximately equal number of cell divisions (25 in mouse; $40-50$ in human) between their origin in the early embryo and fertilization, the stem cells of the male germline continue to divide throughout reproductive life. The sperm of a 45-year-old man have undergone at least tenfold more divisions than those of a 25-year-old man, even allowing for some variation in the rate of cell division across different subtypes of spermatogonia[76]. All classes of mutations that can be attributed to errors of DNA replication in actively-dividing cells are expected to increase linearly with paternal age.

1.4 Genetic variation in populations

The level of genetic variation in populations is governed by four "fundamental forces:" mutation, recombination, drift and selection[77]. So far I have discussed mutation and recombination and their underlying biological basis at the individual level. In this section I introduce the remaining two forces, which are population- rather than individual level processes. This discussion is not meant to be exhaustive but rather to provide a starting point for discussion in the main chapters of this thesis. [1]

1.4.1 Genetic drift

In a population of finite size N, allele frequencies change over time due to the stochastic effects of sampling the next generation's alleles from the current generation's gametes. Every new allele that arises by mutation has initial frequency $\frac{1}{2N}$, and in finite time it will either be lost or fixed. The rate of drift can be summarized by the change in heterozygosity between successive generations (δH). In a population of diploids:

$$\delta H = \frac{1}{2N}$$

That is, genetic drift is more rapid in smaller populations. For neutral alleles, the the probability of fixation is equal to initial frequency ($p_{fix} = \frac{1}{2N}$ [79]) and the expected time to fixation $\tau \approx 4N$ [80]. Genetic drift is accelerated by *inbreeding*, or mating between individuals sharing recent genealogical ancestors.

1.4.2 Genetic diversity and population size

In an idealized population — a population which is constant in size, randomly-mating, sealed off from in- or out-migraiton, and either hermaphroditic or having sex ratio at parity — there exists an equilibrium at which the entry of new alleles from mutation is balanced by the exit of existing alleles by drift. At this equilibrium levels of standing variation are proportional to the product of mutation rate (μ) and population size (N). Formally, we can describe genetic diversity at a locus

[1] With the exception of some results of particular interest for which original references are cited, most of the material here is summarized from [78] and can be found in any introductory text on population genetics.

via the *population-scaled mutation rate* (θ)):

$$\theta = KN\mu$$

where K is a scaling factor. This scaling factor arises because the fundamental unit of inheritance is the chromosome, not the individual: mutations arise on chromosomes, and they are transmitted on chromosomes. Population size therefore depends on the number of transmissible chromosomes at a locus. For a diploid organism with an X-Y sex chromosome system, $K = 4$ for autosomes, since each sex can transmit either member of each chromosome pair); $K = 3$ for the X chromosome, since females can transmit either of their two Xs and males can transmit one; and $K = 1$ for the uniparentally-inherited Y chromosome and mitochondrial genome.

When a population deviates from idealized assumptions, we can replace the census population size N by the *effective population size*, N_e. To a first approximation this parameter absorbs many possibly-unknown demographic factors into a single value that allows us to describe the levels of genetic diversity and drift we would expect if the true population were replaced by one of size N_e that conformed to idealized assumptions[81,82].

Under the "infinite-sites" assumption, mutations do not occur at the same site twice, so every new mutation creates a new *segregating site*. The number S of segregating sites in a sample of chromosomes from the population can therefore be used to estimate genetic diversity[83]:

$$\theta = \frac{S}{\sum_{i=1}^{2N-1} 1/i}$$

Under the assumption of random mating (*i.e.* random union of gametes), θ is also an estimator of H, the expected heterozygosity at a locus.

1.4.3 Effects of natural selection

Natural selection decreases genetic diversity around the selected locus. Ongoing negative selection against deleterious alleles purges those alleles and any linked variants from the population. Positive selection increases the frequency of a beneficial allele; as the allele increases in frequency, so do any linked variants on the same haplotype, leading to a local reduction in diversity termed a *selective sweep*[84]. The effect of selection — whether negative or positive — at a linked neutral locus

is mitigated by recombination between that locus and the target of selection.

There are two notable exception to these rules. The first is so-called *balancing selection*, in which selection favors the existence of multiple alleles in the population rather than any single allele *per se*. This may be the case at, for example, loci important in kin recognition and mate choice, for which diversity at the population level promotes the avoidance of inbreeding[85]. Loci under balancing selection and any linked haplotypes show an excess of polymorphism relative to neutral expectations. The second exception is *meiotic drive*, the transmission of alleles at non-Mendelian frequencies. An allele subject to meiotic drive may rise in frequency and crowd out other haplotypes, leaving a footprint indistinguishable from positive selection, while decreasing host fitness. An example of just such a *selfish sweep* is discussed in **Chapter 5**.

1.5 Methods for characterizing genetic variation

Empirical investigation of genetic variation in populations requires techniques to identify variant sites and genotype them in multiple individuals. In this section I briefly review some relevant methods for characterizing genetic variation, with a focus on methods for ascertaining variants at whole-genome scale with high throughput. I emphasize the strengths and weaknesses of different technologies and approaches rather than the details of their implementation.

1.5.1 The role of reference genomes

A *reference genome sequence*, or *reference assembly*, is a single, haploid, linear sequence taken to be (in some way) representative for the organism under study. Because genetic variation is intrinsically relative, a reference sequence facilitates description of genetic variants by providing both a common coordinate system for describing the location of alleles and a common baseline against which to describe their sequence content. Reference genomes vary widely in quality and completeness; the current references for mouse[55] and human[86] are the highest-quality references among vertebrates. Both were constructed by capillary sequencing of $700 - 1000$ bp reads from genomic fragments cloned into bacterial or yeast vectors. These reads were assembled into *contigs* on the basis of sequence, and contigs into *scaffolds* on the basis of orthogonal physical mapping techniques. Scaffolds were grouped into chromosomes and ordered on the basis of physical mapping and genetic mapping.

Although reference sequences are fundamental to the current practice of genomics, they may also introduce bias into interpretation of experimental and population data. This bias arises from

several sources. First, for repetitive sequences with high copy number and high mutual similarity — especially when the repeat period is longer than a sequencing read — there may not exist a single optimal linear representation that can be extracted from sequencing data. High-identity repeats may be collapsed to fewer than the true number of copies[87]. Second, some genomic regions segregating in the population may be completely absent from the reference even in mature assemblies such as human[36]. Third, regions that are present in the reference but have accumulated substantial divergence from the individual(s) under study may lead to difficulties in interpreting the true level of sequence polymorphism[88]. One of they main contributions of this thesis is to demonstrate that several classes of biologically-important sequences are poorly represented in the mouse reference assembly, and that interpreting variation in these regions requires looking beyond the reference genome.

1.5.2 Microarrays

Genotypes can be ascertained for thousands to a few million biallelic SNVs at a time in a single individual using oligonucleotide microarrays ("SNP chips.") Genotyping arrays have been used with great success in humans[89] and laboratory model organisms such as mouse[90,91,92], and the increasing availability of custom-designed platforms has expanded their utility to organisms of agricultural or ecological interest[93]. Although there exist several competing technologies, all genotyping arrays exploit the specificity of binding between synthetic olignonucleotide probes immobilized on a chip and complementary sequences in sample DNA washed over the chip's surface. Relative binding of DNA fragments with the two possible alleles (labelled A or B by convention) is converted to a fluorescence readout, and offline signal-processing algorithms are used to render one of three possible genotype calls (AA, AB, BB). Assuming that most target loci are present in the expected diploid copy number in a sample, the relative fluorescence intensity from the A and B alleles can be used to infer the presence of copy-number variants[94]. [2]

Genotyping arrays are cost-effective and robust but their utility for population genetics is

[2]Design of microarray platforms tailored to diverse laboratory and wild mouse populations is a major research activity in the Pardo-Manuel de Villena lab. The group participated in the design of the Affymetrix Mouse Diversity Array[92] (2008) and led the design of three versions of the Illumina Mouse Universal Genotyping Array (MUGA)[95] (2010, 2012, 2015). I contributed to the design and led the validation of the third generation of MUGA[95], and developed a software package for exploratory analysis of data from Illumina SNP arrays[96].

limited by the fact that they interrogate only known variants. When array content is ascertained in a representative sample with similar ancestry to the population under study, SNP genotypes provide an accurate estimate of haplotype diversity and are useful for assessing population structure and relatedness. Otherwise arrays are subject to ascertainment bias that may be quite strong[97,98]. Array genotypes are generally not appropriate for estimating quantities that depend on knowing the total number of segregating sites in a population or the full site-frequency spectrum, which includes many estimators for genetic diversity.

1.5.3 Whole-genome sequencing

Direct sequencing of unselected genomic DNA is the workhorse technique of modern genetics. Although it is common to refer to this family of assays as "whole-genome sequencing" (WGS), it is more accurate to describe them as *re-sequencing* because they sidestep the extensive assembly and validation efforts involved in the creation of a high-quality genome sequence. The aim of re-sequencing is inherently comparative: the end product is an alignment between the sequences of two or more chromosomes, from which variant sites and their respective alleles can be identified. WGS has transformed our understanding of genetic variation in the human population[27] and among mouse strains[29].

The typical WGS protocol involves four steps: (1) fragmentation of chromosomes into smaller pieces collectively known as a *library*; (2) generation of sequence reads from the fragments; (3) alignment of the reads to the reference, *or* assembly of the fragments *de novo*; and (4) identification of sequence variants between samples and/or the reference genome[99]. The first two steps are experimental and the last two purely computational. Nearly all sequencing protocols take a shotgun approach, randomly fragmenting the whole genome and and sequencing the entire library in parallel. With sufficient redundancy at the read-generation step, coverage of the entire genome by at least some reads can be assured to a reasonable approximation[100]. Redundancy between sequence reads also mitigates the effect of sequencing errors and allows accurate identification of heterozygous sites in diploid samples. The level of sequencing redundancy is expressed as a *coverage factor*: for example, $4\times$ coverage means that each site is covered by 4 reads, on average.

WGS protocols can be broadly divided into two classes: short-read and long-read. The dominant short-read platform at time of writing is Illumina. These instruments are capable of generating hundreds of millions to billions of sequencing reads in parallel, with lengths $50 - 250$ bp each

and per-base error rates on the order of 10^{-3}. Reads may be *single-end*, representing one end of a genomic fragment; or *paired-end*, representing opposite ends of a single fragment (with or without unknown sequence in between). Because of their short length, individual Illumina reads are relatively uninformative. The first step in the analysis of Illumina data is therefore to produce alignments of reads (read pairs) to a reference sequence. Variant sites are then identified from the read alignments of one or more individuals.

Short reads from complex, repeat-rich genomes like those of mammals have important limitations. First, unambiguous alignment of a read (read pair) to a single location in the reference genome may not be possible when the read originates from repeat sequence. The problem is exacerbated by sequencing error and by divergence between the reference genome and the sequenced sample. Second, even paired-end reads provide limited information about sequence organization at scales larger than a few kilobases. Deviations from expected read depth at a locus are informative for copy number, but more detailed characterization of the order, orientation and divergence between individual copies of repeats is rarely possible from short reads alone.

Long-read platforms include traditional capillary sequencing (low-throughput) and high-throughput platforms such as PacBio and Oxford Nanopore. Read lengths are not uniform and range from 700 bp (capillary) to 20 kb (PacBio). Unfortunately read length is inversely correlated with throughput, per-base accuracy, or both. I do not make use of long-read sequencing in this thesis, but I note here that long reads and short reads have complimentary properties. Long reads are extremely valuable for *de novo* assembly, especially in structurally-complex regions of the genome.

CHAPTER 2

Introduction II: The mouse as a model for genome evolution

The house mouse (*Mus musculus* Linnaeus 1758) — hereinafter simply "the mouse" — has been a workhorse of both basic and applied biomedical research since the beginning of the twentieth century. Mice have many favorable qualities for a model system: they are readily bred in captivity, require little space, have short generation time, and are amenable to experimental manipulation. Many aspects of mouse physiology and behavior are readily translated to humans. Mice have been instrumental in the study of diverse physiological phenomena: transplant rejection[101]; self-vs-nonself recognition in the immune system[102]; skin pigmentation and coat color[103,104,105]; sex determination[7]; dosage compensation on the X chromosome of females[106]; resistance to oncogenic viruses[107]; hormonal regulation of energy balance[108,109]; mutagenesis by ionizing radiation[110]; and many others.

Mice have been a particularly important model system in genetics. Early crosses between mouse stocks with different coat colors helped demonstrate that Mendel's laws applied to animals as well as plants and complemented ongoing work in *Drosophila* to demonstrate the chromosomal theory of heredity (reviewed in[111]). (In fact, Mendel's first experiments were with mice — until his superiors deemed them unfit company for a monk[112].) Establishment of the first inbred strains in the early 1900s facilitated studies of the mechanisms of inheritance and linkage. The mouse genome was only the second vertebrate genome to be sequenced, in 2002[55]. Since that time an extremely deep and continually-growing catalog of functional annotation has been overlaid on this reference sequence[113,114,115].

In addition to their merits as a laboratory organism, mice are also a valuable model for evolutionary studies[116]. Traditional laboratory strains span relatively little genetic variation, and with idiosyncratic distribution (discussed at length below), but wild mice and wild-derived strains substantially expand the space of genetic and phenotypic variation available to researchers[117,118,119]. The subspecies of the house mouse along with its sister taxa cover a more or less continuous

gradient of genetic and ecological differentiation from island populations split within the past three centuries[120,121,122] to reproductively-incompatible species separated by millions of years of evolution[123].

In this chapter I outline the evolutionary history of house mice and the relationships between wild mice and their laboratory relatives. The material reviewed here serves as a primer for discussions in the main chapters of this thesis. [1]

2.1 The mouse among rodents

Mus musculus is only one of more than 570 species of murid rodents (Old World mice and rats[124], **Figure 2.1A**). The subfamily Murinae is sister to gerbils and counts among its members the Old World rats (genus *Rattus*) and field mice (genus *Apodemus*) in addition to house mice. The genus *Mus* can be subdivided further into several subgenera which began to diverge approximately 9 million years ago (Mya). House mice are members of the subgenus *Mus* (**Figure 2.1B**). Their closest relatives, with last recent common ancestor approximately $1-2$ Mya, are a sister taxon consisting of *Mus spretus* (the Algerian feral mouse), *Mus spicilegus* (mound-building mice), *Mus cypriacus* (the Cypriot mouse) and *Mus spicilegus* (the Balkan short-tailed mouse)[125]. A clade of Asian species including *Mus caroli* (the rice-field mouse) and *Mus famulus* (the servant mouse) is somewhat more distantly-related, having a last common ancestor with house mice around $3-5$ Mya.

2.2 Ancestry and diversity of wild house mice

The house mouse began to diverge from its sister species approximately $1-2$ Mya. Fossil and genetic evidence indicates that its ancestral range lay in central Asia, an area roughly corresponding to modern-day India, Pakistan, Afghanistan and Iran[126,123,127,128,129]. Around $500,000$ years ago the species began to split into three genetically- and morphologically-differentated lineages[125,130]: *Mus musculus domesticus*, *M. m. musculus* and *M. m. castaneus*. (A fourth lineage, *M. m. molossinus*, is endemic to the Japanese archipelago and is best understood as a hybrid between *M. m. musculus* and *M. m. castaneus*[131].) Within the last $50,000$ years the lineages began to disperse from their

[1] For this section I am indebted to an excellent review published by John Didion, a recent alumnus of the Pardo-Manuel de Villena lab[118].

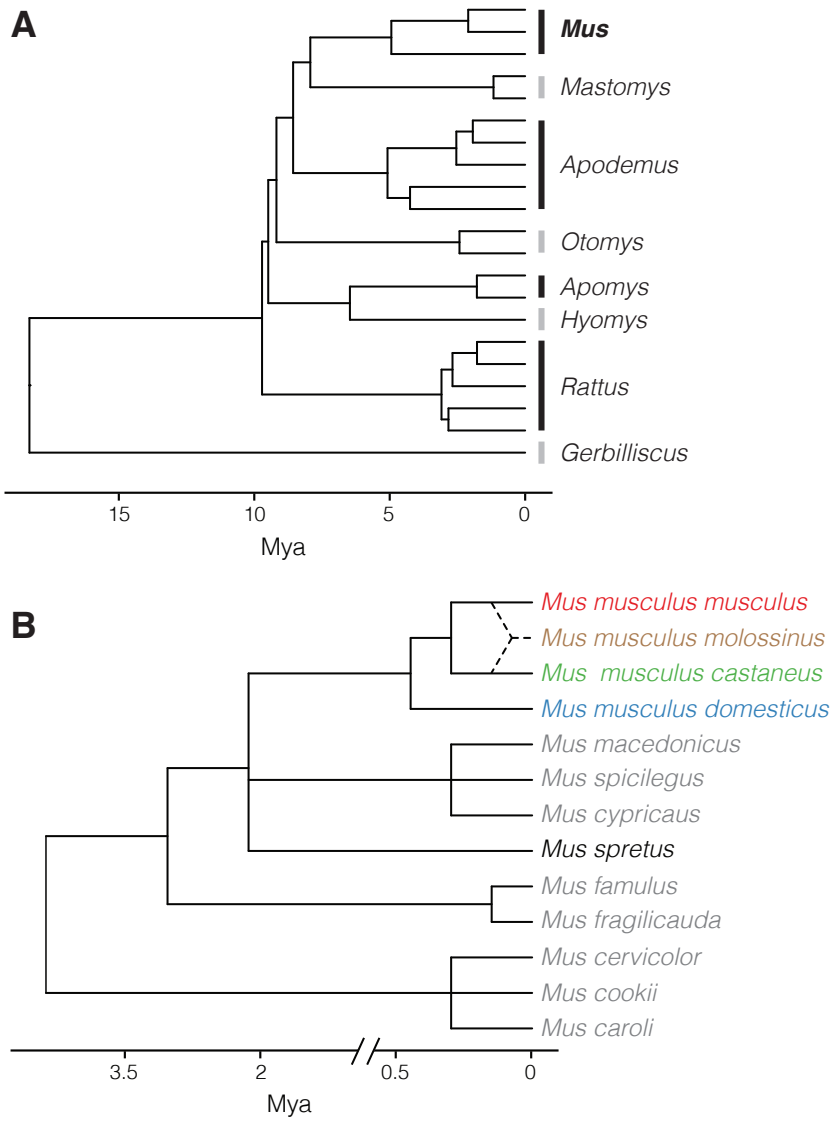

Figure 2.1: Phylogenetic tree of selected rodent genera (**A**) and of the *Mus* genus (**B**). Trees are drawn approximately to scale. Mya, millions of years ago.

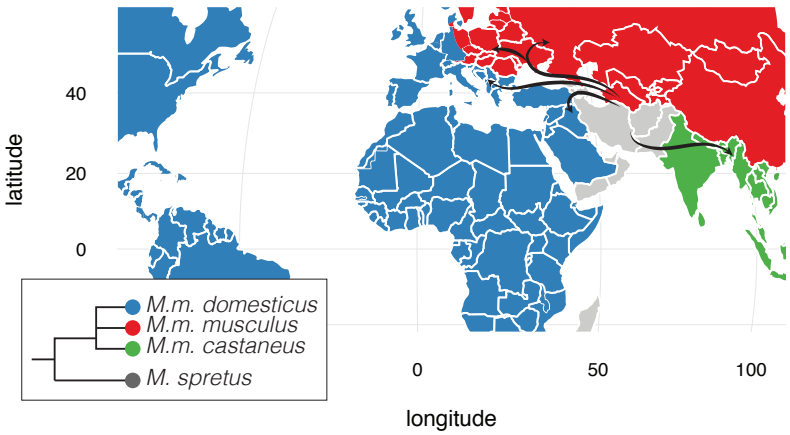

Figure 2.2: Dispersal of mouse subspecies from their ancestral range in the Near East and central Asia. Map is colored according to the dominant ancestry of wild mice in each region; populations in areas shaded in grey are either of uncertain taxonomic origin or significantly admixed. Adapted from[118].

ancestral range: *M. m. domesticus* into the Middle East, Europe and the Mediterranean basin; *M. m. musculus* to eastern Europe, Scandinavia and northern Asia; and *M. m. castaneus* to the Indian subcontinent and southeast Asia (**Figure 2.2**). During the same period house mice became commensal with humans, and their spread was facilitated by human migration[132]. Mice are now found on nearly every landmass visited by humans[133].

2.2.1 Taxonomic status of mouse lineages

Morphological and biochemical similarities between the major lineages of house mice and their close relatives have bedeviled systematists since the mid-twentieth century[118]. An examination of the literature on the phylogenetic history of mice reveals a confusing web of nomenclature, with lineages in *Mus musculus* alternately deemed "subspecies", "semi-species", full species or a "species complex" possibly extending to *Mus spretus* and *Mus spicilegus*. In keeping with the consensus among most authors, I refer to the three major lineages of house mice as *subspecies*[123], but some authors contend that the full species designation is more appropriate given reproductive barriers between lineages[134].

In any case it is clear that *incomplete lineage sorting* — discordance between phylogeny at a locus and species-level relationships — is widespread across the mouse genome[135,136,130], consistent with

nearly-simultaneous divergence of the three subspecies from each other. Discordance between matrilineal ancestry (inferred from mitochondrial DNA) and patrilineal ancestry (inferred from Y-linked markers) supports this view[137]. Likewise it is clear from transects in zones of secondary contact between subspecies that there is ongoing gene flow between subspecies[138,139].

2.2.2 Speciation and hybrid zones

Pairs of subspecies have come into secondary contact in several locations around the globe within the past several millenia, forming geographically-restricted *hybrid zones*. The best-studied hybrid zone lies along a narrow (< 30 km wide) strip which meanders from Denmark through Germany and the Czech Republic[132]. To the west lies the territory of *M. m. domesticus*, and to the east, *M. m. musculus*. Moving from west to east across the hybrid zone, the average ancestry of mice traces a smooth cline from pure *domesticus* to pure *musculus*[138].

Hybrid zones provide a natural experiment for studying the basis of reproductive isolation between incipient species. In the *domesticus-musculus* zone, hybrids of both sexes may have reduced fertility[140], but hybrid male sterility is the dominant mode of reproductive incompatibility[141,142]. The degree of sterility is variable in the laboratory and in nature[143,140,132]. Its proximate cause is defects in chromosome pairing and synapsis[142,144], leading to disruption of meiotic sex chromosome inactivation[145,146], arrest late in meiotic prophase and eventually germ cell death. Dozens of loci associated with sterility-related morphological, histological or functional phenotypes have been mapped in laboratory crosses[141,147,148,149] and wild populations[150,151]. The loci of largest effect are on the X chromosome[152] and on chromosome 17. The chromosome 17 locus was subsequently mapped to a single gene, *Prdm9*, which also serves as a master regulator of the rate and spatial distribution of recombination[153]. Links between speciation and the meiotic machinery are further explored in **Chapter 3** and **Chapter 6**.

2.3 Origins of laboratory mice

Laboratory mice, for all their experimental utility, are a synthetic construct with little resemblance to any single population that exists in nature[118]. Their origin is a tale of contingency and historical accident[154]. Understanding this history is critical to the well-informed design of experiments to test evolutionary hypotheses.

2.3.1 Ancestry of classical inbred strains

The majority of inbred strains and outbred stocks of mice used in laboratory studies can be traced to a small population of "fancy mice" propagated by hobbyist breeders in Japan and Europe through the late 1800s[155,156]. Small colony sizes and selective breeding meant that fancy mice were already fairly inbred[155]. These stocks formed the basis for the so-called *classical inbred strains* developed by the first mouse geneticists in the early twentieth century. Genetic diversity in the classical inbred strains is limited: over most of the genome, classical strains can be projected onto just 5 founder haplotypes, and 97% of the genome can be explained by no more than 10 haplotypes[157].

Classical laboratory mice and commercial outbred stocks have long been known to be three-way hybrids between the three major subspecies[137,158,159], but the first genome-wide studies of polymorphism in inbred strains led to conflicting inferences on the relative contribution of each subspecies[160,161,162]. That confusion was due almost entirely to faulty assumptions regarding the purity of a few strains used as phylogenetic reference points (see below). Subsequent analyses calibrated with data from wild-caught mice of known ancestry clearly demonstrated that classical strains are derived primarily from *M. m. domesticus* (94% of the genome) with minor contributions from *M. m. musculus* (5%) and *M. m. castaneus* ($< 1\%$)[157]. Furthermore, the non-*domesticus* regions are correlated across strains and represent regions inherited from Japanese *M. m. molossinus* and shared identical by descent (IBD) between extant strains.

2.3.2 Wild-derived strains

In contrast to classical inbred strains, so-called *wild-derived strains* are generated by inbreeding among mice trapped in a single geographic location. A wild-derived strain represents, in effect, one haploid draw from the pool of chromosomes in a given population of wild mice. Although some selection for docility and ease of husbandry is inevitable, wild-derived strains are much more representative than are classical strains of the phenotypic profile of wild mice and of the level of variation segregating within and between subspecies in nature[163,164].

Wild-derived strains are an important resource for testing evolutionary hypotheses in *Mus*[119]. However, progress in the field has been hampered by a reliance on the strong assumption that wild-derived strains are pure representatives of a single subspecies. In fact this is rarely the case: of 62 wild-derived strains analyzed with the $600,000$-SNP Mouse Diversity Array, only 9 had pure

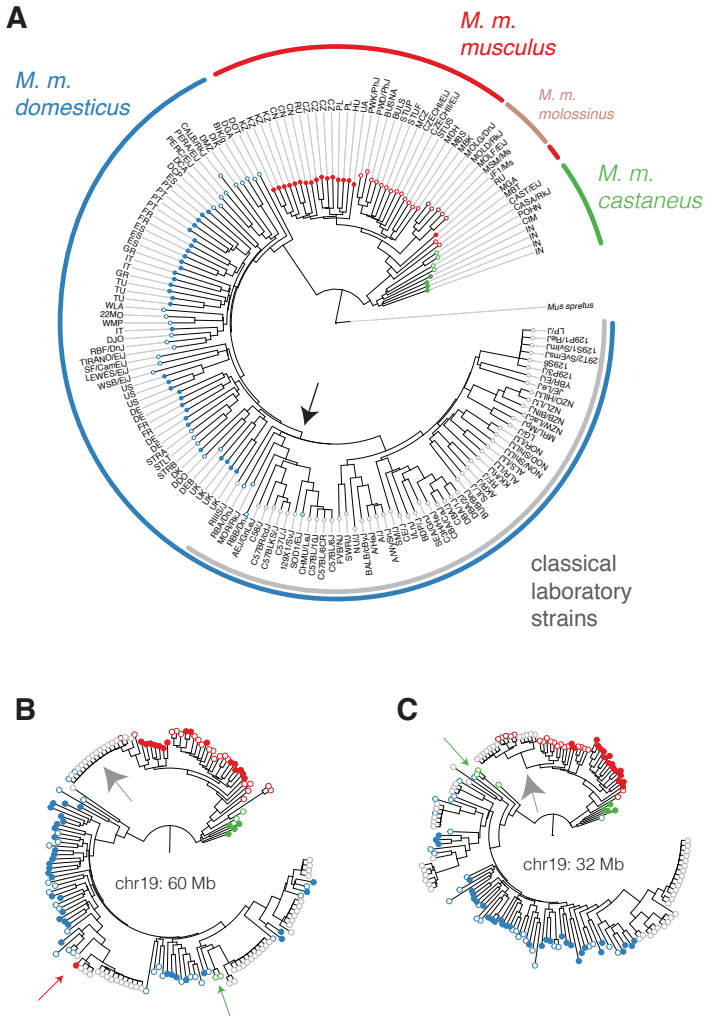

Figure 2.3: Global versus local phylogenetic relationships between laboratory strains and wild mice. (**A**) Phylogenetic tree for chromosome 19 showing relationships between laboratory inbred strains (open circles) and their wild relatives (closed circles). Inbred strains are labelled by strain name, and wild mice by their country of origin. Arrowhead points to the last common ancestor of classical laboratory strains. Tree was constructed from genotypes at $13,739$ biallelic SNP markers on the Mouse Diversity Array[92,157] and rooted rooted using five representatives of *Mus spretus* as an outgroup (not shown.). (**B,C**) Local trees showing evidence for gene flow or incomplete lineage sorting. Grey arrowheads, IBD between classical inbred strains and *M. m. molossinus*; green and red arrowheads, discordant samples from *M. m. castaneus* and *M. m. musculus*, respectively.

ancestry and 35 had some contribution from all three subspecies[157]. Inter-subspecific introgression in wild-derived strains — whether due to naturally-occurring gene flow or contamination in the laboratory — confounded earlier estimates of the relative contribution of each subspecies to classical laboratory strains[161,29] and created the false impression that levels of inter-subspecific differentiation were extremely variable along the genome[165,166].

The phylogenetic relationship between classical inbred strains, wild-derived strains and a representative sample of wild mice is illustrated in **Figure 2.3**. Panel A represents the maximum-likelihood toplogy over 61 Mb of sequence from chromosome 19; at this level, the three cardinal subspecies are readily differentiated. But at local scale, the expected relationships do not necessarily hold. Discordance between the global and local phylogeny may arise for three reasons: (1) contamination by accidental cross-breeding in the laboratory, in the case of inbred strains; (2) admixture in the wild; or (3) incomplete sorting along subspecies lineages of haplotypes polymorphic in the ancestral population of *M. musculus*.

CHAPTER 3

Structural variation and recombination in the mouse germline

3.1 Introduction

[1] Recombination — the exchange of genetic material between homologous chromosomes, which is the essence of sex — is a ubiquitous feature of meiosis in eukaryotes. Although the origin of sex remains one of the great mysteries (and controversies) of evolutionary biology[167], it is clear from its persistence across the tree of life that it must serve extremely important roles in individual fitness and the maintenance of genetic diversity in populations[168].

Recombination and its dual, genetic linkage, were discovered in the early twentieth century as exceptions to the Mendelian principles of segregation and independent assortment[169]. William Bateson, Reginald Punnett, T.H. Morgan and others had identified groups of traits whose inheritance appeared to be "coupled" to varying degrees. Morgan offered the elegant interpretation that coupling arose because the underlying "factors" were physically located on the same chromosome, and proposed that the cytological phenomenon of crossing-over could explain the degree of coupling. This and subsequent experiments by Sturtevant[170] solidified the chromosome theory of heredity and provided evidence that "factors" — what we would now call genes — are linearly arranged. Sturtevant made the important observation that the frequency of recombination between pairs of genetic markers could be interpreted as distance (in the metric sense) and used this fact to construct the first *linkage map*, in *Drosophila melanogaster*[170]. Linkage maps have since proven invaluable as tools for understanding genome organization.

[1] A portion of the results presented in this chapter are published in:

> Liu EY*, Morgan AP*, Chesler EJ, Wang W, Churchill GA, Pardo-Manuel de Villena F. High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. *Genetics*: **197**: 91–106. PMID 24578350.

Important contributions were made by Eric Yi Liu, Wei Wang and Gary Churchill.

In most eukaryotes, the formation of crossovers between homologous chromosomes provides mechanical support to allow the pair to orient perpendicular on the meiotic spindle and be segregated properly to daughter cells at meiosis I. (There are exceptions — meiosis in male *Drosophila melanogaster* is achiasmate, for example — but they are rare.) Failure to form crossovers and to place them appropriately along chromosomes is associated with aneuploidies which are invariably deleterious[11,171]. Meiotic recombination also provides an opportunity to repair accumulated damage to germline DNA[172]. From these points of view, recombination is critical to the maintenance of genome stability in the germline and subsequently in gametes.

But recombination also has important roles at the population level. By allowing beneficial mutations to dissociate from linked deleterious mutations and to then associate with beneficial mutations on other haplotypes, recombination mitigates the phenomenon known as Hill-Robertson interference[2]. The net effect is to allow the population to achieve higher mean fitness. Likewise, recombination limits the extent of *selective sweeps* — the loss of genetic diversity at sites linked to a mutation under positive selection[84]. Recombination thus enhances both the efficiency of selection in the present and the population's capacity to adapt in the future.

Despite the centrality of recombination in meiosis and its evolutionary conservation in general, the overall rate and spatial distribution of recombination vary widely between and even within species. In mammals this variation is intimately linked to the formation of reproductive barriers that separate species[173], a process that has been characterized in detail in mouse[141,147,144]. Recombination-rate variation in mouse is heritable[174] and under selection[175,176].

In this chapter we present a comprehensive analysis of crossover recombination in several experimental populations of laboratory mice. Our results replicate several well-known results regarding sexual dimorphism in the rate and distribution of crossovers in mammals. We make several novel observations:

- Both the X and Y chromosomes harbor modifiers of the overall recombination rate with large effect sizes that are segregating between and within mouse subspecies.
- Advanced paternal age is associated with an increase in the overall recombination rate in multiple and diverse genetic backgrounds. The effect is independent of crossover interference.
- Crossovers are strongly suppressed around clusters of large copy-number variants (CNVs)

that we term *coldspots*. Haplotypes with equal copy number at a given coldspot are only marginally more likely to recombine than haplotypes with different copy number, suggesting that these loci represent complex rearrangements and not simple changes in dosage. Coldspots have a transcriptional and epigenetic profile in male germ cells consistent with closed chromatin.

The remainder of this section introduces the molecular events and possible outcomes of meiotic recombination; genetic and non-genetic factors influencing the overall rate and spatial distribution of crossovers; and experimental approaches for studying recombination in mammals, with a focus on the mouse populations used in the present investigation.

3.1.1 Molecular basis of recombination

Meiotic recombination begins during the earliest substage of the first meiotic prophase with the programmed introduction of double-strand breaks (DSBs)[177]. (Refer to diagrams in **Figure 3.1**.) These DSBs will subsequently be repaired using the homologous chromosome as a template, leading either to a *crossover* (CO) — if the repair involves exchange of flanking regions of the chromosome — or a *non-crossover* (NCO) if it does not. Because the repair process is not symmetric with respect to the chromatids involved, both COs and NCOs also entail *gene conversion*. In mouse, approximately 200 DSBs occur per meiosis, of which $20 - 30$ will lead to COs and the remainder to NCO products[1].

The key steps in recombination are synchronized with the gathering and organization of chromosomes before the first meiotic division. DSBs occur after the pre-meiotic round of replication but before homologous chromosomes pair; in fact, DSBs facilitate the "homology search" by which homologs find their partners, and are required for pairing[178]. Before and during the formation of DSBs, the telomeric ends of chromosomes gather into a structure known as the "bouquet" — conserved across the eukaryotes — that anchors them to the nuclear envelope and is thought to facilitate pairing[179,180]. Pairing is followed by synapsis, the formation of the proteinaceous synaptonemal complex (SC) between the homologs (the *bivalent*). At this stage chromatin is arranged in loops projecting outward from the SC. These loops are tethered to the SC by a suite of proteins that bind to DSBs[181]. The SC provides a physical scaffold that facilitates repair of DSBs by the homologous recombination machinery. (Although the fine biochemical details of recombination

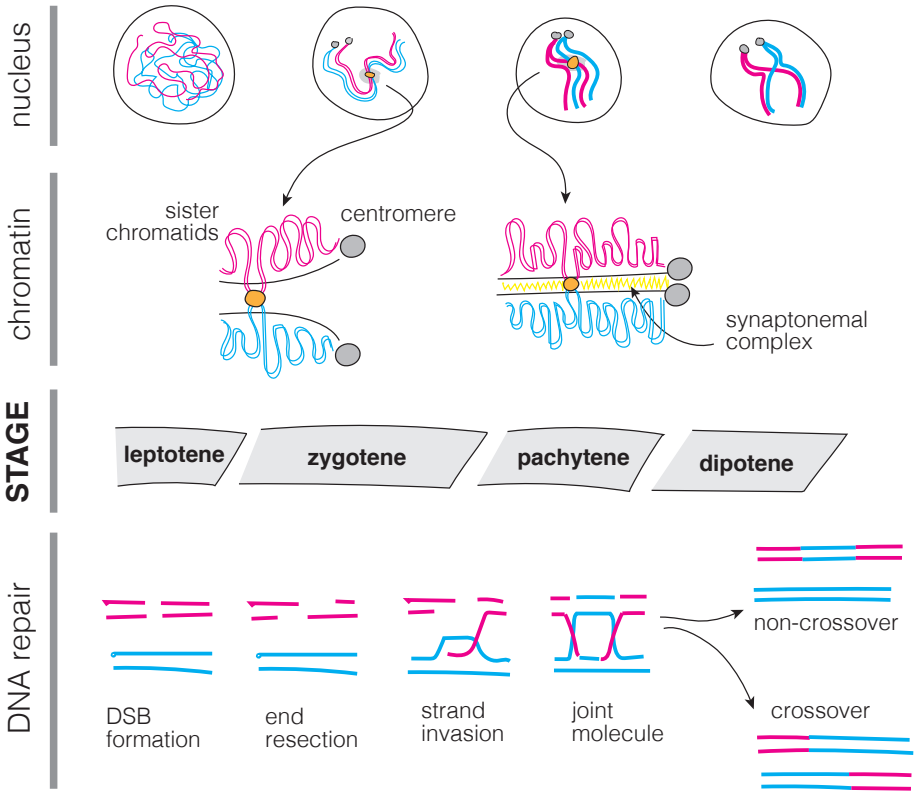

Figure 3.1: Nuclear appearance, chromatin configuration and steps in double-strand break (DSB) formation and repair as crossovers or non-crossovers during the first meiotic prophase. Substages are named in the center track. Orange dot marks the site of a DSB which will be processed into a crossover or non-crossover. Figure adapted from[167].

are beyond the scope of this chapter, it suffices to say that DSB repair proceeds through a "joint molecule" that involves one or more chromatids from each member of the homologous pair.) When DSB repair is complete the SC dissociates, and homologs are physically attached via *chiasmata* at the sites of COs. Chiasmata persist until the end of the first prophase.

Because recombination occurs within the context of meiosis, male and female mammals differ in several important respects. Whereas the primary germ cells in males (primary spermatocytes) do not enter meiosis until the time of sexual maturity, primary germ cells in females (primary oocytes) enter meiosis *in utero*. Primary oocytes arrest and enter a resting state ("dictyate") that is maintained until fertilization. Chiasmata thus persist for up to years in mouse or decades in

longer-lived mammals. By contrast, male meiosis proceeds from start to finish within about 30 days in mouse. On the other hand quality-control processes are much stricter in male than female meiosis. Failures of pairing, synapsis or recombination generally lead to arrest and germ cell death before the end of the first prophase in spermatocytes but not oocytes[1]. Consequently the vast majority of mature sperm are euploid in humans, while as many as 20% of oocytes are aneuploid[182].

3.1.2 Broad-scale control of recombination

Although the existence of a true homeostatic mechanism for regulating the total number of crossovers is still a matter of debate[183], the general requirement for a minimum of one crossover per bivalent has been apparent for decades[184]. In mammals, the number of chromosome arms — not chromosomes — predicts the length of the genetic map, suggesting that there is a requirement for one crossover per chromosome arm[13]. The location of the obligate crossover is important: too near the centromere or telomere predisposes to nondisjunction[171].

When multiple crossovers occur on a single chromosome (arm), they tend to occur further apart than would be predicted if crossing-over operates as a memoryless uniform process along chromosomes. This pattern, known as *interference* (strictly speaking, positive interference), followed quickly from the first models of recombination[185] but its mechanistic basis remains poorly understood. In yeast and plants[186], a minority of crossovers are exempt from interference; because the proteins involved in their execution are conserved in vertebrates, the same is likely true in mouse. Weaker interference in females is thought to contribute to the higher number of crossovers observed in female meiosis in mouse[187].

The overall recombination rate, as measured by the number of crossovers per meiosis (*i.e.* the total length of the genetic map), is generally higher in females than in males[188,189,190] — although there are exceptions, including sheep[191] and opossum[192]. The distribution of crossovers along chromosome arms also differs between the sexes: crossovers in males are enriched in the distal portion of chromosomes and depleted near the centromere in many mammal species including human[189], mouse[190], dog[193] and cow[194].

Males face a distinct challenge in fulfilling the requirement for one chiasma per bivalent: the sex chromosomes. Mammalian sex chromosomes are heteromorphic share homology only at the *pseudoautosomal region* (PAR) at the distal tip of the long arms[195]. (Mice, having an afrocentric karyotype, have one PAR; humans have two, one for each chromosome arm on the X and Y.) Synapsis

on the sex chromosomes is restricted to the PAR, and DSB repair in the PAR occurs later and in a distinct nuclear domain (the *sex body*) from the recombination process on the autosomes[20,196]. Although the sequence of the PAR in the mouse reference genome assembly is still incomplete, it is thought to be at most a megabase in size[197]. Its local recombination rate per base par is thus $10 - 50$ times greater than the genomic average.

Recombination rate is heritable and varies between laboratory strains of mice[174] and between mouse subspecies[175]. Wild-derived inbred strains of predominantly *M. m. musculus* ancestry have recombination rates approximately 30% higher than strains of *M. m. castaneus* or *M. m. domesticus* ancestry. Nearly half of this variation maps to a single locus on the X chromosome in an F_2 cross between CAST/EiJ (*M. m. castaneus*) and PWD/PhJ (*M. m. musculus*)[175]. This finding is supported by differences in local recombination rates on several chromosomes have been detected in reciprocal F_1 hybrids between WSB/EiJ (*M. m. domesticus*) and CAST/EiJ. In crosses between *M. m. domesticus* and *M. m. musculus* wild-derived strains, surrogate phenotypes for hybrid sterility map to the same loci as recombination rate on chromosomes 17 and X[142,152,198].

3.1.3 Fine-scale control of recombination

Early models for recombination posited that crossovers arose via a continuous process along chromosomes or segments thereof. The ability to rapidly type dense panels of genetic markers in populations, in large pedigrees and in pools of sperm has radically altered that view over the past decade[173]. Deep characterization of recombination products at single loci[199] and analyses of patterns of linkage disequilibrium (LD) in unrelated individuals[200] have demonstrated that, at killable scale, the rate of recombination is extremely heterogeneous along the genome. The vast majority of crossovers occur in short (< 1 kb), discrete regions — *hotspots* — spaced tens of kilobases apart, at least in mouse and in human. The background rate of recombination is essentially zero.

Hotspots in primates and mice are defined by a degenerate 13 bp motif (`CCNCCNTNNCCNC`) that serves as the recognition sequence for PRDM9[201], a meiosis-specific enzyme that trimethylates the histone 3 lysine 4 residue (H3K4me3)[202,153,203]. This mark is the activation signal for a hotspot[204]. The structure of PRDM9 contains a tandem array of 12 zinc finger domains whose sequences are highly polymorphic in populations and rapidly evolving between species[205]. The sequences of the zinc fingers determines the specificity of a PRDM9 allelic variant for particular sequence(s) of the degenerate binding motif. Motifs which bind PRDM9 more avidly have higher activity for

both COs and NCOs[206,207]. However, the asymmetric nature of DNA repair in meiosis leads to degradation of the most active hotspots[208], such that PRDM9 and its cognate binding sites are locked in a cycle of repeated evolutionary turnover[209]. Besides PRDM9, the activity of hotspots and their propensity for CO versus NCO outcomes may depend on sex and as-yet known modifiers in the genetic background[210,211].

In taxa that lack a functional PRDM9 ortholog — including birds[212], canids (including the domestic dog)[213] and some yeasts[214] — the recombination landscape is more stable over time. The distribution of recombination is still non-uniform, but is directed towards conserved sequence features (such as promoter regions[215]) rather than labile hotspot motifs.

Curiously, allelic differences in *Prdm9* between *M. m. musculus* and *M. m. doemsticus* are responsible for male sterility in hybrids[216]. The testes of affected males are characterized by meiotic arrest and germ cell death late in meiotic prophase, due to defects in DSB repair, homolog pairing and formation of the sex body[141]. A similar phenotype is observed in a *Prdm9* knockout[202]. Re-targeting PRDM9 to ancestral hotspots via engineering of its zinc-finger arrays reverses sterility[217], directly implicating the co-evolution of PRDM9 and hotspot motifs in the formation of reproductive barriers.

3.1.4 Methods for studying recombination

Approaches to the study of recombination can be divided into three categories: genotyping informative markers in pedigrees; observation of cytological markers for crossovers; and analyses of linkage disequilibrium in populations of unrelated individuals. Each has advantages and disadvantages, which we discuss briefly below.

- *Pedigrees.* Crossovers can be inferred from offspring genotypes when at least one parent is heterozygous. Recombination fractions among progeny are converted to genetic distances via a mapping function to account for interference and the possibility of multiple crossovers between informative markers. The resolution of linkage maps constructed from pedigrees depends on the density of the marker panel — which, in the limit, depends on the level of segregating variation in the parents — and on the number of informative meioses. Pedigree analyses are equally useful for male and female meiosis. However, because the segregation of the recombinant versus non-recombinant chromatids in a four-strand bundle is a stochastic

process, the variance of crossover counts derived from pedigrees is higher than from direct cytological observation.

- *Cytological markers.* Chiasmata can be directly observed under a microscope late in the first meiotic prophase, and can be counted in multiple germ cells per individual to improve the precision of the estimate of recombination rate. However, these methods have limited precision for localizing crossovers and are exceedingly difficult to apply in females due to the developmental timing of crossing-over in the ovary. More generally, antibody pulldown and high-throughput sequencing (ChIP-seq) of sequence fragments bound to key proteins in the recombination process allows the characterization of the number and spatial distribution of recombination intermediates at nucleotide resolution[218]. Interpretation of data from both cytological assays and ChIP-seq relies on the assumption that the distribution of recombination intermediates is an unbiased estimator of the distribution of transmitted crossovers — an assumption which may or may not be met[219].

- *Linkage disequilibrium.* Given sufficient sample size and marker density, population-level polymorphism data can offer an extremely fine-grained view of the (relative) rate and distribution of meiotic recombination along the genome. Polymorphism data capture many thousands of generations of informative meioses, averaged over both sexes. However, LD patterns are confounded with other population-genetic processes besides recombination which may or may not act uniformly along the genome. Recombination rates estimated from LD are fairly robust to model misspecification in simulations, but the degree of departure from their assumptions in real data is unknown *a priori*[219].

3.1.5 The Collaborative Cross and Diversity Outbred populations

Multiparental populations (MPPs) of model organisms are designed exploit the stochastic nature of meiosis to perform a factorial experiment. Breeding schemes for MPPs combine diverse founder genomes at equal frequency and, via recombination, shuffle them into a collection of random mosaics free of population stratification. MPPs thus provides a unique opportunity to study the rate and distribution of recombination against a randomized genetic background.

In this chapter we analyze two mouse MPPs: the Collaborative Cross (CC)[220,221], a panel of recombinant inbred lines derived from eight founder strains; and the Diversity Outbred (DO), an

outbred stock derived from a subset of partially-inbred CC lines[222]. The CC and DO were originally envisioned as platforms for high-resolution genetic mapping of loci underlying quantitative traits. They were designed to address two important shortcomings of existing panels of inbred strains and recombinant inbred lines: low levels of genetic diversity and long blocks of identity by descent among classical inbred strains[157], and cryptic population structure and pervasive long-range LD arising from the idiosyncratic ancestry of these same strains[223]. The eight founder strains for the CC and DO comprise five classical inbred strains of admixed but primarily *M. m. domesticus* ancestry — A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ and NZO/HlLtJ — and three wild-derived strains — CAST/EiJ (*M. m. castaneus*), PWK/PhJ (*M. m. domesticus*) and WSB/EiJ (*M. m. domesticus*). The wild-derived strains contribute vast majority of segregating variants. Of note, there are four *Prdm9* alleles segregating in the CC and DO: *msc* (PWK/PhJ), *cst* (CAST/EiJ), *dom2* (A/J, C57BL/6J, 129S1/SvImJ, NZO/HlLtJ) and *dom3* (NOD/ShiLtJ, WSB/EiJ).

The breeding schemes for the CC and DO are shown in **Figure 3.2**. CC lines were initiated using a "funnel" scheme: two generations of outcrossing, followed by 25 or more generations of inbreeding by sibling mating. Every realization of a funnel yields a unique mosaic of the eight founder haplotypes, but crossovers are constrained by the order of founders in the first generation. Each CC line was therefore assigned a unique funnel order to achieve full randomization, subject to additional constraints to maintain balanced contributions on the sex chromosomes and mitochondrial genome. A set of 144 CC lines at inbreeding generations 4 through 12 (median 5) were selected to found the DO. The DO is maintained by random mating in synchronized generations, 175 pairs per generation, with avoidance of sibling and first-cousin matings. Each mating pair contributes exactly two offspring to the next breeding generation; additional progeny are distributed to investigators for experiments.

3.2 Results

3.2.1 The CC and DO provide complementary views of recombination

The breeding scheme of the Collaborative Cross ensures that every crossover arising in the first two breeding generations can only have arisen during exactly one of six meioses (**Figure 3.3**). By genotyping a sibling pair at the G_2:F_1 generation, the products of eight meioses can be observed — four meioses (in G_1) shared between the members of the pair, and two unique to each sibling

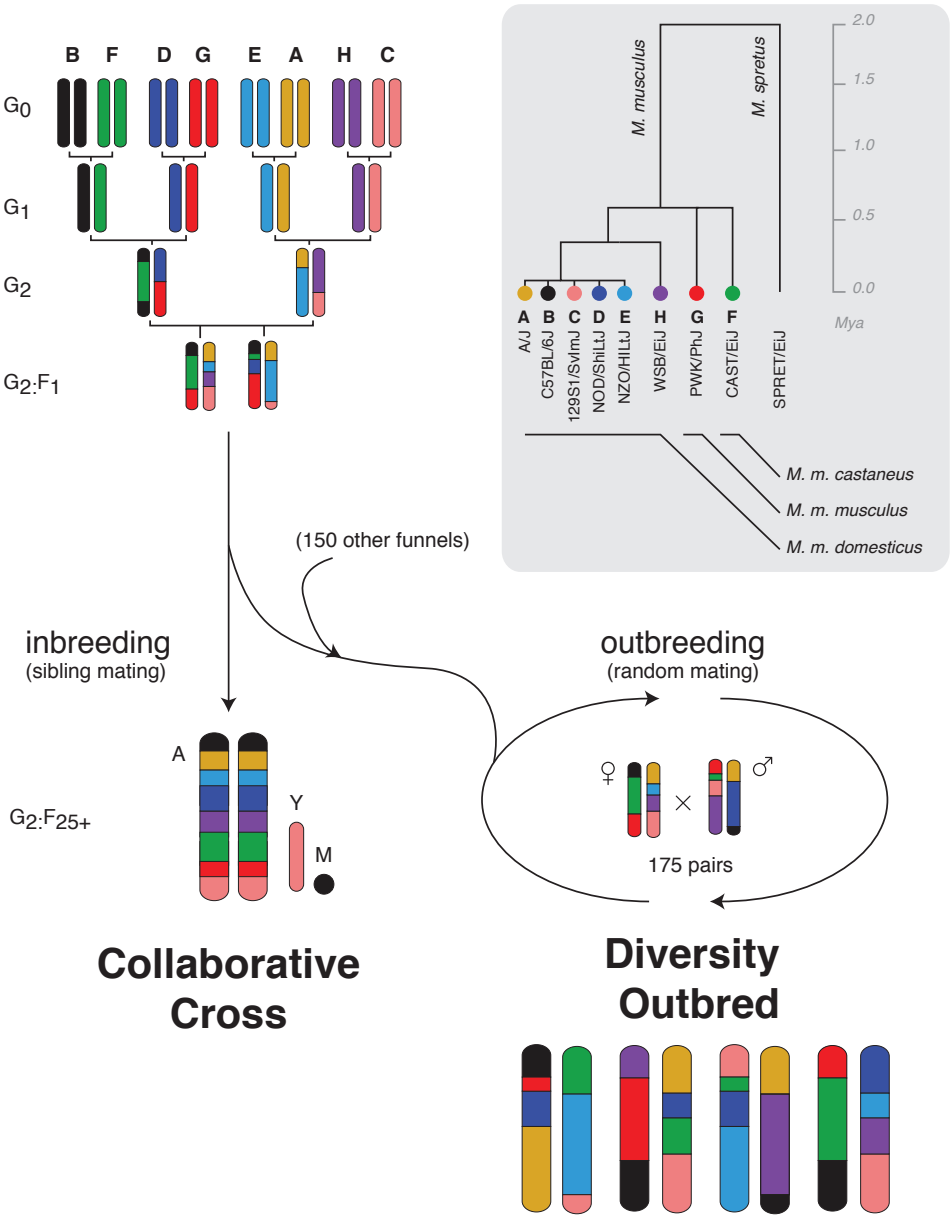

Figure 3.2: Breeding schemes for the Collaborative Cross (CC) and Diversity Outbred (DO) mouse populations. Both populations share the same set of eight founder strains, whose approximate phylogenetic relationship is shown in the inset panel. Throughout this chapter we refer to strains by one-letter codes for brevity. Each CC recombinant inbred line was initiated with two rounds of outcrossing in a "funnel" scheme (left), followed by many generations of sibling-mating to reach inbred status. The DO was seeded from 144 CC funnels early in the inbreeding process (right). It is maintained by random mating with 175 pairs per generation.

(in G_2). Because each founder strain is represented only once among the eight G_0 ancestors in each funnel, the G_1 and G_2 are obligate heterozygotes and all meioses are maximally informative. Randomization of the order of founder strains in each funnel ensures a balanced contribution from each founder genome to the autosomes of the G_2:F_1 generation, and an equal number of informative male and female meioses. Crossovers arising in G_2 are always observable, provided they are transmitted to offspring, but crossovers in G_1 are only observable if they are transmitted both at G_1 and in at least one G_2 meiosis. The probability that an inherited crossover on the autosomes is transmitted through at least one G_2 meioses in a given parent is $\frac{3}{4}$, so the expected number of observable meioses per funnel is $(\frac{3}{4})(4\,G_1 \text{ meioses}) + 1(4\,G_2 \text{ meioses}) = 7$. By a similar logic, the expected number of observable meioses per funnel on the X chromosome is $\frac{7}{2}$.

We genotyped a sibling pair from each of 237 funnels for a total of 474 G_2:F_1 offspring (1659 meioses) and augmented these genotypes with funnel information to infer fully-phased founder haplotypes (see **§ 3.5**). A total of $25,038$ crossovers were identified, of which $18,948$ were singletons and $3,045$ (all from G_1) were observed twice (*i.e* were shared by members of a sib pair.) Among the $21,993$ distinct crossovers, $21,368$ occurred on autosomes and 625 occurred on the X chromomsome for a sex-averaged map length of $1,288$ cM for the autosomes and 75 cM for the autosomes. The autosomal map is 6.7% shorter than the standard genetic map for the mouse[190], and the X chromosome map 1.5% shorter.

The CC breeding scheme provides a rich set of expectations and constraints that we used to confirm the integrity of our set of inferred crossovers. Several of these are shown in **Table 3.1**. In all cases, the observed crossover counts are in close agreement with Mendelian expectations.

Although less straightforward than simply genotyping parent-offspring duos, the CC G_2:F_1 is conceptually not different from a traditional pedigree study of linkage. Crossovers are inferred directly from individual-level genotypes and can be assigned to a specific parent's germline, allowing the construction of sex-specific genetic maps. The resolution of the sex-averaged map — 1 crossover per 114 kb on average — is sufficient to reveal patterns of local variation in recombination rate at the megabase scale, as discussed further below. But the map is still too sparse to examine fine-scale patterns such as hotspot usage.

For a more fine-grained view of recombination in the same outbred background as the CC G_2:F_1 we turned to the Diversity Outbred (DO) population. The DO comprises (at time of this study) 21

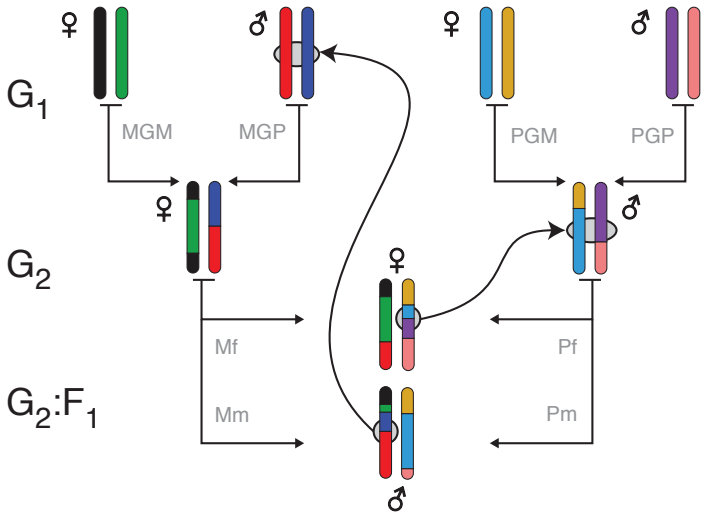

Figure 3.3: G_2:F_1 mice were bred in 237 funnels uniquely defined by the ordering of founder strains in the parental generation. This design balances the contribution of each of the eight founder strains through both the maternal and the paternal lineage. Using funnel order and genotypes for both members of a sib pair, it is possible to assign crossover events inferred in members of the sib pair to one of eight meioses occurring in the germline of a specific ancestor: in the G_1 generation, maternal grandmother (MGM), maternal grandfather (MGP), paternal grandmother (PGM), or paternal grandfather (PGP); or in the G_2 generation, mother (Mf, Mm) or father (Pf, Pm), with independent G_2 meioses contributing to each sib pair. Crossovers arising in G_1 may be shared between members of the sib pair, but crossovers at G_2 are always distinct.

Relationship	Expected	Observed
Observed events in G_1 vs. G_2	$3:4$	$0.747 \quad (p = 0.78)$
Same event type from opposite half-funnels	$1:1$	$1.00 \quad (p = 1.00)$
Mf vs. Mm		1.02
Pf vs. Pm		1.03
MGM vs. PGM		0.975
MGP vs. PGP		0.999
Singleton vs. shared events in G_1	$2:1$	$0.673 \quad (p = 0.17)$
MGM		0.678
MGP		0.672
PGM		0.675
PGP		0.668

Table 3.1: Expectations regarding autosomal recombination in G_2:F_1 pedigrees based on Mendelian rules. All p-values were obtained via a χ_2-test with a single degree of freedom.

generations with 175 breeding pairs ($N = 350$) per generation and thus provides at least tenfold more resolution than the G_2:F_1. But unlike the CC G_2:F_1, crossovers in the DO cannot in general be ascribed to a specific meiosis. There is no simple relationship between observed crossovers and genetic map length in the DO. Crossovers observed in a sample of n DO mice at generation k represent the cumulative products of up to $2N(k-1)+2n$ meioses. The expected number of observed meioses M is much lower due to the stochastic nature of inheritance in the population and of the sampling process, and in fact $M \approx 2n + O(\frac{1}{n})$ — that is, the great majority of crossovers observed in a sample are of recent origin. (See **Chapter A** for derivation and discussion.) But by sampling mice at many generations along the pedigree, we both obtain a very dense sex-averaged map and to observe temporal patterns in the accumulation of crossovers in the population.

We aggregated genotype data from $6,886$ individuals from sixteen breeding generations of the DO[2] and identified $2,242,658$ crossovers (see **§ 3.5**) arising in approximately $15,832$ observed

[2]Credit for this effort is due to Dan Gatti (The Jackson Laboratory) and more than a dozen investigators who contributed

meioses. As expected, the number of crossovers per genome increases linearly over time (at a rate of 27.1 crossovers per genome per generation; **Figure 3.4**), and the size of haplotype blocks decreases — although not linearly.

We used SNP genotypes from the autosomes to estimate pairwise kinship coefficients ($\hat{\pi}$) within generations and from these inferred close pedigree relationships. Knowledge of close relationships is useful for discriminating crossovers shared by descent versus recurrent crossovers, and for improving chromosome phasing. As expected, only a tiny minority of within-generation pairs ($16,335$ of $2,605,768$; 0.63%) can be detected as relatives at a threshold of $\hat{\pi} > 0.125$ and the remainder are effectively unrelated ($\hat{\pi} \approx 0$). Among these, $10,491$ pairs have $\hat{\pi} > 0.2$ and represent possible sibships. The distribution of estimated kinship coefficients by generation is shown in **Figure 3.5**. For the generations with the large sample sizes ($7, 8, 11$) the distribution clearly has two non-zero modes corresponding to relationships of degree 3 (cousins) and 2 (siblings and double-cousins).

Chromosome segments shared identical by descent (IBD) in related pairs were used to prune the set of 2.2 million total crossovers to a set of $749,560$ distinct crossovers (**Figure 3.6**). Among distinct crossovers, $570,113$ were singletons (observed exactly once, private to an individual) and $179,447$ were observed multiple times. To confirm the robustness of our classification of crossovers as shared, we compared estimated kinship from SNP genotypes to the proportion of non-singleton autosomal crossovers apparently shared between members of the same pair and obtained reasonably close agreement (**Figure 3.7**). For a subset of $1,529$ mice corresponding to 508 known sibships, we compared the kinship estimate from shared crossovers to its expected value $\pi = 0.25$. Our estimate from shared crossovers is $\hat{\pi} = 0.225$, or 10% below the expected value. The degree of underestimation is related to the proportion of recombinations that are private, a value which itself decreases with increasing sample size. This suggests that while our power to detect a

genotype data. See also

> Chesler EJ, Gatti DM, Morgan AP, Strobel M, Trepanier L, Oberbeck D, McWeeney S, Hitzemann R, Ferris M, McMullan R, Clayshultle A, Bell TA, Pardo-Manuel de Villena F, Churchill GA (2016) Diversity Outbred Mice at 21: maintaining allelic variation in the face of selection. *G3* early online publication. PMID 27694113.

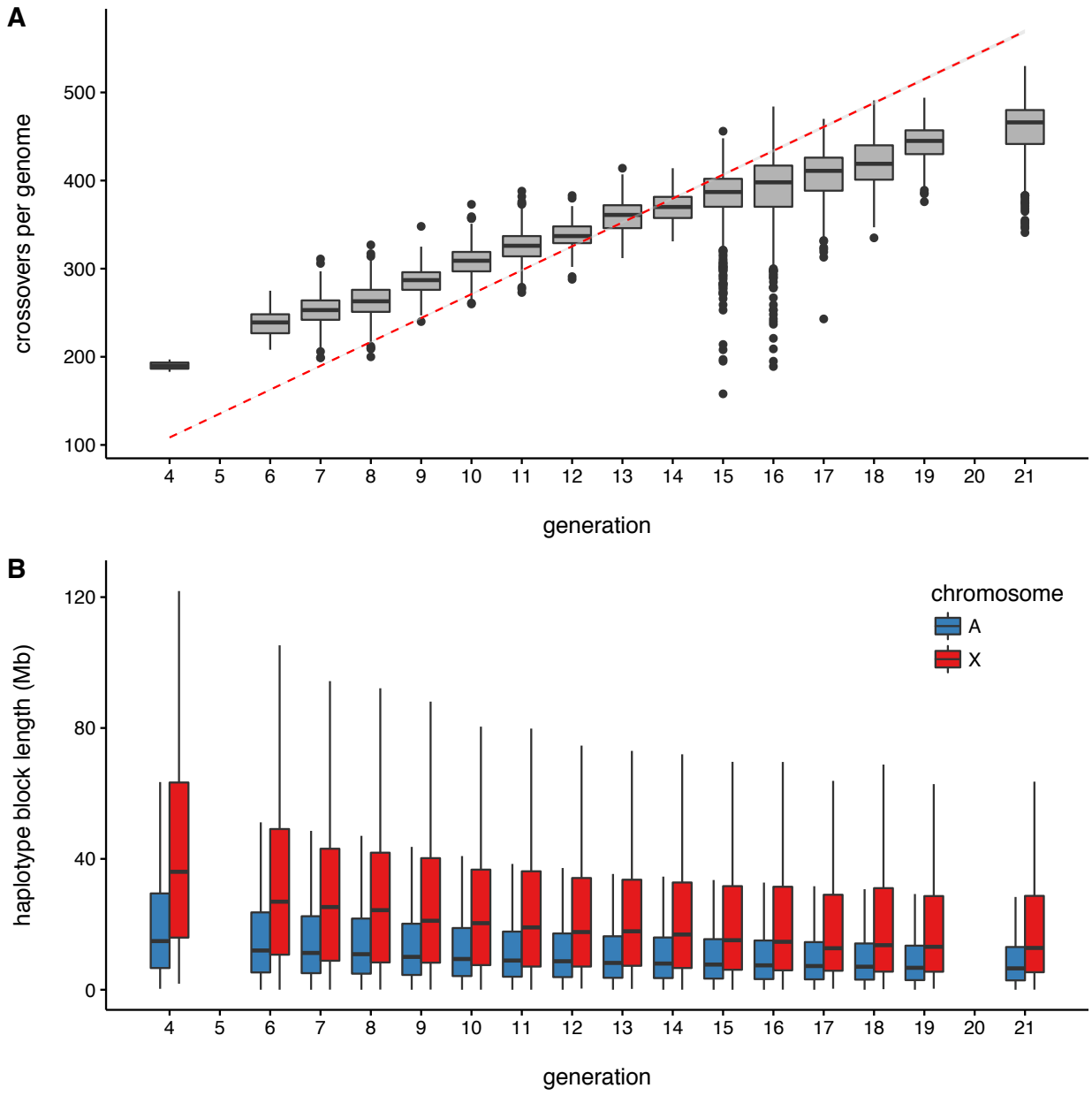

Figure 3.4: Accumulation of crossovers in the genomes of Diversity Outbred (DO) mice. (**A**) Distribution of the number of observed crossovers per genome as a function of generation number. The accumulation of crossovers is linear (regression line shown in red, constrained to pass through the origin), with rate 27.1 additional autosomal crossovers per genome per generation. (**B**) Distribution of haplotype block lengths by generation Note that the decay in the size of haplotype blocks is not linear, unlike the increase in the number of crossovers per genome, suggesting that the recombination map is reaching saturation.

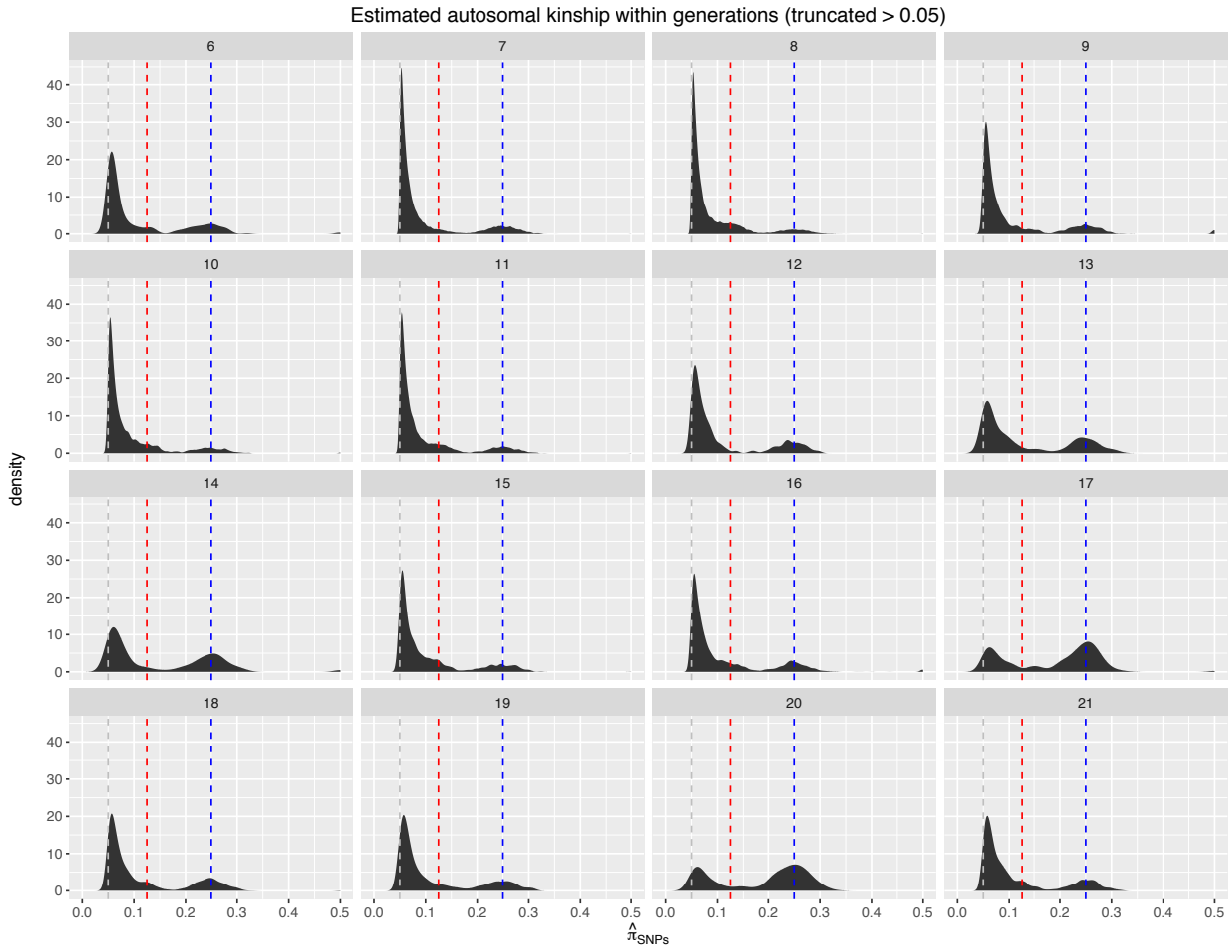

Figure 3.5: Distribution of relatedness within generations in the DO. Kinship coefficients estimated from autosomal SNP genotypes ($\pi_{\hat{SNPs}}$) were computed for all pairs of individuals from the same generation; only values > 0.05 are shown in order to emphasize modes in the distribution corresponding to expected values for cousins (red line; $\pi = 0.125$) and siblings (blue line; $\pi = 0.25$).

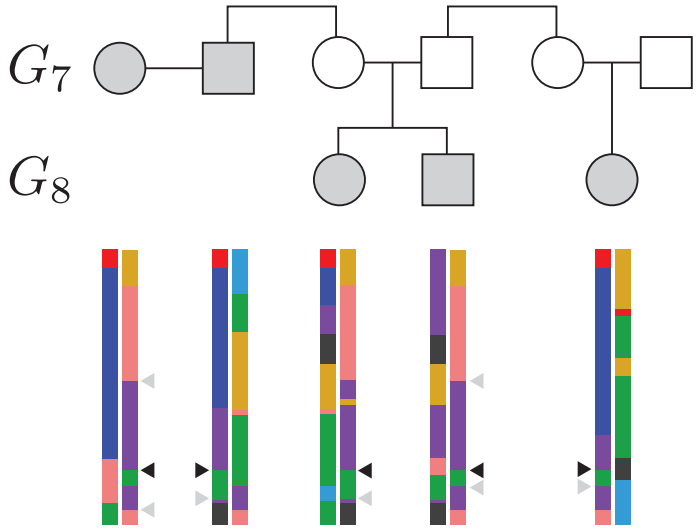

Figure 3.6: Joint inference of DO pedigree and sharing of crossovers. A branch of the pedigree (generations 7 and 8) inferred from SNP genotypes is shown in upper panel; filled shapes are observed individuals and open shapes unobserved. Reconstructed chromosomes are shown for each observed individual. The focal crossover shared IBD by all individuals in this pedigree is indicated with a black arrowhead, and in-phase crossovers used to delineate IBD blocks between individuals are indicated with grey arrowheads. Note that the focal crossover must have occurred no later than generation 6, and that several other crossovers are also shared IBD on these chromosomes.

crossover as shared between *any* two individuals improves with sample size, our ability to detect it as shared in a *particular* pair does not. Consequently our set of $749,560$ distinct crossovers likely contains some duplicates that we have falsely labelled recurrent events, and these duplicates may slightly distort the scaling from crossover counts to centimorgans.

Nonetheless, the cumulative recombination map in the DO is remarkably similar to the G_2:F_1 map when each is plotted in its natural scale (centimorgans for G_2:F_1, crossover counts for the DO; **Figure 3.8**). To convert the DO map to the more interpretable centimorgan units, we smoothed the cumulative map in 100 kb bins on each chromosome and used polynomial regression with degree 3 (see **§ 3.5**) to estimate the relationship between cumulative crossover count and G_2:F_1 centimorgan position, then used the fitted model to predict the centimorgan position of every observed crossover. We use this rescaled map (**Figure 3.9**) for the remainder of the analyses presented here.

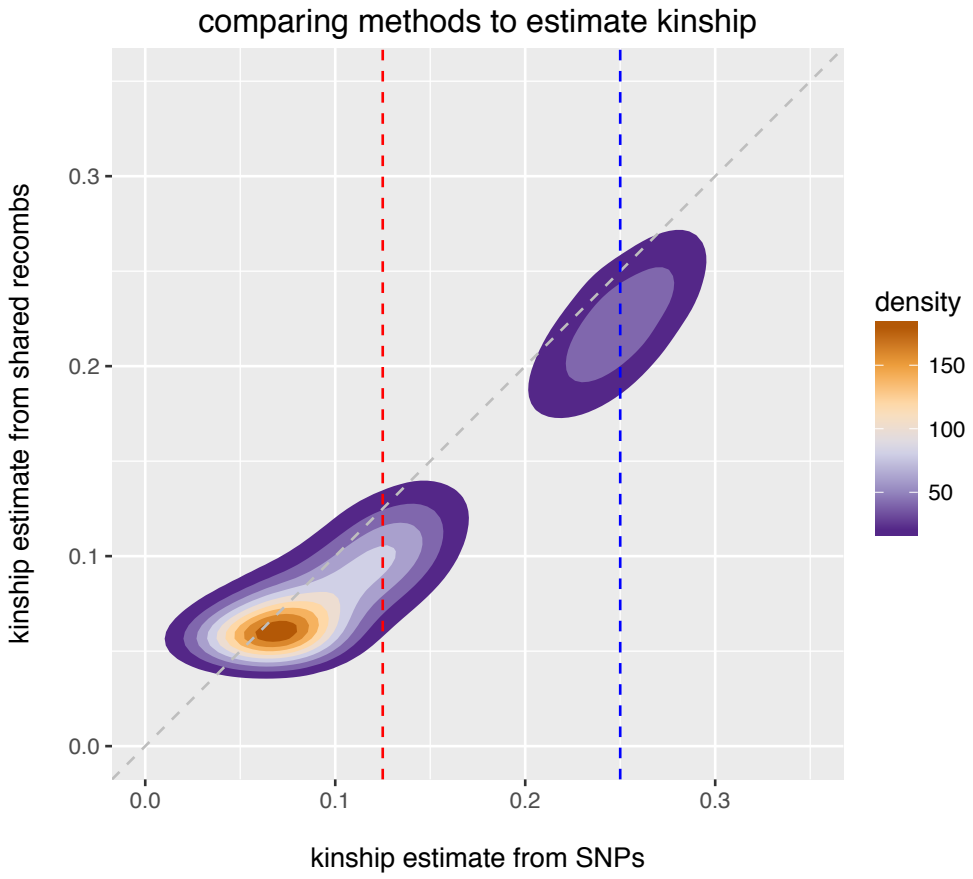

Figure 3.7: Comparison of kinship estimates from genotypes (x-axis) versus the proportion of non-singleton crossovers shared between individuals (y-axis). Red and blue lines represent expected kinship coefficient for cousins and siblings, respectively. Bivariate density is rendered as colors from purple (low) to orange (high).

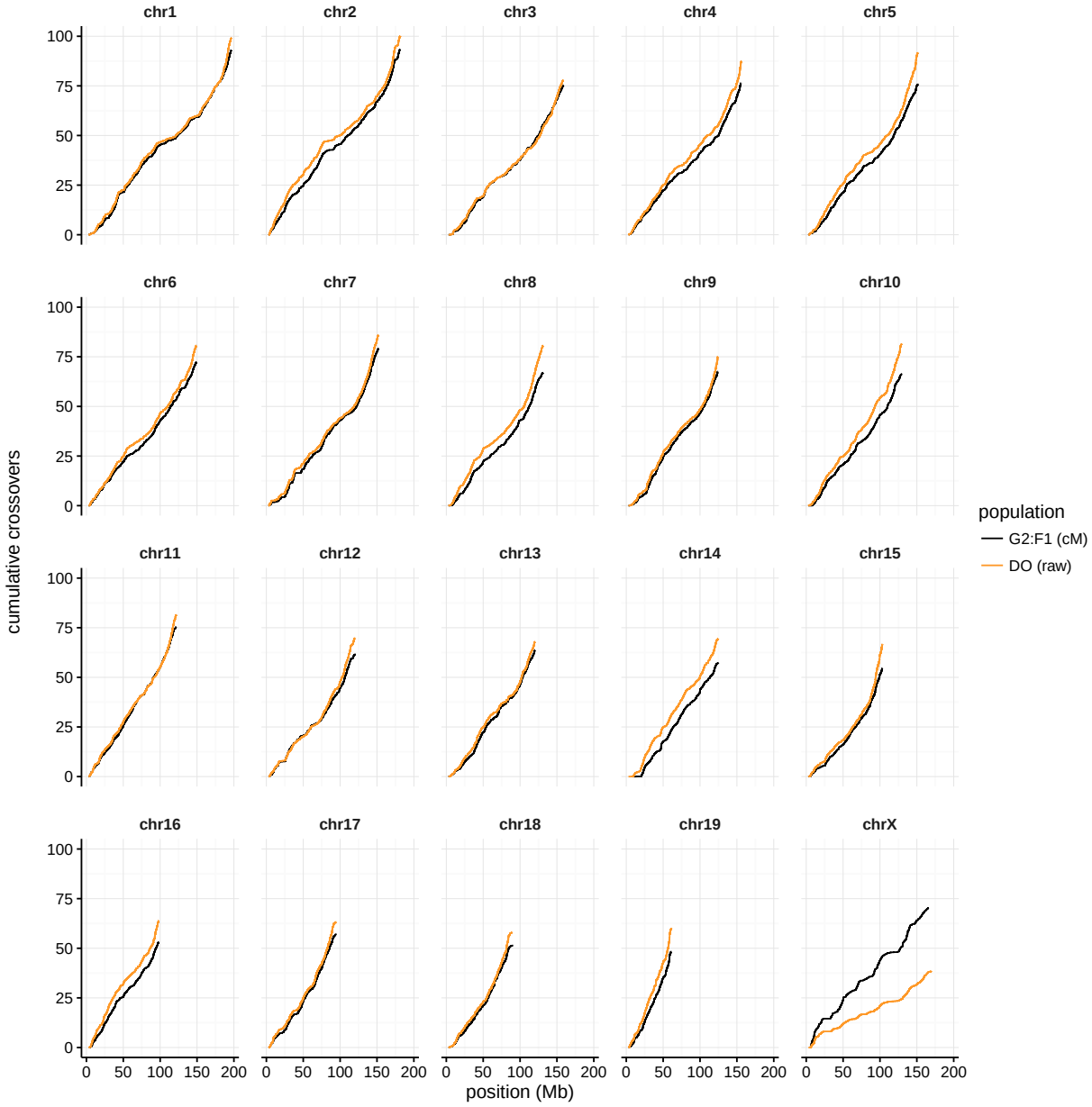

Figure 3.8: Comparison of cumulative recombination maps in CC and DO. Maps are shown as genetic (cM) position versus physical position (Mb) for the CC G_2:F_1 (black), and (cumulative crossover count/500) versus physical position for the DO (grey). The maps are remarkably consistent in length and shape for the autosomes, but differ in length on the X chromosome due to the increased proportion of female meioses contributing to the DO versus the CC map.

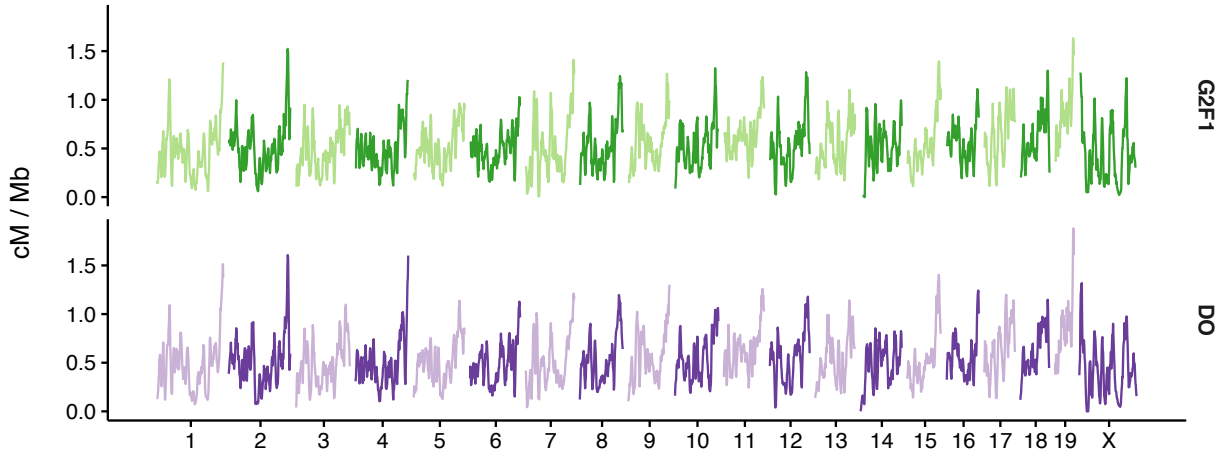

Figure 3.9: Local sex-averaged recombination rates (cM/Mb), calculated in 5 Mb windows with 1 Mb offset between adjacent windows, for the CC G_2:F_1 and DO.

The recombination map in the DO is extremely dense: $749,560$ distinct crossovers equates to approximately 11 crossovers expected between any two of the approximately $70,000$ consecutive markers on the genotyping array. The absence of crossovers in any interval is therefore strong evidence for true local variation in recombination rate rather than sampling artifact. Although local recombination rates (cM/Mb) are broadly similar between the DO and G_2:F_1, the correlation between local recombination rates decreases at finer scales (**Figure 3.10**).

3.2.2 Rate and distribution of crossovers differs by sex

Consistent with existing literature on recombination in mice, humans and many other mammals[?], we found that the rate of crossing-over is greater in females than males: the female G_2:F_1 map for the autosomes ($1,355$ cM) is longer than the male map ($1,221$ cM). There is a small but not significant change in map length from G_1 to G_2 in both sexes (**Table 3.2**). When more than one crossover occurs on a single chromosome (observable only in the G_2 meioses), the distance between crossovers is shorter in females: 40.7 cM (95% CI $39.9 - 41.5$ cM) versus 48.0 ($46.9 - 49.1$ cM) in males (**Figure 3.11A**). We fit the "gamma model" of crossover interference described in[224,225] and found that the strength of interference (parameter ν) is significantly greater in males (12.7 $[11.5,\ 13.9]$) than in females (8.8 $[8.2,\ 9.4]$) (**Figure 3.11B**). Furthermore, the proportion of multiply-recombinant chromosomes is greater in females: 11.8% versus 6.8% ($p < 10^{-5}$, Fisher's exact test). Together these results support the long-standing observation that the recombination rate is higher

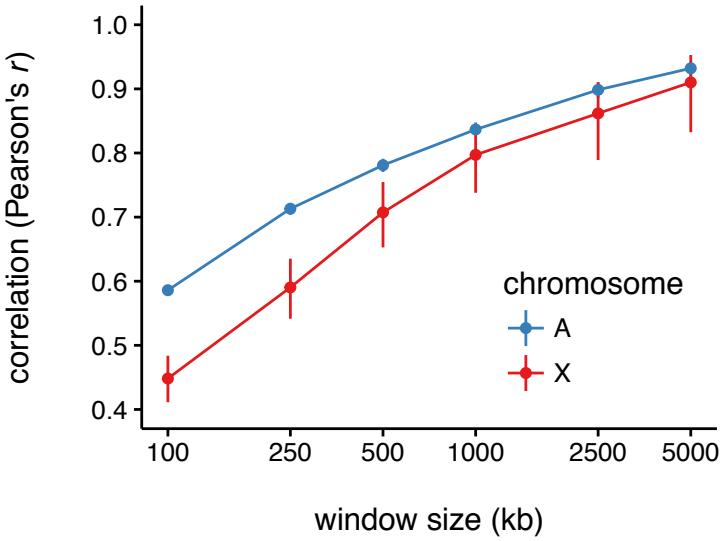

Figure 3.10: Correlation between DO and CC maps as a function of scale. Pearson correlation (r) between local recombination rate (measured as cM/Mb) in the DO and CC G_2:F_1, with correlations computed in windows of increasing size from 100 kb to 5 Mb. For all window sizes, rates were smoothed by allowing adjacent windows to overlap by one-half their width.

Sex	**Female**	**Male**
G_1 (cM)	1330 (1296, 1365)	1279 (1247, 1311)
G_2 (cM)	1381 (1351, 1409)	1211 (1182, 1236)

Table 3.2: Total length of the autosomal genetic map by sex and generation in the G_2:F_1(). Confidence intervals were obtained by non-parametric bootstrap with 100 replicates.

in the female than in the male germline, and that this effect is mediated in part by weaker crossover interference in females[?] .

The spatial distribution of crossovers also differs markedly by sex. Crossovers in males are enriched in the distal portion of all autosomes, while they occur more uniformly across each chromosome in females (**Figure 3.12**). This effect might be a consequence of stronger crossover interference in males: crossovers will necessarily be pushed towards chromosome ends when more than one occurs on a single chromosome. To investigate further we divided the products of G_2 meioses into single- and multiple-recombinant classes and examined the spatial distribution of

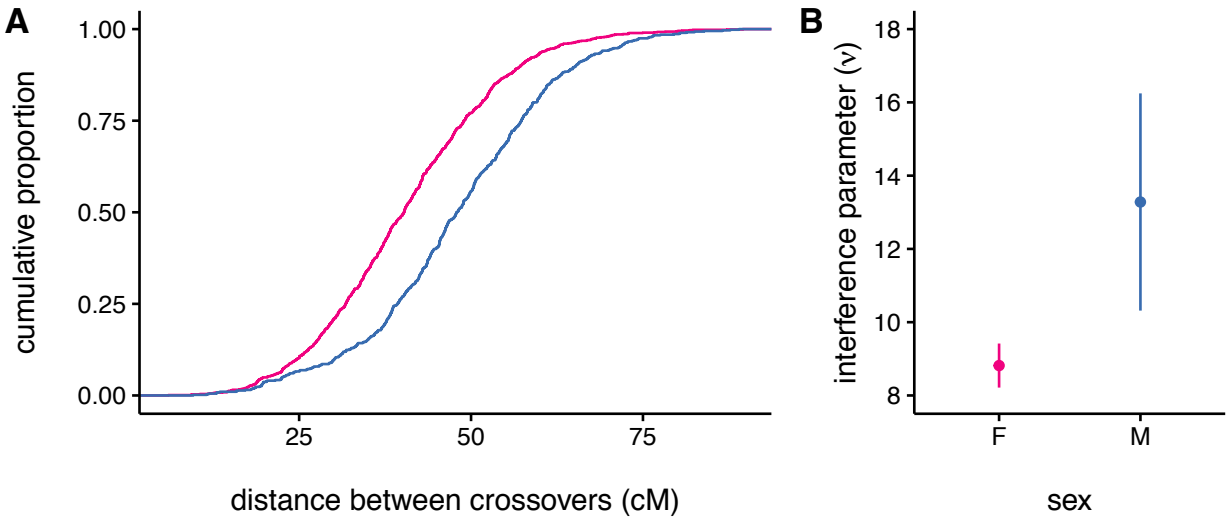

Figure 3.11: Crossover interference differs between males and females in G_2 meioses. (**A**) Distribution of distance (on the genetic map) between adjacent crossovers on the same chromosome in males versus females. (**B**) Estimates of the unitless interference parameter (ν) from the gamma model of crossover interference[224], with 95% confidence intervals, for males and female meioses.

crossovers. We found that the pattern of male-specific enrichment in subtelomeric regions remained even after conditioning on the number of crossovers (**Figure 3.13**).

3.2.3 Sex-linked loci have large effects on recombination rate

To the extent that recombination is implicated in reproductive isolation between subspecies, evolutionary theory predicts that modifiers of the recombination rate should accumulate disproportionately on the sex chromosomes **??**. Several previous studies have identified X-linked loci with dramatic effects on recombination rate in F_2 crosses and in reciprocal F_1 hybrids[175,226,176]. We took advantage of the randomized design of the CC and the ability to assign each G_2:F_1 crossover to a specific meiosis to estimate the marginal effects of the X and Y chromosomes on global recombination rate.

The Y chromosome present at each of the four male meioses in a funnel (MGP, PGP, Pf, Pm; **Figure 3.3**) can be determined without ambiguity from the funnel order because it is non-recombining and hemizygous. The eight founder strains of the CC can be collapsed into just four distinct Y chromosome haplogroups: the strains A/J, C57BL/6J, 129S1/SvImJ and NZO/HlLtJ share the same Y (denoted "ABCE"); NOD/ShiLtJ and WSB/EiJ share a Y (denoted "DH"); and CAST/EiJ

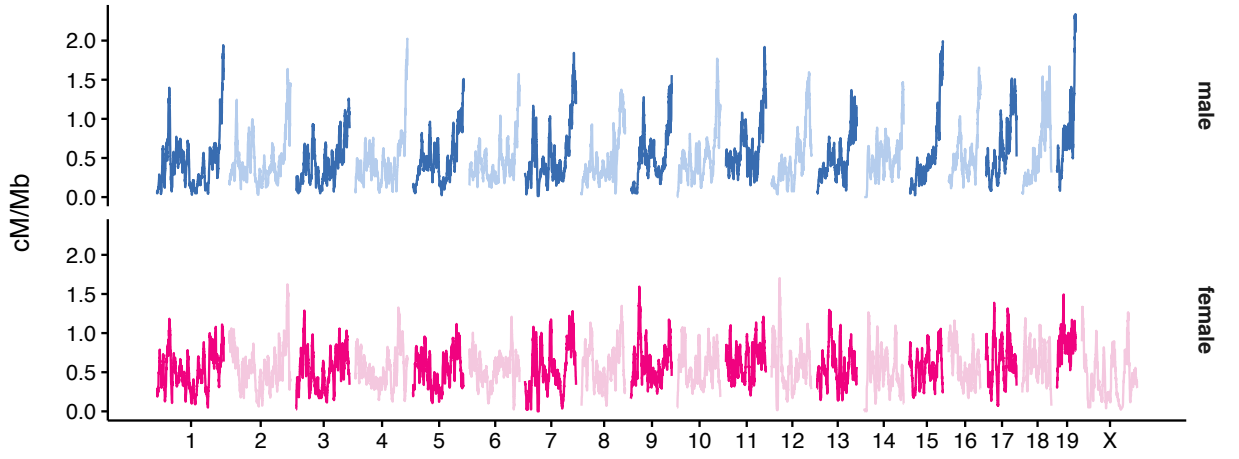

Figure 3.12: Local sex-specific recombination rates (cM/Mb), calculated in 5 Mb windows with 1 Mb offset between adjacent windows, for G_1 and G_2 meioses combined. Note the enrichment of crossovers in the distal portion of all autosomes in males.

and PWK/PhJ have distinct Y haplogroups (denoted "F" and "G" respectively). (See **Chapter 6** for detailed analysis of the ancestry and diversity of mouse Y chromosomes.) We found that Y chromosome haplogroup is significantly associated with the number of crossovers per meiosis in 948 male meioses ($p = 8.8 \times 10^{-7}$). Haplogroup G (PWK/PhJ) is associated with between 1.1 (95% CI $0.4 - 1.7$) extra crossovers per meiosis versus halpogroup ABCE and 2.4 ($1.4 - 3.6$) extra versus haplogroup F (CAST/EiJ). The size and direction of effect is consistent between G_1 meioses — which occur in F_1 hybrids — and G_2 meioses in a four-way mixed background (**Figure 3.14**).

To test the effect of the X chromosome on recombination rate, we used the funnel order to assign to each meiosis a dosage of each of the eight X chromosome haplotypes and used these dosages as predictors in a generalized linear model. Only the CAST/EiJ X chromosome had significant association with number of crossovers per meiosis in both sexes together ($p = 7.1 \times 10^{-5}$, likelihood ratio test with 1 df), and it is associated with an extra 0.9 (95% CI $0.5 - 1.4$) crossovers per meiosis. A test for an (X chromosome) $\times$ sex interaction was not significant ($p = 0.10$), although **Figure 3.15** suggests that the magnitude of the X effect is qualitatively greater in females than in males. These findings are consistent with previous reports of an X-linked locus in CAST/EiJ that explains almost 50% of variation in chiasmata counts in an F_2 cross[175].

Both intragenomic conflict between the sex chromosomes and accumulation of allelic incompat-

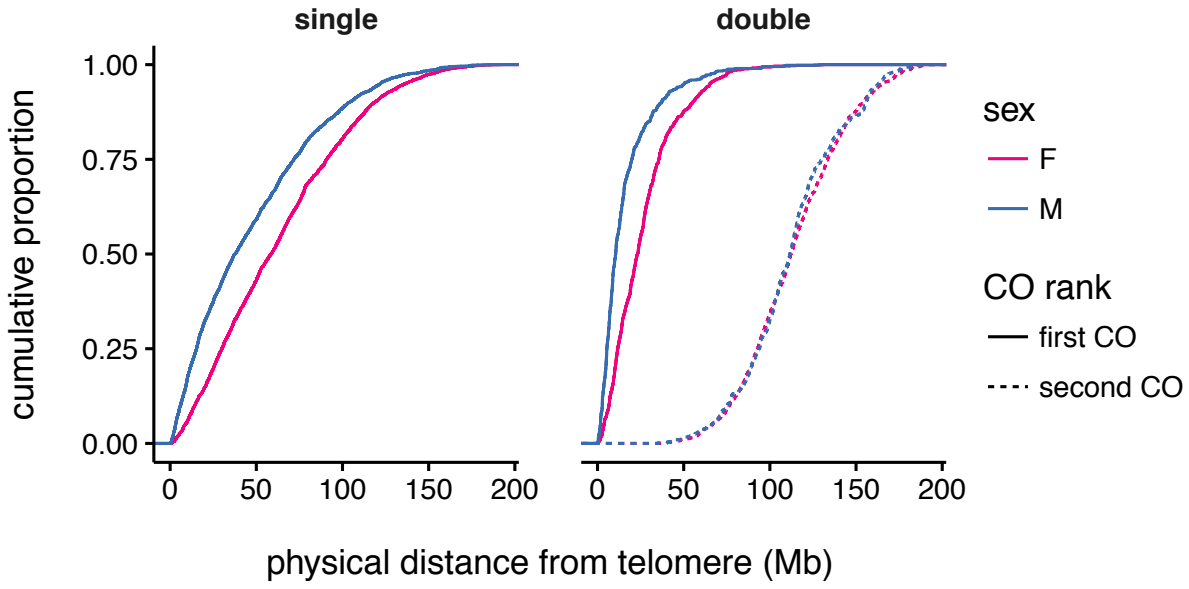

Figure 3.13: Crossovers are enriched in the distal portion of chromosomes in males. Cumulative distribution of physical distance of crossovers from the telomeric end of the chromosome are shown for male and female G_2 meioses. The male distribution is shifted to the left (*i.e.* towards the telomeric end) relative to the female distribution on single-recombinant chromosomes ($p < 10^{-5}$, Kolmogorov-Smirnov test) and for the distal-most crossover on double-recombinant chromosomes ($p < 10^{-5}$), but not for the more proximal crossover on double-recombinant chromosomes ($p = 0.17$).

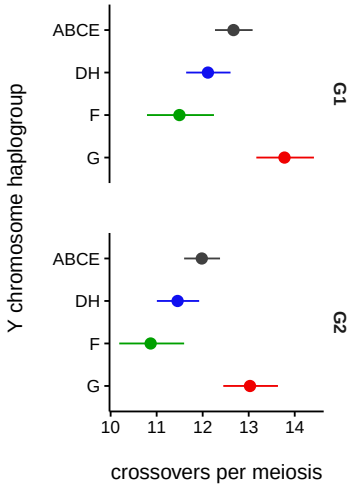

Figure 3.14: Recombination rate, measured as count of transmitted crossovers per meiosis, in G_1 and G_2 males of each of the four Y chromosome haplogroups in the CC. Points are marginal means with 95% CIs from generalized linear model with Poisson response.

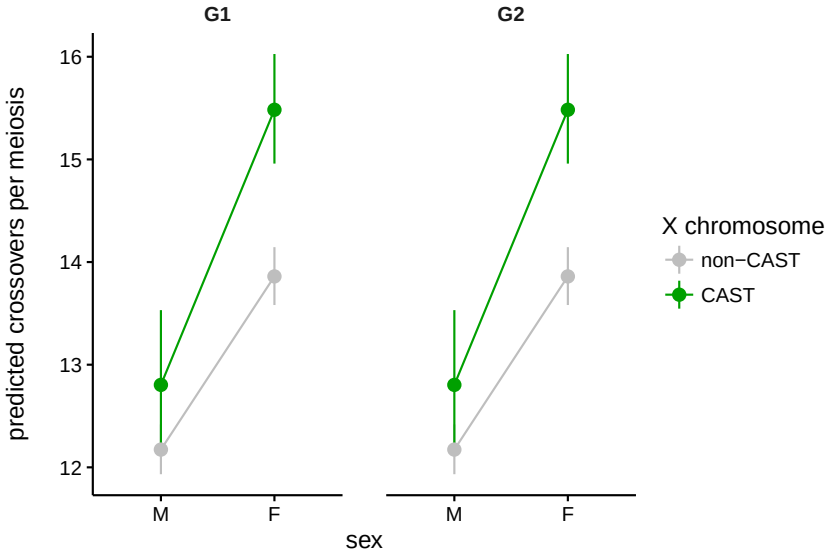

Figure 3.15: Predicted recombination rate, measured as count of transmitted crossovers per meiosis, in males and females with or without a CAST/EiJ X chromosome. Points are fitted values with 95% prediction intervals from generalized linear model with Poisson response.

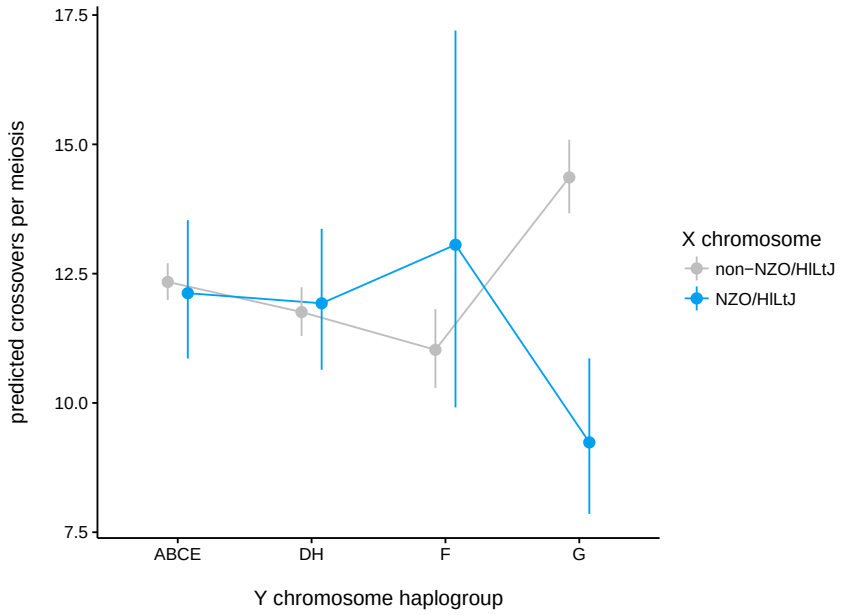

Figure 3.16: Predicted recombination rate, measured as count of transmitted crossovers per meiosis, in males with or without a NZO/HlLtJ X chromosome as a function of Y chromosome haplogroup. Points are fitted values with 95% prediction intervals from generalized linear model with Poisson response.

ibilities between subspecies could plausibly give rise to X-Y interaction effects on recombination rate. We tested for association between X chromosome genotype (dosage) and Y chromosome haplogroup in G_1 and G_2 male meioses jointly and uncovered a significant interaction between Y chromosome and the presence of the NZO/HlLtJ X chromosome ($p = 1.1 \times 10^{-4}$ by likelihood-ratio test with 3 df). When paired with the PWK/PhJ Y chromosome, the NZO/HlLtJ X is associated with a dramatic decrease in the number of transmitted crossovers (**Figure 3.16**). However, all of the information for the X:Y interaction estimate is derived from G_2 males: the (NZO/HlLtJ$\times$PWK/PhJ) cross is unproductive[221], and it was avoided in the design of CC funnels.

3.2.4 Advanced paternal age increases recombination rate

Incidence of most aneuploidy syndromes in humans is strongly associated with advanced maternal age[11]. Classic studies of mouse oocytes noted a change in the distribution of chaismata in older mothers[227], and nondisjunction in human aneuploidies is associated with aberrations in both the rate and spatial distribution of crossovers[228,229,230]. These observations have spurred

great interest in the effect of age on the rate and distribution of recombination in the female germline. Results of several large studies have been surprisingly mixed: evidence has been provided for increased[231,232] and decreased[233] recombination with advancing age, and effects may be confounded with population-specific demographic factors[234]. (It should be noted that these pedigree-based studies measure the number of transmitted crossovers in offspring, not the number of chiasmata in oocytes.)

By contrast, the existence and direction of a paternal age effect on recombination has received less attention. The largest human studies have found no evidence for an age effect on the number of transmitted crossovers[231,233]. The number of chiasmata in diplotene spermatocytes has been shown to increase in aged relative to peri-pubertal male mice from several classical inbred strains[235]. Although chiasmata count is a good predictor of the number of transmitted crossovers in males[219], it is possible that this correlation deteriorates with age. The effect of genetic background — especially of large-effect modifiers of the recombination rate on the sex chromosomes — on the paternal age effect is also unknown.

We used a set of intercross pedigrees (**Figure 3.17**) to simultaneously test the effect of paternal age and the sex chromosomes on the number of crossovers transmitted to progeny.[3] Briefly, reciprocal F_1 hybrid males between wild-derived strains CAST/EiJ, PWK/PhJ and WSB/EiJ were mated at young age to a young, fertile FVB/NJ female; aged for two years; and then mated to a new, young FVB/NJ female. Progeny were collected at genotyped at each timepoint, and the number and location of each crossover tallied. Sample sizes are shown in **Table 3.3**. In total, $4,079$ autosomal crossovers from 301 informative meioses (all male) were identified.

Advanced paternal age is weakly associated with an increase in the number of autosomal crossovers ($F_{1,298} = 5.67$, $p = 0.018$). Offspring of old males inherit an average of 0.7 (95% CI $0.03 - 1.3$) extra crossovers compared to offspring of young males (**Figure 3.18A**). The subspecific origin of the X chromosome has a much stronger effect ($F_{1,298} = 21.3$, $p = 5.8 \times 10^{-6}$): offspring of males with a *M. m. musculus* X chromosome — crosses PWK/PhJ×CAST/EiJ and PWK/PhJ×WSB/EiJ — inherit an average of 1.6 ($0.9 - 2.3$) extra crossovers compared to offspring of males with a

[3]This experiment was designed by Jim Crowley and Fernando Pardo-Manuel de Villena. Mice were bred by Jim Crowley between 2009 and 2012.

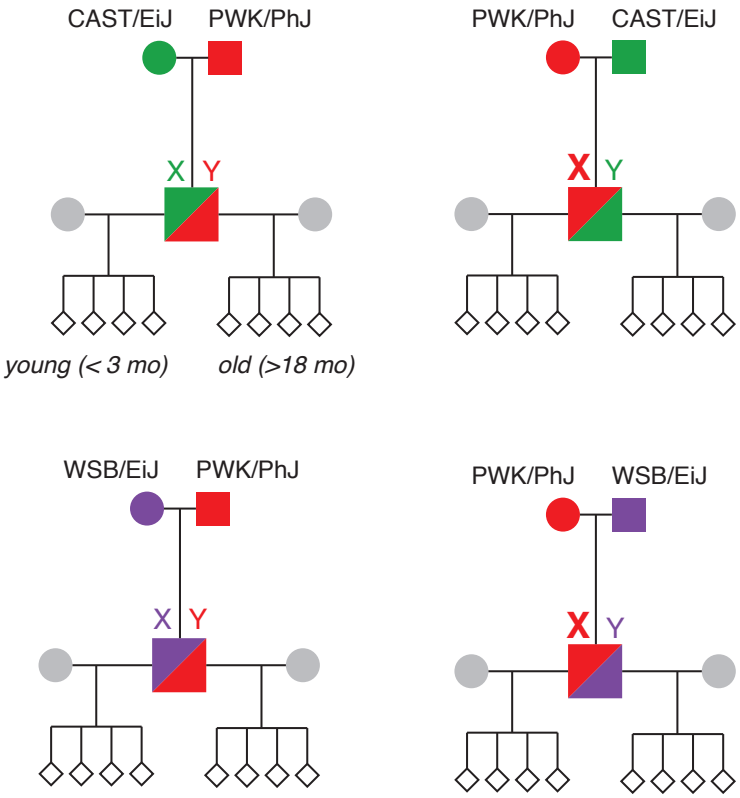

Figure 3.17: Pedigrees used to test paternal age effect on recombination. Reciprocal F_1 hybrids were generated for two of the three possible combinations of the wild-derived founder strains of the CC (CAST/EiJ, PWK/PhJ and WSB/EiJ). Males were bred to a young FVB/NJ at around 3 months of age, and then to a new FVB/NJ female at 18 months of age. X and Y chromosome configurations are shown for the males, and carriers of a *M. m. musculus* X chromosomes are emphasized.

	Paternal age	
Cross	young	old
PWK/PhJ×CAST/EiJ	75	105
PWK/PhJ×WSB/EiJ	12	14
WSB/EiJ×PWK/PhJ	47	48

Table 3.3: Number of genotyped progeny by cross and paternal age.

non-*musculus* X. We could not detect an age×X-chromosome interaction effect ($F_{1,298} = 1.64$, $p = 0.20$). Whereas the increased recombination rate in females relative to males in the CC is due in part to weaker crossover interference in females, we find no effect of paternal age on the interference parameter ν under the gamma model (**Figure 3.18B**). Interference is lower in crosses with a *musculus* X ($\nu = 8.6 \pm 0.7$) than without ($\nu = 11.1 \pm 0.8$), but the difference does not reach significance at the $\alpha = 0.05$ threshold. Instead, the increase in crossover count by both age and X-chromosome genotype appears to be driven by an decrease in the proportion of non-recombinant longer chromosomes (arbitrarily defined here as chromosomes 1 through 12), or equivalently an increase in the proportion of double-recombinant chromosomes, without a change in the number of single-recombinant chromosomes transmitted (**Figure 3.18C**). Among double-recombinant chromosomes, the distance between adjacent crossovers is not different by age ($p = 0.48$, Wilcoxon rank-sum test) but is marginally different ($p = 0.05$) by X chromosome (**Figure 3.18D**).

An excess of X-Y asynapsis has been reported in aged male mice[235] but not humans[236]. Recombination in the mouse PAR has proven difficult to study[237] because the PAR in classical inbred strains is very short compared to other mammals, and composed mostly of repetitive sequences that are difficult to assemble[197]. However, the Y chromosome of the CAST/EiJ strain carries an extra 430 kb of sequence that is X-linked in other strains and just proximal to the PAR boundary[238]. Our genotyping array (MegaMUGA) has several markers in this "extended PAR" informative between CAST/EiJ and PWK/PhJ. We estimate 9.6% (95% CI $5.7 - 14.3\%$) recombination between the distal-most informative marker, UNC31595576 (chrX:169921994) and sex in 176 progeny of the PWK/PhJ×CAST/EiJ cross (**Figure 3.19** and **Table 3.4**). There is no association with paternal age (OR $= 1.07$, $p = 0.99$ by Fisher's exact test.) The recombination rate in the extended PAR is 1% per 44.5 kb, approximately tenfold higher than the genome-wide average of 1% per 500 kb.

3.2.5 Crossovers are enriched in known hotspots

Patterns of linkage disequilibrium in human and great ape populations[219] and maps of recombination precursors in mouse[218] have shown that essentially all crossovers in mammals occur within recombination hotspots. Crossover positions are constrained by the position of programmed DSBs during meiotic prophase, which are in turn dictated by affinity of the interaction between the histone methyltransferase PRDM9 and instances of its degenerate 13 bp binding motif. Tens of

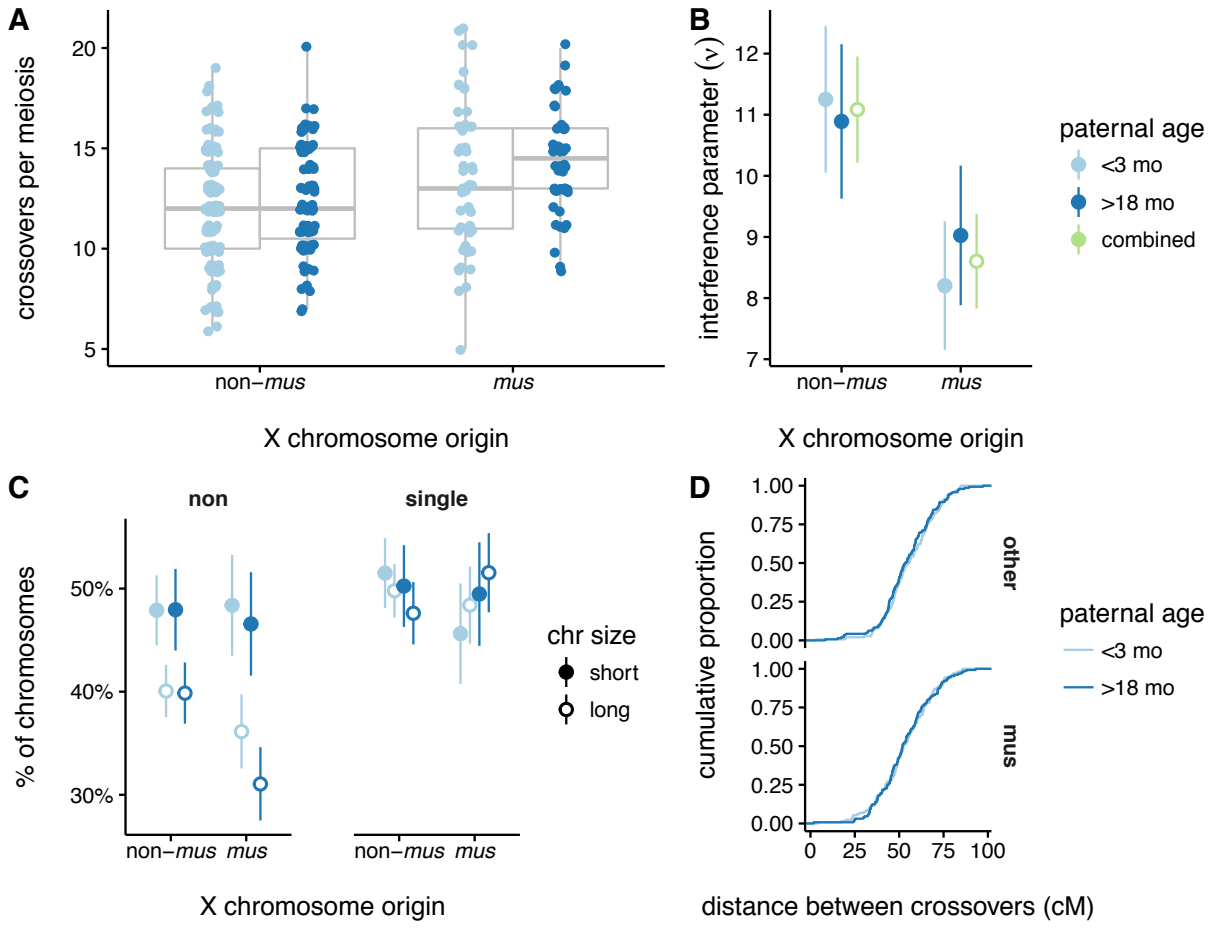

Figure 3.18: Effect of age and X chromosome on male recombination. (**A**) Distribution of the number of transmitted crossovers per meiosis by X chromosome origin and paternal age. (**B**) Estimates of interference parameter ν under the gamma model (± 1 SE) by X chromosome origin and age. (**C**) Proportion of non-recombinant (left) and single-recombinant (right) chromosomes by X chromosome origin, paternal age, and chromosome size. (**D**) Cumulative distribution of genetic distance between adjacent crossovers on double-recombinant chromosomes, by X chromosome origin and paternal age.

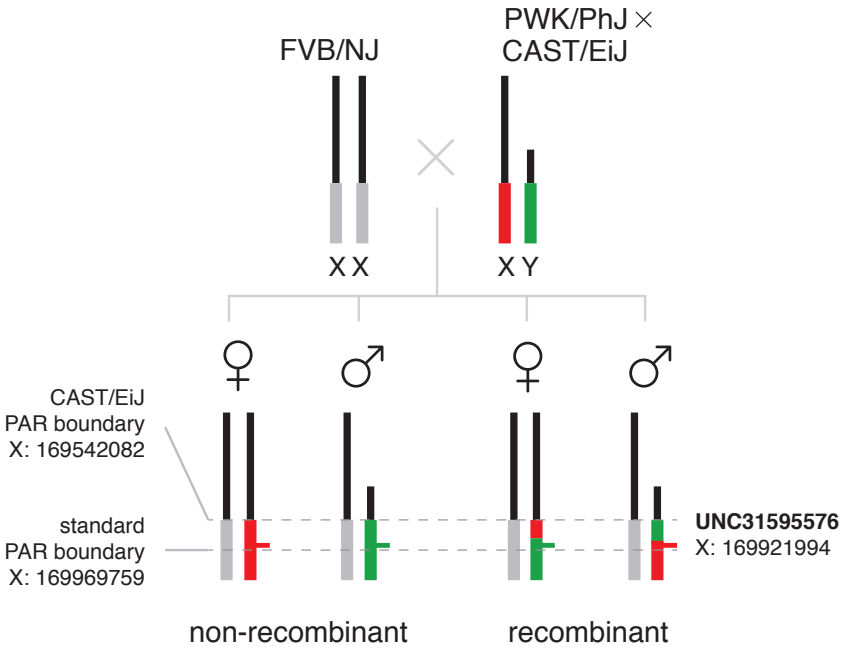

Figure 3.19: Strategy for measuring pseudoautosomal recombination in offspring of (PWK/PhJ×CAST/EiJ)F_1 males. In the absence of recombination, female progeny inherit a PWK/PhJ X and males a CAST/EiJ Y. Recombination can be detected between sex and the distal-most informative marker within the extended PAR of CAST/EiJ (UNC31595576).

	Paternal age		
Product	young	old	**Total**
non-recombinant	91 (90.1%)	68 (90.7%)	159 (90.3%)
recombinant	10 (9.9%)	7 (9.3%)	17 (9.7%)
	101	75	176

Table 3.4: Count of meiotic products according to recombination status in the extended pseudoautosomal region (PAR) of CAST/EiJ, observed in progeny of (PWK/PhJ×CAST/EiJ)F_1 males. There is no association between the rate of crossovers in the PAR and paternal age (OR $= 1.07$, $p = 0.99$ by Fisher's exact test.)

thousands of hotspots have been experimentally predicted in male mice using inbred strains and F_1 hybrids. However, due to the process of "hotspot erosion," it remains difficult to predict which sequences will be targeted for DSBs in a mixed genetic background heterozygous for different *Prdm9* alleles. Little is known about the process through which the small subset of DSBs that will become crossovers are distinguished from the remaining non-crossover products. Furthermore, hotspot maps derived from recombination precursors have so far been constructed only for males because of the difficulty of obtaining tissue at the appropriate developmental stage in females.

Crossovers in the DO can be resolved to a median interval of 29.6 kb (median absolute deviation [MAD] 30.1 kb) on the autosomes and 66.0 kb (72.7 kb) on the X chromosome. While this resolution is too coarse to identify hotspots *de novo*, the total size of the dataset allows a powerful test for sex-averaged usage of known hotspots. We calculated the cumulative density of crossovers (accounting for uncertainty in crossover position) within hotspots ascertained as H3K4me3 peaks in spermatocytes of male offspring of crosses between C57BL/6J, WSB/EiJ, CAST/EiJ and PWD/PhJ (closely related to PWK/PhJ). Experimentally-defined hotspots overlap a median 0.18 crossovers per kb of hotspot in the DO, versus 0.09 crossovers per kb in random genomic intervals of equal size. The enrichment of crossovers in hotspots is similar regardless of the genetic background in which hotspots were ascertained (**Figure 3.20A**). Hotspot strength, as defined by the density of the H3K4me3 signal, is weakly correlated (Spearman's $\rho = 0.24$) with crossover density (**Figure 3.20A**).

Surprisingly, only 48.8% (95% CI $47.8 - 49.7\%$ by non-parametric bootstrap) of crossovers on autosomes and 47.4% ($46.3 - 48.3\%$) of crossovers on the X chromosome overlap any previously-identified hotspot. Although the broadness of our crossover intervals relative to hotspots makes it unlikely that crossover intervals would fail to overlap the underlying hotspot by chance, we sought to rule out artifacts due to the distribution of informative SNP markers on our genotyping platforms. We partitioned crossovers according to the strains at the junction and discovered that hotspot overlap is systematically reduced among crossovers between classical inbred strains versus crossovers involving a wild-derived strain (**Figure 3.21**). This suggests that the apparent lack of overlap with hotspots may be due to a lack of precision in the definition of crossover intervals when the strains at the junction share a haplotype identical by descent (IBD). When crossovers are partitioned according to whether or not they occur in regions where all five classical inbred founder strains are IBD, it is clear that hotspot overlap is reduced in IBD regions, but that IBD regions do

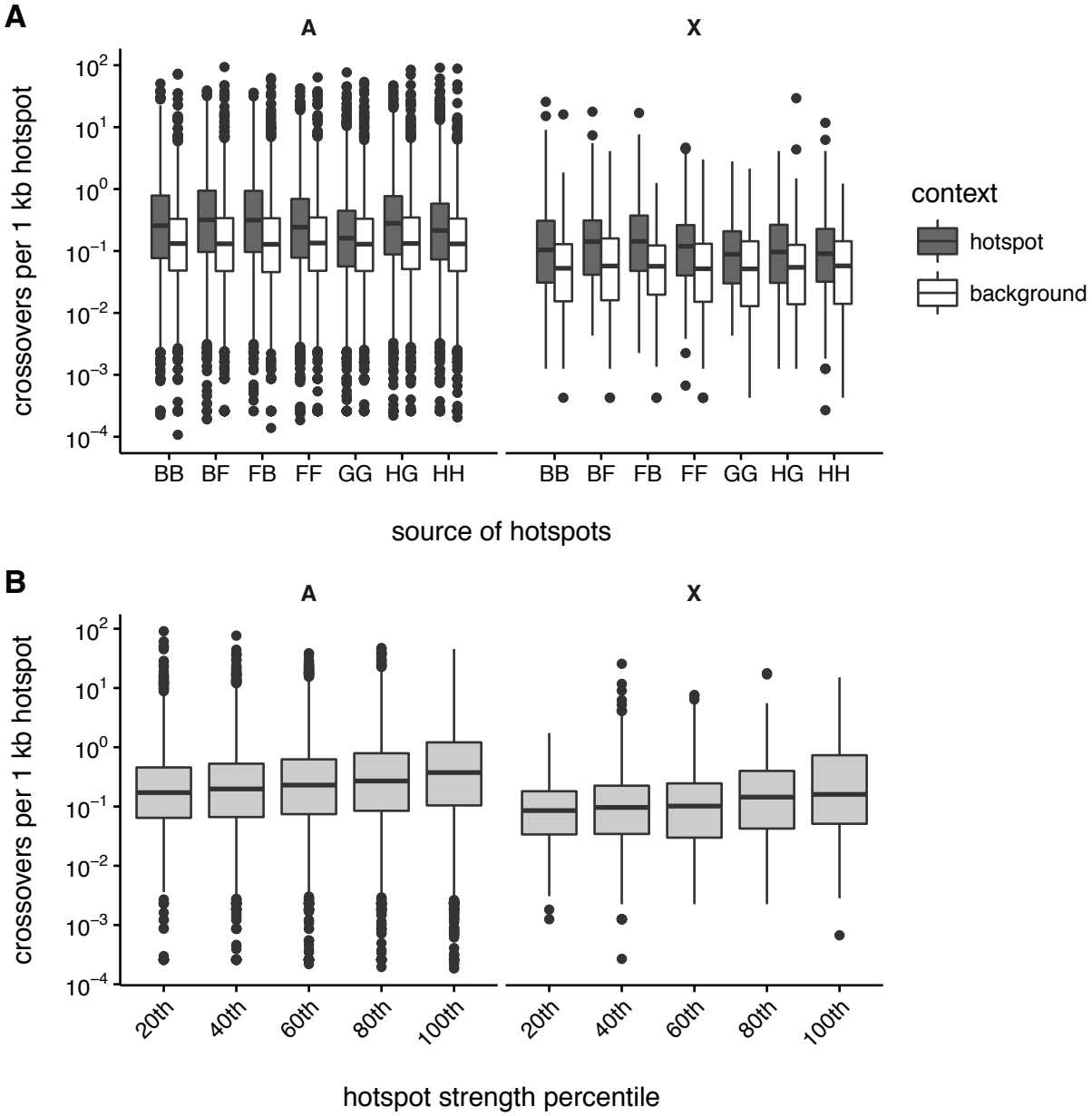

Figure 3.20: Recombination hotspot usage in the DO. (**A**) Distribution of crossover density in hotspots defined by H3K4me3 ChIP-seq in testes from several genotypes, versus random genomic intervals of equal size. Crosses are denoted as (maternal strain)×(paternal strain), and strains denoted by their one-letter codes: B = C57BL/6J, F = CAST/EiJ, G = PWD/PhJ, H = WSB/EiJ. (**B**) Distribution of within-hotspot crossover density by bins of hotspot strength. Overlap is computed separately for the autosomes (A, right) and X chromosome (X, left).

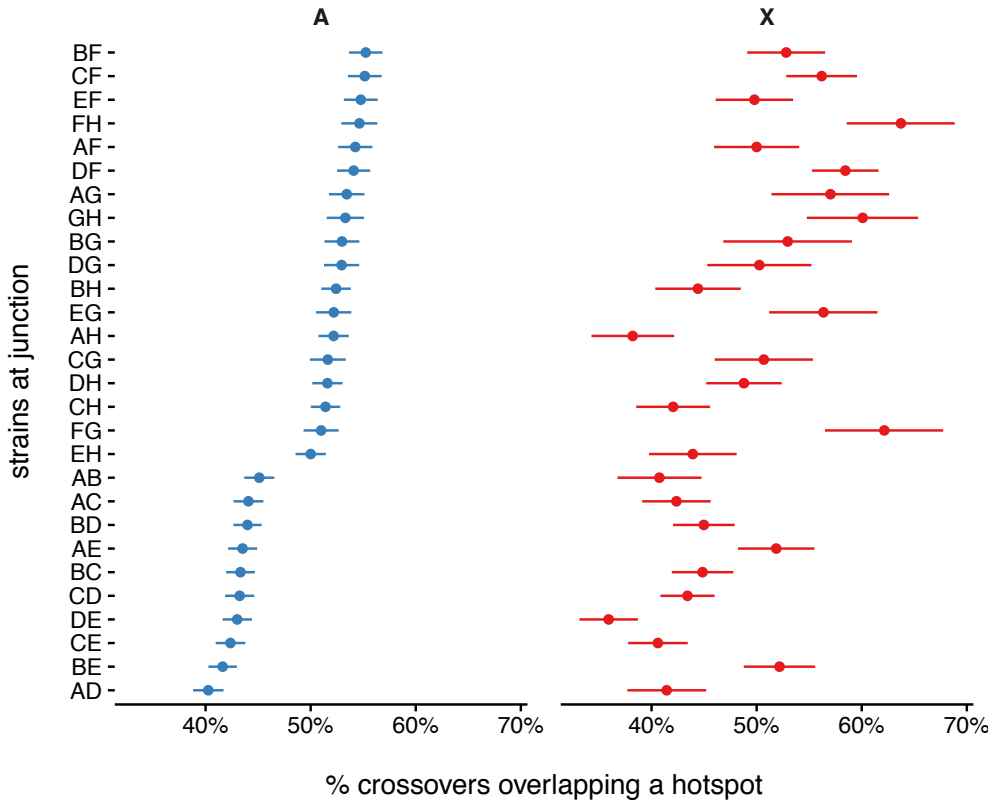

Figure 3.21: Proportion of crossovers overlapping a known recombination hotspot, according to which of the $\binom{8}{2} = 28$ pairs of founder strains are present at the junction. Hotspot usage is shown separately for autosomes (A, left) and the X chromosome (X, right), and pairs are sorted according to hotspot usage on the autosomes.

not fully explain the reduction in hotspot overlap versus crossovers involving a wild-derived strain (**Figure 3.22**).

3.2.6 Crossovers are suppressed near large structural variants

At the megabase scale, genetic and physical distance are well-correlated. Nonetheless, we noticed local plateaus in the cumulative genetic maps for most chromosomes indicative of regions with much smaller genetic than physical size. The effect was present in both the G_2:F_1 and DO maps and was especially obvious on the X chromosome (**Figure 3.8**). Based on this observation we used the extremely dense DO map to systematically define 105 contiguous regions of 100-fold or more reduced recombination rate relative to the chromosome-specific background (see **§ 3.5**). These "coldspots" are not simply the complement of hotspots: whereas the hotspots investigated

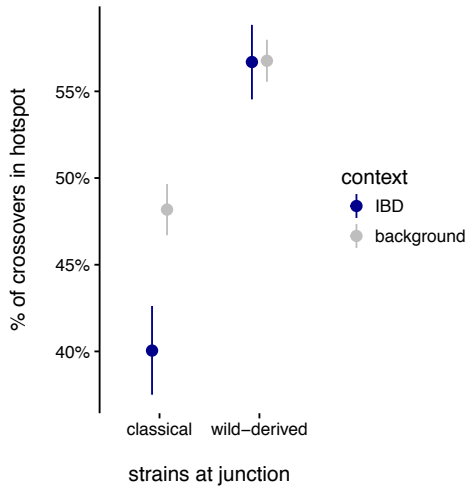

Figure 3.22: Proportion of crossovers overlapping a known recombination hotspot, among crossovers falling within or outside regions shared IBD between all five classical inbred founder strains.

above have median spacing 14.6 kb (MAD 15.9 kb), coldspots span 600 kb to 13.6 Mb. Coldspots are found on all chromosomes but are particularly abundant on the X chromosome (accounting for 22.4% of its length.) An example is shown in **Figure 3.23**.

Given the approximately uniform genomic distribution of DSBs in spermatocytes, coldspots potentially represent the downstream effect of the regulatory process(es) that designate DSBs for crossover versus non-crossover outcomes. We therefore sought to identify the sequence, structural or epigenetic features responsible for suppression of crossovers in coldspots. We first used reference genome annotations and comparative genomics data to define the genomic profile of coldspots (summarized in **Table 3.5**). Coldspots contain fewer protein-coding genes but more pseudogenes than the rest of the autosomes and X chromosome. They are enriched for some classes of transposable elements (LINEs and LTRs) but not others (SINEs). Coldspots lie in evolutionary labile regions of the genome: they are half as likely to contain a conserved element in a 40-way multiple sequence alignment of eutherian mammals than random genomic regions. Most dramatically, coldpsots have 3.6-fold enrichment of segmental duplications (SDs), defined here as duplications longer than 1 kb with $> 90\%$ mutual sequence identity.

Structural variation in populations is concentrated in and around clusters of segmental duplications[42,37,239,43,36]. Although in practice SDs are defined on the basis of a single reference sequence,

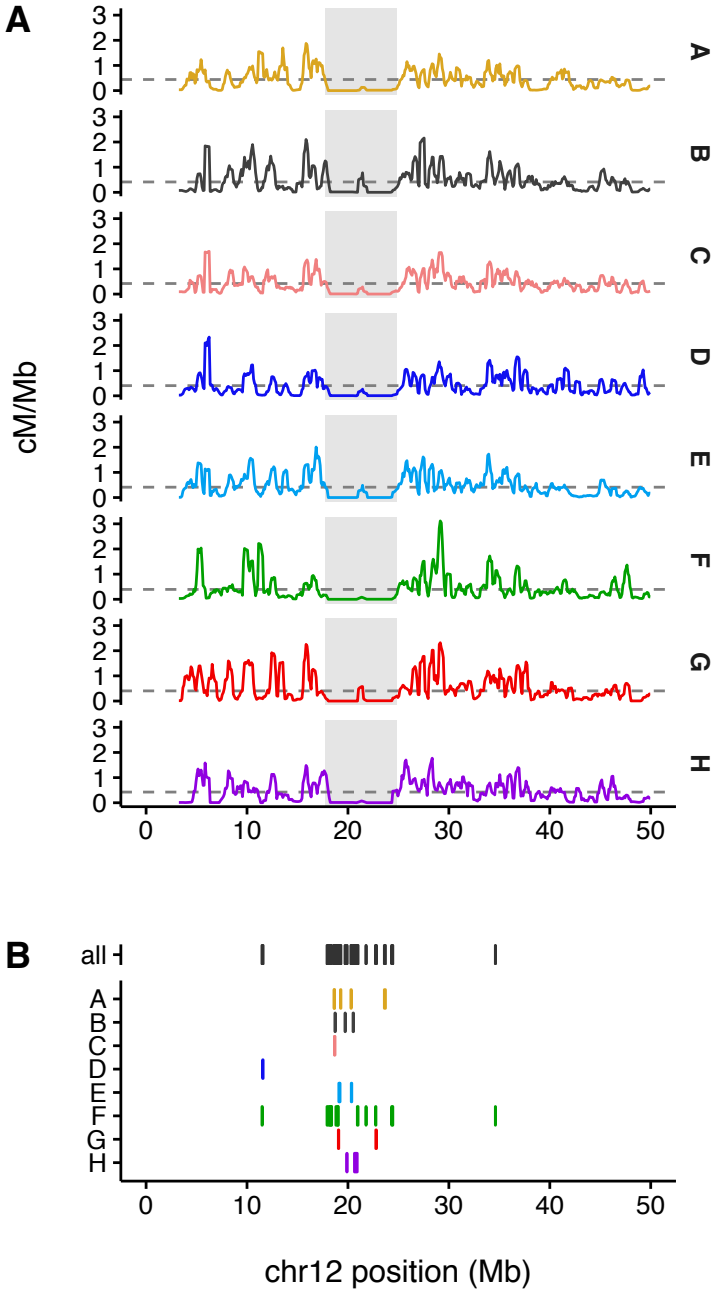

Figure 3.23: Example of a recombination coldspot on chromosome 12. (**A**) Strain-specific local recombination rates (in cM/Mb) across the proximal 50 Mb of chromosome 12. Median chromosome-wide recombination rates for each strain are marked with grey dashed lines. The coldspot is indicated by the grey shaded region. (**B**) CNVs ascertained in the DO. Top line, all CNVs irrespective of strain distribution pattern; remaining lines, CNVs with an allele private to a single strain.

Feature	Enrichment	q-value
Reference genome annotations		
Protein-coding gene	0.60	< 0.001
Pseudogene	1.98	< 0.001
LINE	1.69	< 0.001
SINE	0.69	< 0.001
LTR	1.67	< 0.001
Segmental duplications	**3.66**	< 0.001
Variation between species		
GERP constrained elements	0.46	< 0.001
Variation within species		
IBD among classical strains	0.69	< 0.001
Common CNVs in DO	**3.64**	< 0.001
Multiallelic CNVs in DO	**4.34**	< 0.001

Table 3.5: Enrichment of various genomic annotations in coldspots versus genome background. Significance was computed over 1000 shuffles and is expressed as the q-value, proportion of tests expected to represent false discoveries. GERP constrained elements are sequences conserved across 40 eutherian mammals as defined by the Ensembl Compara pipeline[115]. CNVs are defined on the basis of whole-genome sequencing of 228 DO mice as described in the main text.

this sequence represents just one random draw from the pool of segregating structural variants (SVs) in the population. We posited that the extensive overlap between coldspots and SDs was simply a proxy for an underlying association between coldspots and large SVs segregating in the CC and DO. To test this hypothesis we used low-coverage whole-genome sequencing data from 228 male DO mice to identify copy-number variable regions and examined their overlap with coldspots for recombination. First we calculated normalized read depth (an estimator of copy number) in 25 kb windows across the genome for each individual, and then calculated the coefficient of variation (median / MAD) in the population for each window. In this way we identified copy-number variable regions without attempting to assign genotypes at the individual level. The result is striking: regions with population-level variation in copy number are nearly always coincident with coldspots for recombination (**Figure 3.24**). But the reverse is not true, and suppression of crossovers in some cases extends megabases away from the nearest copy-number variable region (*e.g.* central chromosome 2, distal chromosome X).

Defects in pairing, synapsis or DSB resolution between alleles with unequal copy number in heterozygous individuals could plausibly explain the absence of crossovers in copy-number variable regions. To investigate the relationship between CNV allele-sharing and crossovers, we ascertained and genotyped $1,749$ CNV loci ($1,595$ on the autosomes and 154 on the X chromosome) at least 10 kb in size. The final callset contains only CNVs with minor-allele count > 5 in the sample of 228 individuals whose position and founder allele copy numbers could be confirmed by genetic mapping (**Figure 3.25**). Overlapping loci with identical strain distribution patterns were merged into a single locus. A majority of CNV loci ($1,227$; 71%) have a minor allele private to a single founder strain, and most of these private alleles (862; 70%) are contributed by the wild-derived strains CAST/EiJ, PWK/PhJ and WSB/EiJ. As expected, CNVs cluster near SDs: $1,211$ CNV loci (69%) overlap SDs, and the great majority of these (91%) are multiallelic. Coldspots are enriched in CNVs, as expected, and the enrichment is stronger for multiallelic CNVs (**Table 3.5**). **Figure 3.26** summarizes the properties of CNVs ascertained in the DO.

Our working model predicts that those few crossovers which occur in or near coldspots should be biased towards some pairs of founder haplotypes — those with equal or similar copy number — over others. To measure this bias we calculated an information score, defined as the Kullback-Leibler divergence between the observed frequency of junctions and the expectation based on the

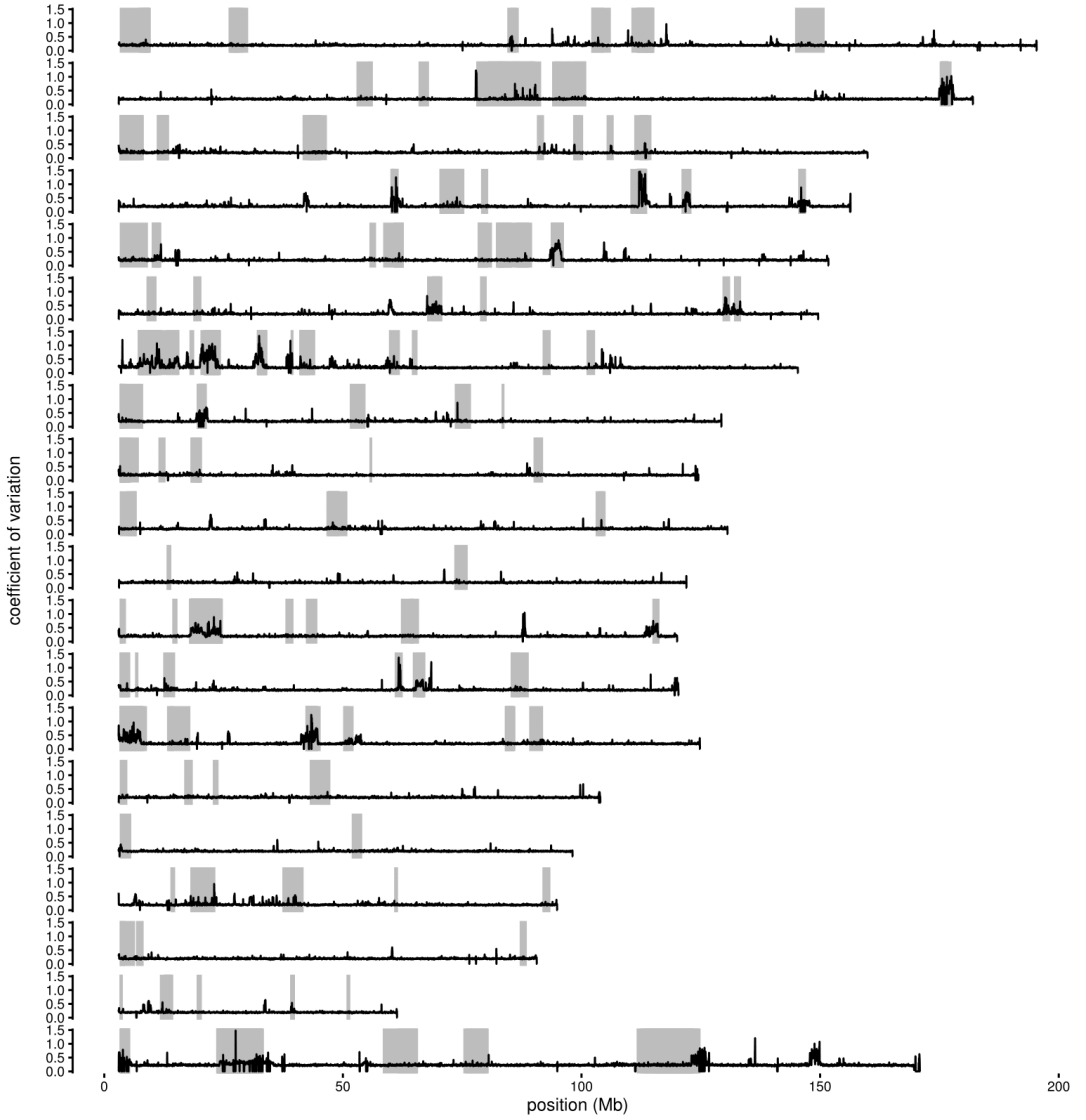

Figure 3.24: Genome-wide view of copy-number variability in the DO. The coefficient of variation (median/MAD) of normalized read depth is shown on all autosomes and the X chromosome. Grey shaded regions are coldspots for recombination.

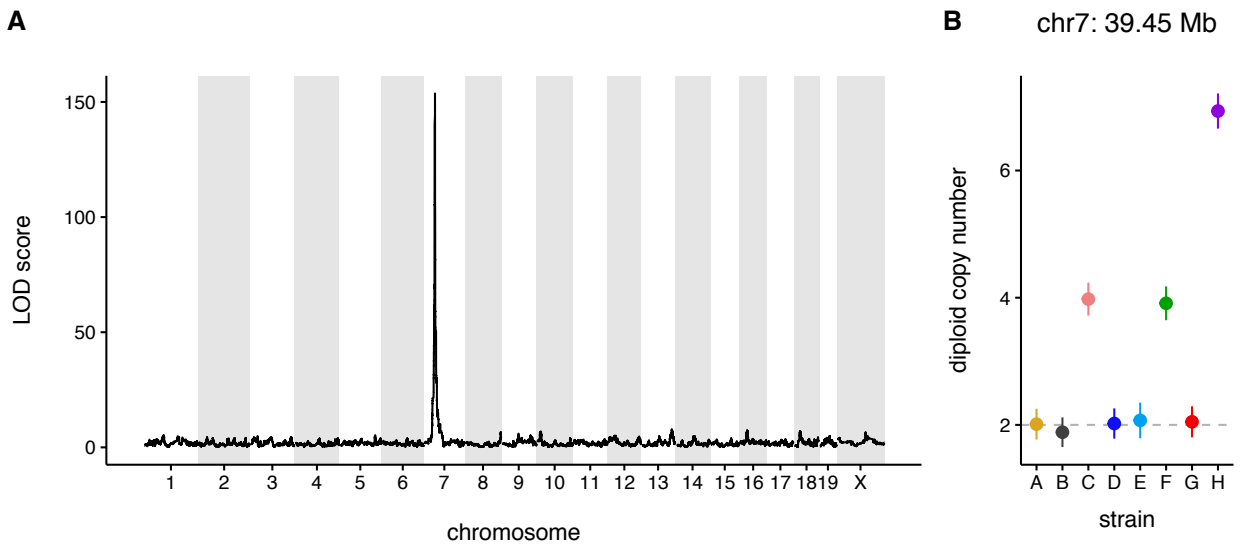

Figure 3.25: Genetic mapping confirms position and allelic configuration of complex CNVs. (**A**) LOD scores from single-locus QTL scan using copy number at chr7: 38.29 Mb as a quantitative trait. (**B**) Founder strain means (± 2 SE) at the QTL peak (chr7: 39.45 Mb), which provide a direct estimate of founder copy number at this triallelic CNV. Note that the QTL peak is not located exactly at the nominal position of the CNV.

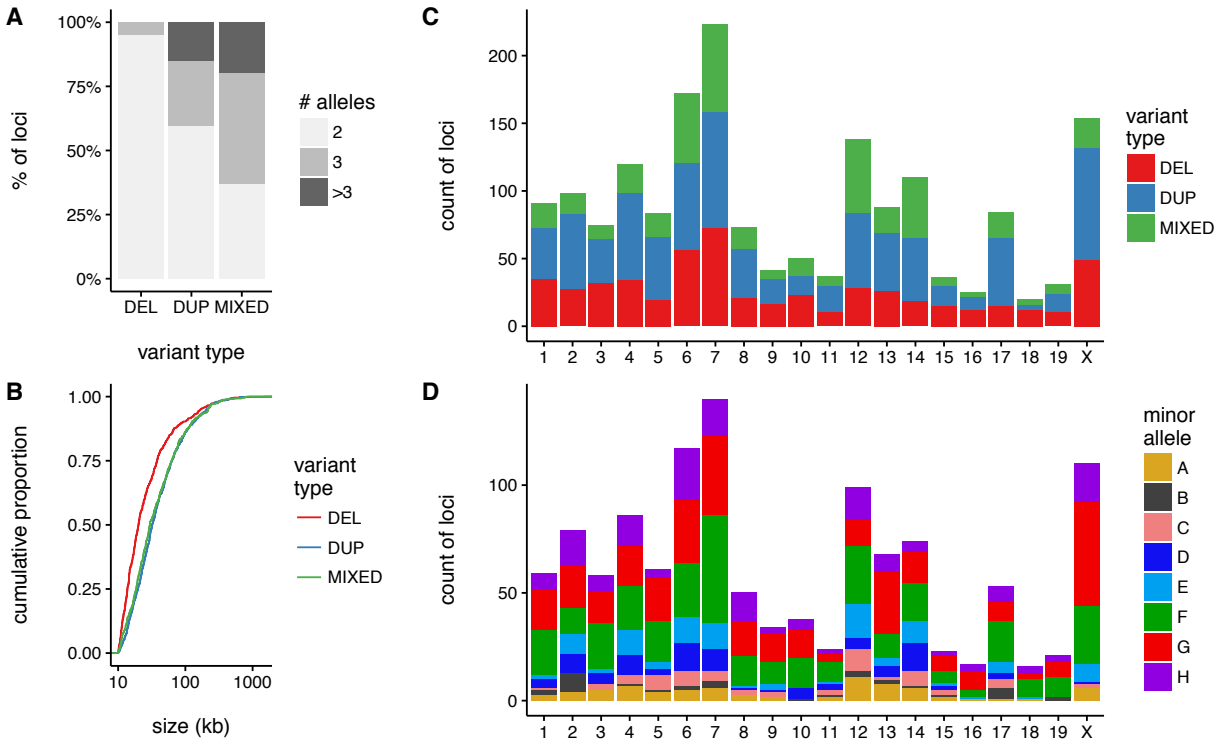

Figure 3.26: Properties of CNVs ascertained in the DO. (**A**) Proportion of CNV loci according to variant type (DEL = deletion, DUP = duplication, MIXED = complex variants) and number of alleles in the DO. (**B**) Cumulative distribution of nominal size of CNV loci by variant type, on log10 scale. (**C**) Count of variants per chromosome, by type. (**D**) Count of private variants (minor allele found in exactly one founder strain) per chromosome.

population frequency of the corresponding haplotypes, in 500 kb windows across the autosomes and X chromosome. The score has expected value 1, and larger values indicate more extreme departures from random joining of haplotypes according to their population frequency. The information score clearly tends to take larger values within cold regions on both the autosomes and the X chromsome ($p < 10^{-5}$, Wilcoxon rank-sum test) (**Figure 3.27A-B**). Within CNV loci that overlap cold regions, we next asked whether haplotypes with the same copy number were more likely to recombine than haplotypes with different copy numbers. This association, measured as the odds ratio, is shown in **Figure 3.27C**. The distribution of odds ratios calculated from observed crossover frequencies and copy numbers is significantly different from that calculated over 1000 random permutations of copy numbers across founder strains ($p < 10^{-5}$, Wilcoxon rank-sum test.) Nonetheless, copy number alone appears to be at best a very weak predictor of crossover patterns in the vicinity of coldspots. This suggests that other forms of structural variation besides simple changes in dosage are important in shaping the landscape of recombination at megabase scales.

3.2.7 Coldspots have epigenetic features of inactive chromatin

The key signals dictating the position of DSBs and ultimately meiotic crossovers are epigenetic. In primates and in mouse, the H3K4me3 mark established by PRDM9 designates a subset of the tens of thousands of available hotspots for DSBs. In taxa without a functional PRDM9 homolog, such as birds[212], dogs[215] and yeast[214], recombination is directed towards gene promoters and CpG islands. The mechanism for this targeting is not clear, but may be related to DNA methylation or to features of chromatin architecture such as nucleosome spacing. We asked whether a complementary suite of epigenetic features could be defined for coldspots, with a view towards identifying properties that could explain the decoupling of the spatial distribution of DSBs from crossovers. Our prototype is the sex chromsomes in male meiosis: DSBs occur on both the X and the male-specific region of the Y, but (under normal circumstances) none of these result in crossovers.

We first examined the association between gene expression and coldspot status in male germ cells. Because DSBs are formed during leptotene and chiasmata are established by pachytene, we re-analzed a published RNA-seq experiment[240] on leptotene/zygotene and pachytene/diplotene spermatocytes isolated by fluorescence-activated cell sorting. Median expression of genes in coldspots is almost ten-fold lower than genes outside coldspots ($p < 10^{-5}$, Wilcoxon rank-sum test), and the effect holds for both the autosomes and the X chromosome (**Figure 3.28**). This may be

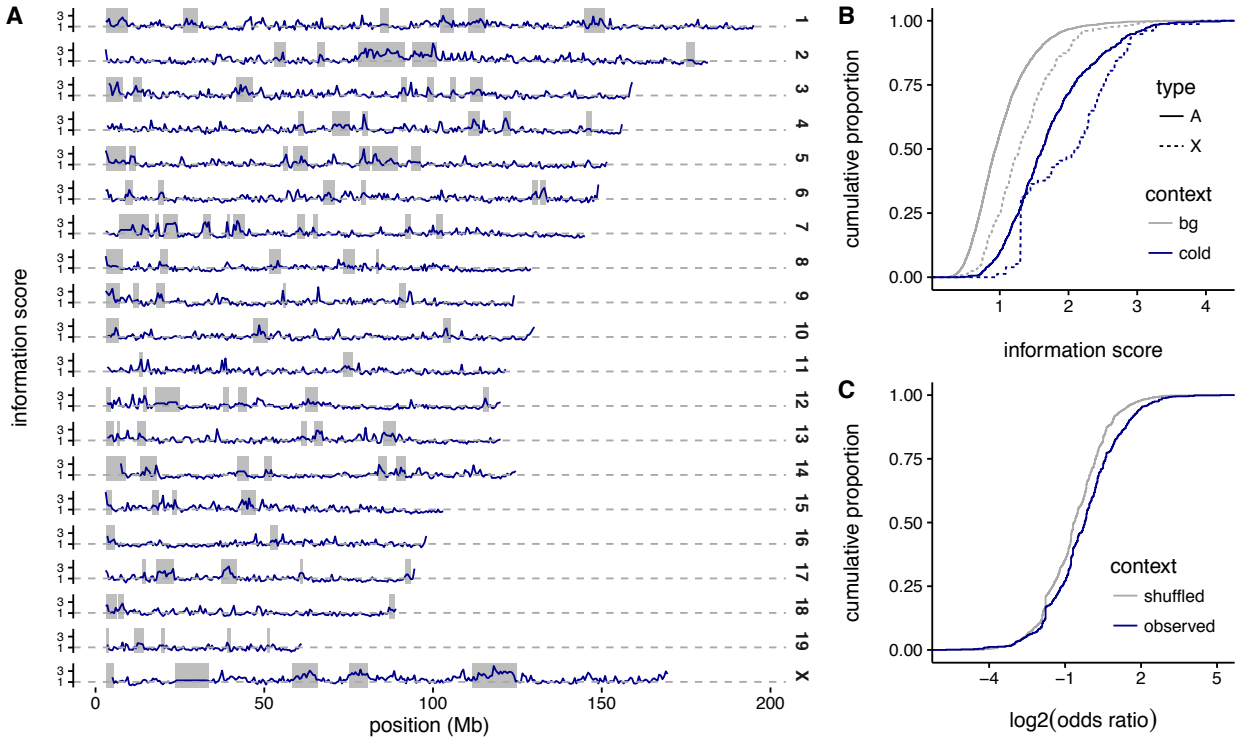

Figure 3.27: Biased distribution of crossovers across strain pairs in the vicinity of coldspots. (**A**) Crossover information score, measuring departure from expected frequency of crossovers with respect to founder strain pairs, in 500 kb windows across the genome. (**B**) Cumulative distribution of the information score in cold regions (blue) versus the remainder of the genome (grey), calculated separately for the autosomes (A) and X chromosome (X). (**C**) Cumulative distribution of odds ratios for association between copy number and crossover incidence (see main text) for CNV loci in cold regions, compared to 1000 permutations.

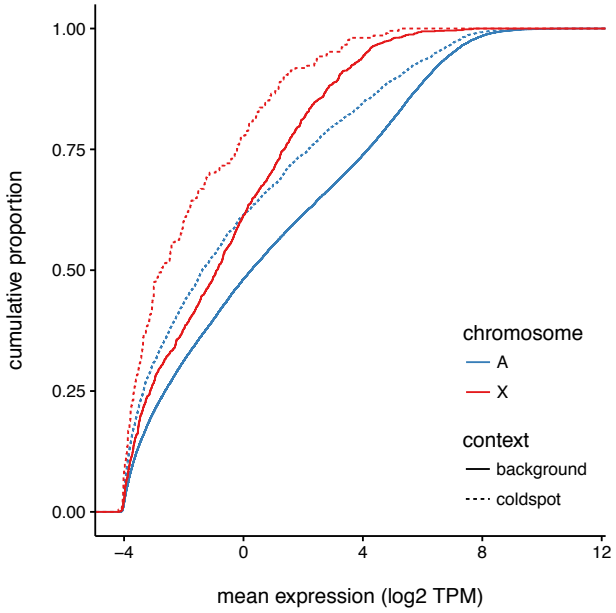

Figure 3.28: Distribution of normalized expression values (transcripts per million, TPM) for genes within (dashed lines) and outside (solid lines) coldspots, measured by RNA-seq in leptotene/zygotene and pachytene/diplotene spermatocytes.

due in part to meiotic silencing of unpaired chromatin (MSUC), which has begun to take hold by late pachytene.

It has been shown that PRDM9 binding is decreased in regions with features characteristic of heterochromatin, including the H3K9me2 mark and association with the nuclear membrane[207]. We hypothesized that coldspots are depleted in crossovers in part because they are heterochromatic or heterochromatin-like. To this end we re-analyzed two ChIP-seq experiments performed in spermatocytes, one against the PRDM9-mediated H3K4me3 mark and one against the H3K9me2 heterochromatin mark, for evidence of enrichment or depletion in coldspots. Rather than identify discrete methylation peaks — which are prone to technical artifacts in duplicated and repetitive sequence — we compared the density of ChIP-seq reads in coldspots versus random genomic intervals of equal size (**Figure 3.29**). As expected, H3K4me3 signal associated with recombination hotspots and active promoters is decreased in coldspots on both the autosomes and sex chromosomes ($p < 10^{-5}$, Wilcoxon rank-sum test.) Likewise the heterochromatin-associated H3K9me2 signal is increased in coldspots compared to the genomic background — but only on autosomes ($p < 10^{-5}$,

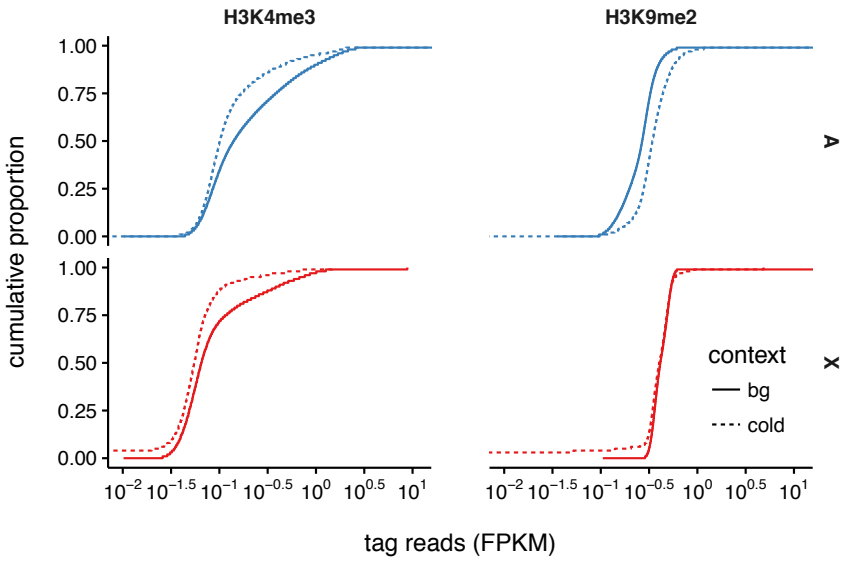

Figure 3.29: Distribution of ChIP-seq signal density, measured as reads per million base pairs of target (FPKM), for the hotspot-associated H3K4me3 mark and the heterochromatin-associated H3K9me2 mark. Dashed lines, distribution in coldspots; solid lines, distribution in random genomic windows having the same size distribution as coldspots. Autosomes and X chromosome are plotted separately because DSB repair on the sex chromosomes is temporally and spatially distinct from the autosomes during male meiosis.

Wilcoxon rank-sum test.) The discrepancy between the X chromsome signal for H3K4me3 and H3K9me2 likely reflects the sequestering of the X chromosome into the transcriptionally-inert sex body over the course of meiotic prophase. Taken together, levels of gene expression and these two canonical histone marks indicate that coldspots correspond to transcriptionally-repressed, closed chromatin during male meiosis.

3.2.8 Coldspots are not unique to the rodent lineage

Both the global rate and the fine-scale spatial distribution of recombination are rapidly evolving in the mouse lineage. Coldspots might be a byproduct of this process rather than a common feature of mammalian meiosis. To test the generality of coldspots we sought a second mammal for which there exists both a high-quality reference genome assembly and a dense pedigree-based genetic map. (Genetic maps derived from patterns of LD in populations are prone to artifacts in SDs and other repetitive sequences where it is difficult to ascertain variants using short-read sequencing.) We

chose the domestic dog (*Canis lupus familiaris*): large pedigrees are available[241], and like the mouse, the dog has an all-acrocentric karyotype, mitigating the possibility of confounding of coldspots with the centromere effect[242]. More interestingly, domestic dogs and other canids apparently lack a functional PRDM9 ortholog[213], providing an opportunity to test whether coldspots are independent of PRDM9 and therefore of a lineage-specific spatial distribution of recombination hotspots.

We re-analyzed a published genetic map derived from a golden retriever pedigree spanning approximately 408 effective meioses[241]. Local sex-specific recombination rates across the 38 dog autosomes are shown in **Figure 3.30**. The dog map recapitulates the major feature of the mouse map: elevated recombination rate in the distal portion of chromosomes in males but not females. Applying the same strategy as we used for the sex-averaged mouse map from the DO, we identified 66 coldspots on 13 chromosomes in the sex-averaged dog map. They are larger than coldspots in mouse — ranging in size from 400 kb to 11.4 Mb — and cover a slightly smaller fraction of the autosomes (3.9%) as a consequence of the lower density of the dog map, and therefore lower power to discriminate true coldspots from stochastic variation in the background recombination rate. As with mouse, coldspots in the dog genome are 2.5-fold enriched for SDs ($q < 0.001$). Example coldspots on dog chromosomes 9 and 22 are shown in **Figure 3.31**.

It is well-known that the broad-scale features of the recombination map are conserved across mammals[243]. Similarity in the size and sequence features of coldspots between dog and mouse, whose last common ancestor lived approximately 55 million years ago[244], suggests that coldspots associated with SDs and/or SVs are also a general feature of meiosis in mammals. They are independent of PRDM9 and of the fine-scale distribution of DSBs.

3.3 Discussion

3.3.1 Sex differences in recombination

Two general rules seem to apply to sex-specific linkage maps from many animal taxa: the overall recombination rate is lower in the heterogametic sex (in mammals, the male); and recombination in the heterogamtetic sex is biased away from the centromere[245,246]. We have replicated this result quite robustly in the G_2:F_1, and provide evidence that weaker crossover interference in females is at least partly responsible for the differences in overall recombination rate between the sexes,

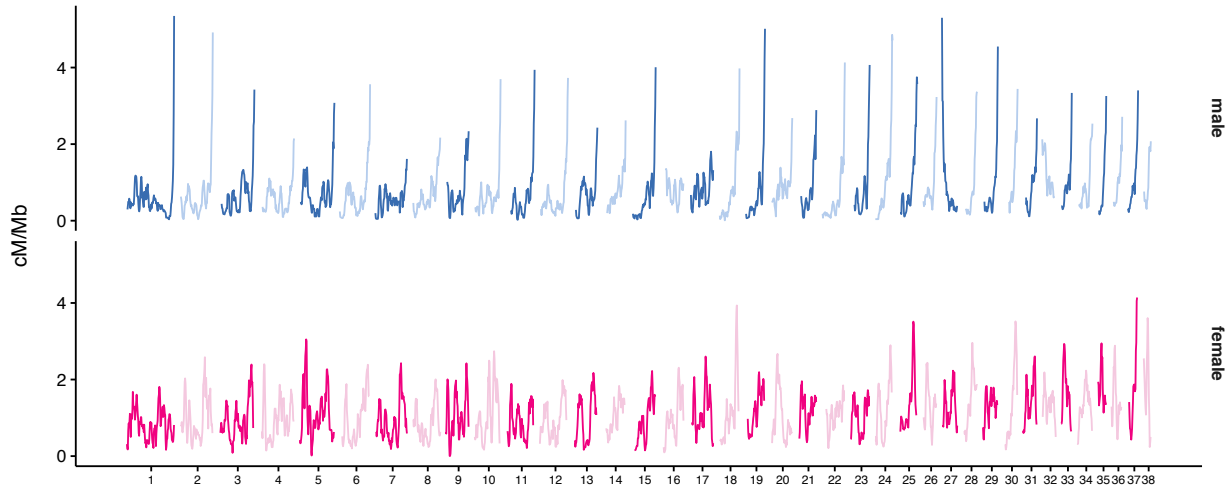

Figure 3.30: Sex-specific recombination rates in the domestic dog. Local sex-specific recombination rates (cM/Mb), calculated in 5 Mb windows with 1 Mb offset between adjacent windows, in a golden retriever pedigree[241]. Note that the orientation of chromosomes 27 and 32 appear to be reversed with respect to the centromere.

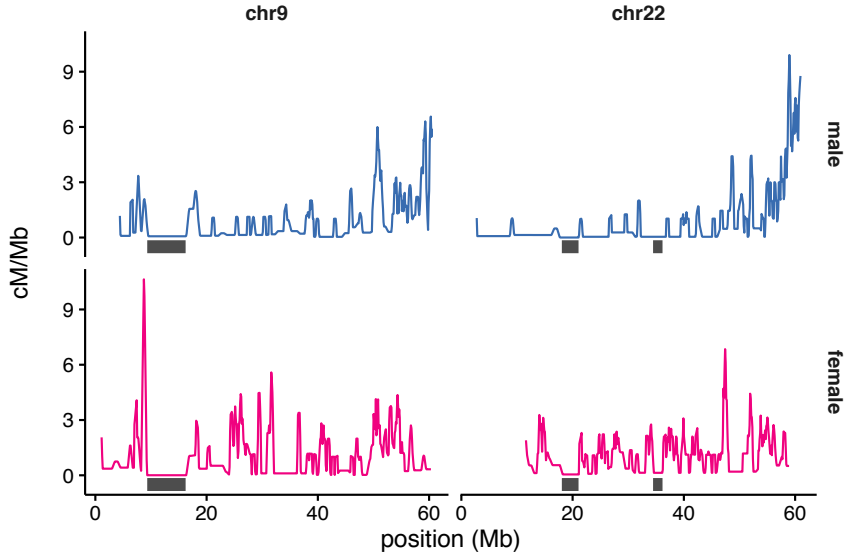

Figure 3.31: Recombination coldspots in the domestic dog. Local sex-specific recombination rates (cM/Mb), calculated in 500 kb windows, on portions of dog chromosomes 9 and 22. Coldspots ascertained on the sex-averaged map are indicated by grey bars along the x-axis. Due to variation in the distribution of informative markers in the pedigree, male and female maps do not necessarily cover the same genomic territory.

consistent with other studies[187]. The dramatic sex difference in spatial distribution of crossovers remains to be explained.

We have previously argued that the tendency of males to recombine in subtelomeric regions is a side effect of the requirement for a crossover in the pseudoautosomal region(s) at the distal tips of the X and Y chromosomes[247]. That now seems unlikely. Sex-specific linkage maps from multiple species of birds[212], fish[248,249] and frogs[250] show sex differences similar to linkage maps in mammals. The distally-biased pattern thus appears to be the ancestral one. If this is the case, then it must be independent of the morphology of the sex chromosomes, which differ widely between and within major vertebrate taxa. Redistribution of crossovers more uniformly along the chromosome in females might be an adaptation to maintain the stability of bivalents during the extended arrest that characterizes oogenesis. However, crossovers too near the centromere are associated with aneuploidy[12].

An alternative explanation invokes meiotic drive. Drivers closely linked to the centromere can distort meiosis I only, while those not linked to the centromere can drive at meiosis II only[14]. Population genetic models predict that alleles that modify the recombination rate can increase in frequency by ablating meiotic drive[251,252]. When meiotic drive is sex-limited — as is the case for true meiotic drive during female meiosis — selection favors sex-limited modifiers of the recombination rate and thereby leads to differences in recombination rates between males and females. When drive occurs at meiosis II, selection favors alleles unlinked to the drive that increase recombination between the drive and the centromere; all other things being equal, the net effect will be a more homogeneous distribution of crossovers along chromosome arms[246].

3.3.2 Effects of paternal age on recombination

Well-powered pedigree studies in humans have shown that the paternal age effect, if one exists at all, must be much smaller than the maternal age effect[231,233]. By contrast, cytological studies of mouse spermatocytes demonstrated a robust increase in the number of chiasmata per cell among older compared to younger males from the same strain[235]. The rate of asynapsis, especially for the sex chromosomes, appeared to be elevated in older male mice during meiosis I. Comparatively low rates of aneuploidy in meiosis II led the authors to conclude that cells with aberrantly paired chromosomes were effectively eliminated by the end of meiosis I irrespective of age. Our study is the first to demonstrate an effect — albeit a weak one — of paternal age on

the number of transmitted crossovers. We show that the effect can be attributed to a mild increase in the number of double-recombinant chromosomes rather than to any detectable change in the strength of interference. Absence of an age effect on the recombination rate in the pseudoautosomal region is consistent with a strict quality-control mechanism during the first meiotic prophase.

Older males could transmit more crossovers for several reasons. Expression of *trans*-regulators of DSB formation such as *Prdm9* might increase with age. Balance between the crossover and non-crossover pathways might be altered such that the number of DSBs is equal between old and young males, but more are repaired as crossovers in older males. Each of these hypotheses is readily testable. Alternatively, epigenetic changes in the germline of older males might act in *cis* to increase the rate of recombination. There is some evidence that the methylation decreases with age around the periphery of CpG islands and in sites of nucleosome binding[253]. Hypomethylation might be propagated, via changes in nucleosome affinity into increased binding avidity for PRDM9, to net increase in the strength of recombination hotspots. If so, we would predict changes in the fine-scale distribution of DSBs in younger versus older males.

It should be noted that, whatever the mechanism for the paternal age effect on recombination, it is qualitatively different from the maternal age effect. Spermatogenesis occurs continuously over the reproductive lifetime of male mammals, so sperm of aged males are the product of meioses initiated in old age. This leaves little opportunity for selection against a class of spermatocytes with fewer chiasmata. By contrast, eggs that complete meiosis in aged females represent the survivors among a limited pool of primary oocytes. The kinetics of the oocyte pool remain the subject of some controversy[227,254]. Nonetheless, a positive correlation between female recombination rate and reproductive success[231] argues that the maternal age effect on recombination can be attributed at least in part to selection within the germline for oocytes with more chiasmata. This explanation is compatible with the (strong) maternal age effect on aneuploidy.

3.3.3 Recombination-rate variation and speciation

The two hybrid sterility loci of largest effect in intersubspecific crosses — on chromosomes 17 (*Prdm9*[153]) and X (*Hstx1*[152]) — are both modifiers of the recombination rate. Important post-zygotic reproductive barriers in mouse are thus mediated by the recombination machinery. The CC and DO, which mix the genomes of all three subspecies, provide a unique opportunity to estimate the effects of modifiers of the recombination rate in a fully randomized genetic background.

Population genetic theory predicts that hybrid breakdown should first affect the heterogametic sex — Haldane's rule[255] — and that incompatibilities between diverging taxa should accumulate disproportionately on the sex chromosomes when partially recessive[256]. Most theoretical and empirical attention has focused on the role of the X chromosome[257,258,259]. We confirm the presence of an X-linked locus in CAST/EiJ that increases the rate of recombination[175].

Our study is the first report of a Y-linked modifier of the recombination rate: in the G_2:F_1 population, the PWK/PhJ Y is associated with an extra 1.1 crossovers per meiosis relative to the Y chromosome of classical inbred strains, and an extra 2.4 crossovers relative to the Y chromosome of CAST/EiJ. The interpretation of these results deserves some caution, as the design of the CC ensures that a Y chromosome is never tested against an X chromosome from the same strain. In particular this means that the effect of the PWK/PhJ and CAST/EiJ Y chromosomes cannot be tested in the presence of an X of the same subspecific ancestry. Given the existence of large-effect modifiers of recombination on the X we cannot rule out the possibility that what we observe in the G_2:F_1 as a marginal effect of the Y is in fact a byproduct of X-Y interactions. The role of the Y chromosome in hybrid sterility appears to be minor[260] and post-meiotic[261].

We identified one X-Y interaction effect on recombination rate: for CC funnels in which an NZO/HlLtJ X and a PWK/PhJ Y can meet in G_2, the rate of recombination is reduced by nearly 30%. A direct test in G_1 is not possible because the NZO/HlLtJ$\times$PWK/PhJ cross (but not the reciprocal cross) produces few, if any, viable offspring, the reasons for which are unknown [4]. Our result raises the possibility that there may be general developmental defects in the (NZO/HlLtJ$\times$PWK/PhJ)F_1 due to antagonistic interactions between the sex chromosomes. These may be unmasked in G_2 after being uncoupled from the autosomal background. If this is true, we would predict negative LD between the Y and one or more loci on the X in the DO. We caution that the models fit for G_2 recombination rates use *expected* X-chromosome dosage based on funnel order, since G_2 genomes are not directly observed. It is possible that, in the 9 funnels in which the NZO/HlLtJ X and the PWK/PhJ Y could meet in a G_2 male, the NZO/HlLtJ X was not transmitted by the G_1 female.

In addition to testing hypotheses regarding genetic control of the overall recombination rate in

[4]Fernando Pardo-Manuel de Villena and Ginger Shaw, personal communication

the G_2:F_1, we have investigated the fine-scale distribution of recombination in the DO. Despite our limited resolution to define the exact position of crossovers, only around half of crossovers in the DO have any overlap with known recombination hotspots. This reflects the incompleteness of our catalog of hotspots: high-throughput analyses of DBSs via ChIP-seq have provided strong evidence that essentially all DSBs occur in hotspots[218]. Our study is the first to analyze an extremely dense recombination map constructed from crossovers rather than recombination precursors (H3K4me3 peaks, DSB sites or chiasmata) in a genetically-diverse population. It is clear that the fine-scale landscape of recombination is distinct from crosses between two inbred strains. Four *Prdm9* alleles are segregating in the DO and most individuals will be functionally heterozygous. When two PRDM9 variants are present in heterozygosity, the "stronger" variant acts in a dominant fashion to activate its own cognate hotspots[262], at least in F_1 hybrids. Hotspot usage in a DO mouse will thus depend on its *Prdm9* diplotype as well as the frequencies of the cognate haplotypes in the rest of the genome. In this setting we predict that most crossovers are initiated at ancestral rather than co-evolved hotspots[209,217]. Given sufficient resolution on a large sample of crossovers — for instance, using whole-genome sequence data — we could obtain the distribution of PRDM9 binding motifs and estimate which PRDM9 allele(s) are most active in the DO.

3.3.4 The relationship between structural variation and recombination

The degree of suppression of crossovers near clusters of CNVs in both the CC and DO is striking. Some of these coldspots, such as the example on chromosome 12 shown in figure **Figure 3.23**, have only one or two crossovers in the more than $15,000$ effective meioses in the DO. Overlap with repetitive sequences makes coldspots suspect for technical artifacts: do we simply lack the power to observe crossovers in these regions of the genome? It is the case that SNP markers on the genotyping arrays used in this study are based away from duplications and repetitive sequences[95]. However, unlike ChIP-seq based DSB maps, we require only information at flanking markers to identify a crossover. Provided crossovers are at least a few megabases apart — as is clearly the case, at least through generation 21 (**Figure 3.4**) — the probability of completely missing a crossover is negligible. (Even for closely-spaced crossovers in successive generations, the probability that flanking haplotypes are the same is only $\frac{1}{8}$.) This is the special value of genetic data and the design of the DO. An alternative explanation for coldspots might be systematic genotyping error in and near repetitive regions. But genotyping error tends to increase, not decrease, the number of inferred

crossovers in the DO and other multiparental populations[263].

The classical explanation for regional absence of crossovers in pedigrees is inversion[51]: the reciprocal products of (odd numbers of) crossovers in inversion heterozygotes are acentric and dicentric, respectively, and cannot be properly segregated to gametes. Inversions large enough to by cytologically visible have been described in many species[264]. Besides the *t*-haplotype and major histocompatibility complex (MHC) on proximal chromosome 17[265], there are few reports of inversions segregating in natural populations of mice. We hypothesize that recombination coldspots correspond to regions of complex structural rearrangement that are highly polymorphic within and between mouse subspecies. High-throughput sequencing with short reads, especially at sparse coverage, is well-suited to identifying changes in copy number but not to characterizing changes in kilobase- to megabase-scale rearrangements. The issue is further complicated by reliance on a reference genome assembly from C57BL/6J that may bear little resemblance to the underlying genomic organization in other strains. What we have detected as CNVs are likely, at least in part, a proxy for more complex rearrangements; consequently, copy number alone is only a weak predictor of crossover patterns. When different structural alleles meet at meiosis, we predict that pairing and synapsis are disrupted. Coldspots are thus excluded from the chromosomal axes where crossovers are formed. The qualitative result is the same as for classical inversions but at smaller scale.

Other explanations are possible. Coldspots have epigenetic features consistent with closed chromatin, and therefore may simply not be accessible to the protein complexes that generate DSBs. But this fails to account for the degree of overlap between coldspots and CNVs, unless CNVs tend to be confined to closed chromatin — in which case the underlying association is again with structural variation.

Enrichment of SDs and CNVs in coldspots presents a paradox: SDs are thought to be sites of recurrent rearrangement because they are prone to non-allelic homologous recombination (NAHR) between paralogous copies of a repeated sequence, yet we have shown that recombination is suppressed across SDs. How, then, does the structural complexity of coldspots arise? We offer two hypotheses to resolve the paradox that are not mutually exclusive. First, recombination in coldspots may be biased towards non-crossover outcomes. In the presence of duplicated sequences, non-homologous exchanges can occur between sister chromatids or within the same chromatid, and some of the products will be mutations[47]. Defects in pairing between homologous sequence in

heterozygotes or hemizygotes for different structural alleles may also predispose to mutation, as on the human Y chromosome[266]. Second, mutations in coldspots may be associated with errors of DNA replication rather than meiosis. This class of mechanisms is triggered either by DNA damage or stalling of replication forks and can give rise to novel alleles of arbitrary complexity[44]. That SVs in human have predominantly paternal origin also suggests that they are disproportionately associated with replication[75]. Replication-based mutations cannot explain the congregation of new SVs near existing duplications. In any case, the structural mutation rate in coldspots must be quite high and the resulting alleles not too deleterious: mutliallelic SVs are common in both mouse[43,267,268] and human populations[39,36]. We show evidence for an elevated mutation rate at one large SV in mouse, *R2d2*, in Chapter 4.

3.4 Conclusions and future directions

Detailed characterization of the sequence and organization of highly polymorphic, structurally complex duplication clusters would shed light on the nature of recombination coldspots. This will require a combination of third-generation sequencing using longer reads and more traditional molecular biology. We predict that these analyses will reveal a level of polymorphism that dwarfs the variability of unique sequence[33]. Without reasonably complete sequence of a representative panel of alleles, our inferences regarding their evolution and functional relevance will be limited.

The combination of genetic mapping with whole-genome sequencing (WGS) in a segregating population like the DO is a powerful approach for learning about genome organization. Genetic data has special value in that it is robust to many of the artifacts to which high-throughput sequencing is vulnerable. Alignment of sequencing reads implies a strong assumption regarding the correctness of the reference genome assembly, as well as its collinearity with the haplotypes segregating in the population under study. Genetic mapping provides orthogonal proof of the location of variants detected by sequencing. We have used genetic mapping to confirm the position of CNVs and to overcome noise in assignment of copy number at loci overlapping segmental duplications. This approach could be extended to any variant that is (1) detectable in sequencing reads, with or without alignment to a reference; and (2) polymorphic in the population. For example, the mouse reference genome includes 44 contigs or scaffolds not placed in the main assembly path, mostly corresponding to clusters of large segmental duplications. If these have variable copy number among the DO founder strains, as is likely, the DO WGS data could be used

to localize them. The same could apply to sequence families that are completely unrepresented in the reference assembly such as telomeric and pericentromeric satellite repeats.

These prospects underscore the enduring utility of classical genetic data for studying the structure of the genome and the basic biology of the germline.

3.5 Materials and methods

3.5.1 Mice

Collaborative Cross G_2:F_1 population. The G_2:F_1 mice used in this study were bred at Oak Ridge National Laboratory (ORNL) beginning in 2005 as described in detail elsewhere[269,221].

Diversity Outbred population. Breeding and maintenance of the DO at the Jackson Laboratory is described elsewhere[222,270].

Aged male pedigrees. Intercrosses used to test for the effect of advanced paternal age on recombination were generated beginning in 2009 at the University of North Carolina (UNC). Briefly, all possible reciprocal crosses were performed between CAST/EiJ, PWK/PhJ and WSB/EiJ. Male F_1s were singly housed and bred at approximately $8-12$ weeks of age and again at >18 months of age to young, fertile FVB/NJ females. Progeny were sacrificed at birth by cervical dislocation.

All mice were treated in accordance with the recommendations of the Institutional Animal Care and Use Committee of ORNL, the Jackson Laboratory, and UNC, respectively.

3.5.2 DNA preparation

High molecular weight DNA from G_2:F_1 mice was extracted from tail tissue at UNC with a standard phenol-chloroform method . For intercross and DO mice, low molecular weight DNA was extracted by many different investigators using several standard methods as described in[270].

3.5.3 Genotyping

G_2:F_1 mice were genotyped at the Jackson Laboratory using the Mouse Diversity Array[92]. DO mice were genotyped on either the MegaMUGA (77,808 markers) or GigaMUGA ($143,259$ markers arrays[95] by the commercial service of Neogen/Geneseek, Inc (Linclon, NE). Genotypes for DO mice were generously contributed by many investigators and curated at the Jackson Laboratory as described in[270]. Samples with $>10\%$ missing calls were excluded.

3.5.4 Haplotype reconstruction

Analyses of recombination in multiparental populations depend on *haplotype reconstruction* — inference of the mosaic structure of individual genomes in terms of the founder individuals. In the G_2:F_1 all individuals are obligate heterozygotes, so there exist $8 choose 2 = 28$ possible diploid genotype (diplotype) states. For the DO an additional 8 homogzygous states are possible. We used hidden Markov models (HMMs) to estimate the posterior probability of each of the 28 (G_2:F_1) or 36 (DO) possible diplotype states for each individual, given that individual's observed genotypes and the observed genotypes of multiple biological replicates of founder strains and F_1s. The most likely founder mosaic was obtained by selecting the diplotype state with the highest posterior probability at each marker. Diplotypes were phased to haploid chromosomes using pedigree information in the G_2:F_1 or a heuristic in the DO (see below). Junctions between haplotype blocks in this phased mosaic represent crossovers. Each crossover in our dataset is represented as the interval between the last informative marker in the proximal haplotype block and the first informative marker in the distal haplotype block.

Collaborative Cross G_2:F_1 population. The funnel breeding scheme of the CC imposes a rich set of constraints on the possible diploid genotypes at the G_2:F_1 generation. The software `GAIN`[271] exploits these constraints to infer fully phased haplotypes given genotypes for a pair of G_2:F_1 siblings and the funnel order at G_0. The inference procedure was split into two steps: a first pass using only $121,504$ markers successfully typed in $> 99\%$ of individuals; and a second pass using an additional $549,595$ markers to refine crossover boundaries.

Diversity Outbred. Haplotypes for DO mice were reconstructed using the HMM module of `DOQTL`[263] with genotypes from MegaMUGA ($68,268$ QC-passing markers) or GigaMUGA ($120,789$ QC-passing markers) as input. Diplotypes were "pseudo-phased" using a greedy algorithm: moving left to right along each chromosome, choose the configuration that minimizes the total number of crossovers. In sibships identified based on kinship estimates from SNP genotypes, we attempted to improve phasing using a dynamic programming algorithm. For a given chromosome pair (*e.g* chromosome 1), there are $2k$ chromosomes in a group of k siblings. An individuals chromosomes can have at most two phasing configurations (only one when homozygous), so there are 2^k possible configurations in the sibship. We used a scoring function that gives equal weight to every crossover and used dynamic programming to choose the state path that minimizes the

total number of crossovers in a chromosome pair across the k siblings. Only 0.5% more crossovers were shared between siblings after phasing improvement. We concluded that greedy phasing is an acceptable heuristic for our purposes.

3.5.5 Pedigree reconstruction in the DO

Kinship coefficients were estimated for all individuals within generations from SNP genotypes using `KING` v1.4[272]. Markers with $> 10\%$ missing data or $< 5\%$ minor allele frequency were removed. Unlike some other kinship estimators, the estimator $\hat{\pi}$ implemented in `KING` does not require that markers be in linkage equilibrium and its sampling variance decreases as the number of markers increases. Approximately $66,000$ autosomal SNPs were used at each generation. We used $\hat{\pi} > 0.15$ as a cutoff for siblings (relationship degree 2) and $\hat{\pi} > 0.10$ as a cutoff for cousins (relationship degree 3), based on inspection of the distribution of pairwise kinship coefficients across all generations.

3.5.6 Estimation of genetic maps in the G_2:F_1

Cumulative genetic maps for the CC funnels were computed directly from the interval representation of crossovers by integration across each chromosome to account for uncertainty in localization of crossovers. Maps were obtained separately across several slices of the data — males versus females, G_1 versus G_2 — in addition to the overall sex-averaged map. Crossover counts were converted to centimorgans using the formula

$$\text{cM} = 100 \times \left(\frac{\text{number of crossovers}}{\text{number of meioses} \times \text{number of funnels}} \right)$$

The effective number of meioses varies over different slices of the data. At G_2, for example, all four meioses in the funnel are fully observed, but only an average of three G_1 meioses can be observed per funnel.

3.5.7 Estimation of genetic maps in the DO

Estimation of the genetic map in the G_2:F_1 is far more challenging than in the G_2:F_1. Traditional approaches to the construction of linkage maps in pedigrees assume that every crossover is distinct and can be attributed to at most two specific meiosis (in the case of unknown phase.) In the DO, however, crossovers cannot be uniquely assigned to a specific meiosis; the number of effective meiosis is unknown; and the same crossover may be observed multiple times if it is

shared IBD between two or more individuals. We first sought to reduce our total dataset of 2.2 million crossovers to a set of distinct crossovers. To do so, we identified crossovers with the same haplotypes at the junction (*e.g.* HF) and overlapping coordinates in overlapping windows of two generations (G_{n-1},G_n). For each overlap, we identified the individual *chromosomes* on which the crossovers were identified, and tested whether the next crossover or the previous crossover on the same chromosomes were also shared. If at least one other neighboring crossover was shared between the chromosomes, we considered the entire set of crossovers — both the focal pair and the neighboring pair(s) — as shared. (See **Figure 3.6** for an example.) After performing this analysis in adjacent generations, we constructed a graph of shared crossovers across all generations ($4-21$). Nodes in the graph are individual crossovers, and edges represent sharing between chromosomes. Connected components in the graph correspond to distinct crossovers transmitted over multiple generations; nodes with no incoming or outgoing edges correspond to singletons that are by definition distinct

Next we constructed a cumulative map by integrating across all distinct crossovers on each chromosome. We noticed that the shape of this cumulative map was remarkably similar to the shape of the sex-averaged cumulative map in the G_2:F_1. The two populations share the same eight founders at the same expected allele frequencies, so we reasoned that the G_2:F_1 map could be used as a scaffold for approximating the relationship between centimorgans and distinct crossover counts in the DO. To obtain this approximation we fit polynomial regressions of degree k (using least squares) each chromosome as follows:

$$\text{cM}_{\text{G}_2:\text{F}_1} = 0 + x_{\text{DO}}\beta_1 + x_{\text{DO}}^2\beta_2 + \cdots + x_{\text{DO}}^k\beta_k + \epsilon$$

where x_{DO} is cumulative crossover count in the DO. (Note that the model lacks an intercept term because the genetic map must begin at zero.) The fitted values from this regression were taken as the centimorgan positions on the DO map. We found that rescaling with polynomials of even k or $k = 1$ overestimated map length on every chromosome. Polynomials of odd degree could better accommodate the enrichment of crossovers in subtelomeric regions. For parsimony we chose $k = 3$. Note that it is possible to obtain a non-monotone function from these regressions, which violates a fundamental property of the genetic map, that it be non-decreasing. We confirmed that none of

the fitted models showed any evidence for violation of this property. Strain-specific maps were estimated similarly. Every distinct crossover contributes to two strain-specific maps, corresponding to the two haplotypes at the junction.

If all crossovers in our dataset were truly distinct, then the appropriate scaling would be linear, and the slope would provide an estimate of the number of effective meioses we can observe in our cross-section of the DO. We view the higher-order terms in the polynomial as correction factors for the inclusion of cryptic duplicate crossovers in the map. In practice, however, the linear term dominates: its median value across chromosomes is 1.4×10^{-3}, while the quadratic and cubic terms are $O(10^{-8})$ and $O(10^{-13})$ respectively. Assuming all crossovers can be observed and ignoring interference, the map function is linear and we can define genetic distance as

$$\text{cM} = 100 \times \left(\frac{\text{number of crossovers}}{\text{number of meioses}} \right)$$

And in the rescaled DO map

$$\text{cM} \approx \beta_1 x$$

where x is the number of observed crossovers. By substitution and elimination of terms we obtain

$$\text{number of meioses} \approx \frac{100}{\beta_1} \approx 67 \times 10^3$$

3.5.8 Estimation of genetic maps in intercrosses

Genotype data from the intercrosses used to investigate the paternal age effect were analyzed using `R/qtl` and treated as a backcross, since only one parent (the sire) was segregating. The number of informative markers differed by cross ($13,355$ for CAST/EiJ×PWK/PhJ; $12,510$ for PWK/PhJ×WSB/EiJ) but far exceeded saturation for a cross of this size. Genotyping errors were identified and removed using the `cleanGeno()` function, and the positions of crossovers identified using `findXO()`. Crossovers in the pseudoautosomal region were identified by manual inspection.

3.5.9 Models for crossover interference

Although several statistical models of the crossover interference process have been proposed, we chose to fit the *gamma model* as described in[224]. This model has a single parameter ν for which larger values correspond to stronger positive interference. Briefly, chiasmata are modelled as

arising by a stationary renewal process. The increment (in morgans) between adjacent chiasmata on the same chromosome is distributed as $\Gamma(\nu, 2\nu)$; the expected increment is $\frac{1}{2}$, or 50 cM. We fit the gamma model to crossover-location data from G_2 and intercross meioses using the `xoi` package for `R`[225].

3.5.10 Models for recombination rates

Tests for the effect of X chromosome, Y chromosome and sex on recombination rate in the G_2:F_1 were performed using Poisson regression — a generalized linear model with Poisson-distributed response — with crossover count per meiosis as the dependent variable. Model comparisons were performed by the likelihood-ratio test, and confidence intervals obtained from profile likelihoods. For testing X-chromosome and age effects on intercross meioses, we first fit a linear mixed model to account for repeated measures on the same sire. The variance component attached to the random effect was not different from zero, so we fell back to a simple linear model. (For the intercross data, the linear model and Poisson regression gave a very similar distribution of residuals.)

3.5.11 Identification of recombination coldspots

We identified cold regions using a one-dimensional dynamic programming algorithm to identify regions with 10- old reduction in frequency of crossovers via a generic scoring scheme[273]. Briefly, we first compute local crossover density ρ_i in windows of 500 kb with 100 kb offset between adjacent windows. Those densities are converted to an *excursion score* e_i as follows:

$$e_i = \lambda(1 - \theta) + \rho_i \log \theta$$

where λ is the mean crossover density per chromosome and θ is a pre-specified enrichment or depletion factor. (Tenfold reduction corresponds to $\theta = 10^{-1}$.) Then a forward pass is made over the excursion scores to calculate the final score E_i

$$E_i = \max\{0, E_i + e_{i+1}\}$$

with $E_0 \equiv 0$. Coldspots are finally extracted by performing a traceback over the E_i. This method avoids the need for a fixed-size sliding window.

3.5.12 Whole-genome sequencing in the DO

Whole-genome sequencing of 228 male DO mice from generations $12-17$ was performed at the Wellcome Trust Sanger Institute (Hinxton, Cambridge, UK).[5] Barcoded libraries were prepared from fragmented genomic DNA using the Illumina TruSeq kit and pooled. Paired-end reads (2×125 bp) were generated using 14 lanes of an Illumina HiSeq 2500 instrument, for an approximate coverage of $4\times$ per sample. Integrity of raw reads was confirmed using `FastQC`. Reads for each sample were realigned to the mm10 reference using `bwa-mem` v0.7.12 with default parameters[274]. Optical duplicates were removed with `samblaster`[275].

3.5.13 Discovery and genotyping of CNVs

Multisample CNV discovery and genotyping was performed with `GenomeSTRiP`[276]. Briefly, this software uses depth of coverage and paired-end mapping patterns in multiple samples to identify candidate CNV regions and infer alternate copy-number allele(s) which are then genotyped in each individual sample. We used the `CNVDiscoveryPipeline` module with default settings except as follows: window size 10 kb for initial discovery of candidate variants; minimum mapping quality (MQ) $= 0$ at both the discovery and genotyping stages; and MQ> 10 at for refining candidate CNV boundaries. These MQ settings are much more permissive than the defaults and permit discovery of CNVs over SDs with high pairwise identity, over which few or no reads align with MQ> 0.

CNV discovery in a natural population is complicated by the fact that most novel alleles are rare. The only means of control of the false discovery rate is to apply strict filters for genotype quality — filters which naturally create bias against novel alleles overlying the regions of the genome most prone to structural mutation. We therefore took as our initial CNV callset the output of the penultimate stage of the pipeline, prior to the application of filters for missingness and genotype quality.

Ultimate we sought to assign copy numbers to the eight founder alleles of the DO, not to individual samples. (*De novo* CNVs were beyond the scope of the present investigation.) Using

[5]Samples for whole-genome sequencing were provided by Allan Pack, John Miclot Professor of Medicine at the Perelman School of Medicine of the University of Pennsylvania. Sequencing was performed on behalf of Richard Mott, a principal investigator at the Wellcome Trust Centre for Human Genetics and the University of Oxford. Data was generously made available via a collaboration with Gary Churchill of the Jackson Laboratory.

the copy number assigned to each sample by `GenomeSTRiP` as a quantitative trait, we genetically mapped all candidate CNVs using `R/qtl2` and estimated the founder strain copy numbers as the best unbiased linear predictors (BLUPs) at the QTL peak. CNVs mapping with LOD score < 10 or minor allele count < 5 were excluded as likely false positives or rare variants. Next we merged CNVs with overlapping coordinates and identical strain distribution pattern into single loci. This yielded a final set of $1,749$ CNVs segregating in the DO.

3.5.14 Analyses of ChIP-seq data

ChIP-seq data for the H3K4me3 mark of active recombination hotspots was obtained from the NCBI Short Read Archive, accession #SRP045879[209], and for the H3K9me2 mark of heterochromatin, accession #SRP059590. Reads for each sample were realigned to the mm10 reference using `bwa-mem` v0.7.12 with default parameters[274]. Optical duplicates were removed with `samblaster`[275].

3.5.15 Test for enrichment of sequence features

We used the `GenomicAssociationTester` package for `Python`[277] to test for enrichment of various sequence features (*i.e.* genes, conserved elements, repeat elements) in defined intervals (*i.e.* coldspots.) `GAT` estimates enrichment by comparing observed overlap to overlap between defined intervals and randomly-sampled intervals from the genome having the same size distribution as the query intervals. Annotations for repeat elements (LINE, SINE, LTR) were obtained from the UCSC Genome Browser (`http://genome.ucsc.edu/`). Gene annotations were obtained from Ensembl v86 (`ftp://ftp.ensembl.org/pub/release-86/gff3/mus_musculus/`), and evolutionarily constrained elements from the alignment of 40 eutherian mammal genomes generated by the Ensembl Compara team (`ftp://ftp.ensembl.org/pub/release-86/maf/ensembl-compara/`). Tracts of IBD in classical inbred strains were obtained from the UNC Mouse Phylogeny Viewer (`http://msub.csbio.unc.edu`).

CHAPTER 4

Evolutionary fates of a large segmental duplication in *Mus*

4.1 Introduction

[1] Duplication is an important force shaping the evolution of plant and animal genomes: it provides a substrate for evolution that is, by right of being redundant, transiently free from selective constraint[278]. Segmental duplications (SDs), defined as contiguous sequences which map to more than one physical position[239], are a common feature of eukaryotic genomes and particularly those of vertebrates.

Like any sequence variant, a duplication first arises in a single individual in a population. The distinction between such copy-number variants (CNVs) and SDs is fluid and somewhat arbitrary: tracts of SDs are highly polymorphic in populations in species from *Drosophila*[279] to mouse[43] to human[239]. Studies of parent-offspring transmissions have shown that SDs are prone to recurrent *de novo* mutations including some implicated in human disease (reviewed in[280]). Bursts of segmental duplication have preceded dramatic species radiations in primates, and more broadly, blocks of conserved synteny in mammals frequently terminate at SDs[281,239]. This suggests that SDs could mediate the chromosomal rearrangements through which karyotypes diverge and reproductive barriers arise.

Notwithstanding their evolutionary importance, SDs are difficult to analyze. Repeated sequences with period longer than the insert size in a sequencing library and high pairwise similarity

[1] The results presented in this chapter are published in:

> Morgan AP, Holt JM, McMullan RC, Bell TA, Clayshulte AM, Didion JP, Yadgary L, Thybert D, Odom DT, Flicek P, McMillan L, Pardo-Manuel de Villena F (2016) The evolutionary fates of a large segmental duplication in mouse. *Genetics*: **204**: 267–285. PMID 27371833.

Important contributions were made by John Didion, Rachel McMullan, Matt Holt, Leonard McMillan, Tim Bell, Amelia Clayshulte and Liran Yadgary. Whole-genome sequence data from *M. caroli* was generously shared before publication by David Thybert, Duncan Odom and Paul Flicek.

are likely to be collapsed into a single sequence during genome assembly. Efficient and sensitive alignment of high-throughput sequencing reads to duplicated sequence remains challenging[87]. Genotyping of sites within SDs is difficult because variants between copies (paralogous variants) are easily confounded with variants within copies between individuals at a given copy (allelic variants). Latent paralogous variation may bias interpretations of sequence diversity and haplotype structure[282].

Paralogy also complicates phylogenetic inference. Ancestral duplication followed by differential losses along separate lineages may yield a local phylogeny that is discordant with the genome-wide phylogeny[283]. Within each duplicate copy, local phylogenies for adjacent intervals may also be discordant due to non-allelic gene conversion between copies[284,285].

In this chapter we present a detailed analysis of a segmental duplication, *R2d*, in the house mouse (*Mus musculus*). *R2d* is a 127 kbp unit which contains the protein-coding gene *Cwc22* and flanking intergenic sequence. Although the C57BL/6J reference strain and other classical laboratory strains have a single haploid copy of the *R2d* sequence (in the *R2d1* locus), the wild-derived CAST/EiJ, ZALENDE/EiJ, and WSB/EiJ strains have an additional 1, 16 and 33 haploid copies respectively in the *R2d2* locus. *R2d2* is the responder locus in a recently-described meiotic drive system on mouse chromosome 2 but is absent from the mouse reference genome[55,286]. We draw on a collection of species from the genus *Mus* sampled from around the globe to reconstruct the sequence of events giving rise to the locus' present structure (**Figure 4.1**). Using novel computational tools built around indexes of raw high-throughput sequencing reads, we perform local *de novo* assembly of phased haplotypes and explore patterns of sequence diversity within and between copies of *R2d*.

Both phylogenetic analyses and estimation of mutation rate in laboratory mouse populations reveal that *R2d2* and its surrounding region on chromosome 2 are unstable in copy number. Cycles of duplication, deletion and non-allelic gene conversion have led to complex phylogenetic patterns discordant with species-level relationships within *Mus* which cannot be explained by known patterns of introgression between *Mus* species[136,157].

4.2 Results

4.2.1 Duplication of *R2d* in *Mus* ancestor

In order to determine when the *R2d* CNV arose, we used quantitative PCR and/or depth of coverage in whole-genome sequencing to assay *R2d* copy number in a collection of samples spanning the phylogeny of the genus *Mus.* Samples were classified as having diploid copy number 2 (two chromosomes each with a single copy of *R2d*) or > 2 (at least one chromosome with an *R2d* duplication).

We find evidence for > 2 diploid copies in representatives of all mouse taxa tested from the Palearctic clade[125] (**Figure 4.1**): 236 of 525 *Mus musculus*, 1 of 1 *M. macedonicus*, 1 of 1 *M. spicilegus*, 1 of 1 *M. cypriacus* and 8 of 8 *M. spretus* samples. However, we find no evidence of duplication in species from the southeast Asian clade, which is an outgroup to Palearctic mice: 0 of 2 *M. famulus*, 0 of 2 *M. fragilicauda*, 0 of 1 *M. cervicolor*, 0 of 1 *M. cookii* and 0 of 1 *M. caroli* samples. Outside the subgenus *Mus*, we found evidence for > 2 diploid copies in none of the 9 samples tested from subgenus *Pyromys*. We concluded that the *R2d* duplication most likely occurred between the divergence of southeast Asian from Palearctic mice (~ 3.5 million years ago [Mya]) and the divergence of *M. musculus* from *M. spretus* (~ 2 Mya)[125,287], along the highlighted branch of the phylogeny in **Figure 4.1A**. If the *R2d* duplication was fixed in the ancestor of *M. musculus*, then extant lineages of house mice which have only 2 diploid copies of *R2d* — including the reference strain C57BL/6J (of predominantly *M. musculus domesticus* origin[162]) — represent subsequent losses of an *R2d* copy. Alternatively, the *R2d* duplication may have been polymorphic in the ancestor of *M. musculus* and then continued to segregate in *M. musculs* and *M. spretus*.

Duplication of the ancestral *R2d* sequence resulted in two paralogs residing in loci which we denote *R2d1* and *R2d2* (**Figure 4.1B**). Only one of these is present in the mouse reference genome, at chr2: 77.87 Mbp; the other copy maps approximately 6 Mbp distal[286], as we describe in more detail below. The more proximal copy, *R2d1*, lies in a region of conserved synteny with rat, rabbit, chimpanzee and human[288] (**Figure 4.2**); we conclude that it is the ancestral copy.

The sequence of the *R2d2* paralog was assembled *de novo* from whole-genome sequence reads[29] from the strain WSB/EiJ (of pure *M. m. domesticus* origin[157]), which has haploid *R2d* copy number ~ 34[286]. We exploited the difference in depth of coverage for *R2d1* (1 haploid copy) and *R2d2* (33

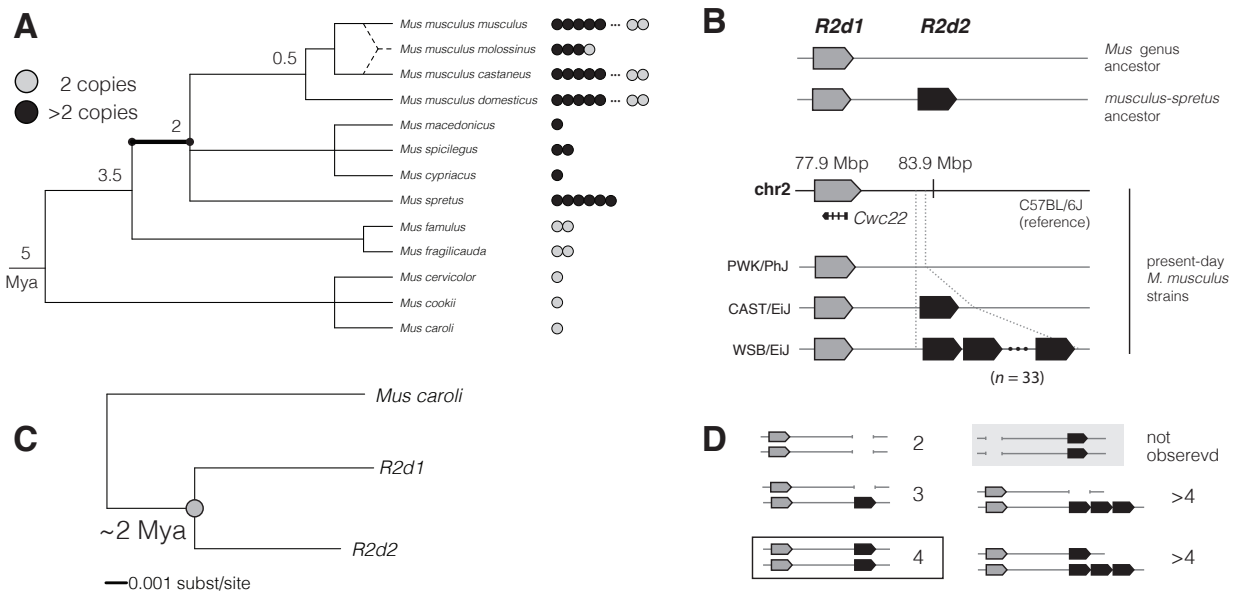

Figure 4.1: Origin and age of the *R2d2* duplication. (**A**) *R2d* copy number across the phylogeny of the genus *Mus*. Each dot represents one individual; grey dots indicate diploid copy number 2 and black dots copy number >2. The duplication event giving rise to *R2d1* and *R2d2* most likely occurred on the highlighted branch. Approximate divergence times (REF: Suzuki 2004) are given in millions of years ago (Mya) at internal nodes. (**B**) Schematic structure of the *R2d1-R2d2* locus. The mouse reference genome (strain C57BL/6J, *M. m. domesticus*) contains a single copy of *R2d* at *R2d1*. Wild-derived inbred strains vary in haploid copy number from 1 (PWK/PhJ, *M. m. musculus*) to 2 (CAST/EiJ, *M. m. castaneus*) to 33 (WSB/EiJ, *M. m. domesticus*). *R2d1* is located at approximately 77.9 Mbp and *R2d2* at 83.8 Mbp. (**C**) Concatenated tree constructed from *R2d1* (reference genome) and *de novo* assembled *R2d2* and *M. caroli* sequences assuming a strict molecular clock. The duplication node is indicated with a grey dot. (**D**) Relationship between observed *R2d* copy-number states and inferred structure of the *R2d1-R2d2* locus. The configuration of the *M. spretus – M. musculus* common ancestor (4 diploid copies) is boxed in black. We have yet to identify samples with diploid copy number 2 but no *R2d1* (grey shaded box).

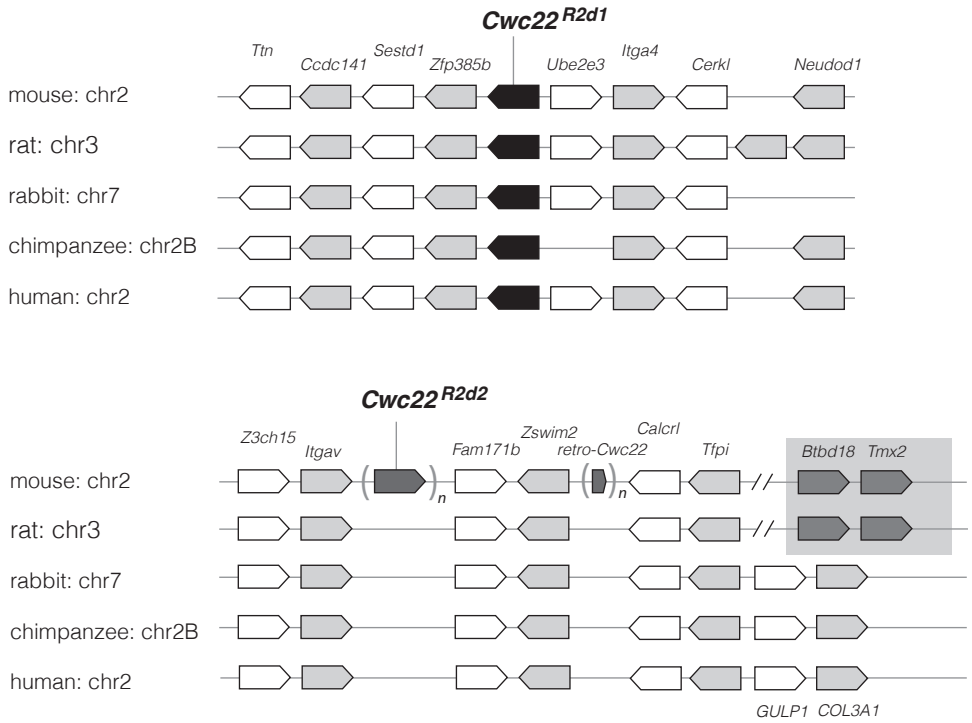

Figure 4.2: Conservation of synteny between mouse and four other mammals around *Cwc22*^{R2d1} (upper panel) indicates that the *R2d1* sequence remains in its ancestral position. Chevrons represent genes, alternating white and grey, and are oriented according to the strand on which the gene is encoded. *Cwc22*^{R2d2} is novel in the mouse but its position relative to genes with conserved order is shown in the lower panel. Note that synteny is disrupted in mouse and rat distal to *R2d2*.

Sequence	Start	End	Strand	Type	Family
R2d2	45011	45166	+	LTR/ERVK	RLTR10
R2d1	1881	3074	+	LINE	L1Md_T
	73374	73896	-	LINE	MT2_Mm
	82227	89459	-	LINE	L1Md_A
	94179	100406	-	LINE	L1Md_F2
	126201	127206	+	LINE	L1Md_F2
	127995	134405	+	LINE	L1Md_Gf
	143125	150112	+	LINE	L1Md_A

Table 4.1: Transposable-element insertions private to *R2d1* or *R2d2*. Coordinates are offsets with respect to the start position of *R2d* (for *R2d1*: chr2: $77,869,657$ in the reference genome; for *R2d2*: the beginning of the *de novo* assembled contig.)

haploid copies) to assign variants to *R2d1* or *R2d2* (**Figure 4.3**). Pairwise alignment of the *R2d2* contig against *R2d1* is shown in **Figure 4.4**. The paralogs differ by at least 8 transposable-element (TE) insertions: 7 LINE elements specific to *R2d1* and 1 endogenous retroviral element (ERV) specific to *R2d2* (**Table 4.1**). (Due to the inherent limitations of assembling repetitive elements from short reads, it is likely that we have underestimated the number of young TEs in *R2d2*.) The *R2d1*-specific LINEs are all $< 2\%$ diverged from the consensus for their respective families in the RepeatMasker database (`http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker`), consistent with insertion within the last 2 My. The oldest *R2d2*-specific ERV we could detect is 0.7% diverged from its family consensus. TE insertions occurring since the ancestral *R2d* duplication are almost certainly independent, so these data are consistent with duplication < 2 Mya. The *R2d* unit, minus paralog-specific TE insertions, is 127 kbp in size. *R2d* units in the *R2d2* locus are capped on both ends by $(\texttt{CTCC})_n$ microsatellite sequences, and no read pairs spanning the breakpoint between *R2d2* and flanking sequence were identified.

In order to obtain a more precise estimate of the molecular age of the duplication event we assembled *de novo* an additional of 16.9 kbp of intergenic and intronic sequence in 8 regions across the *R2d* unit from diverse samples and constructed phylogenetic trees. The trees cover 17 *R2d1* or *R2d2* haplotypes, 13 from inbred strains and 4 from wild mice. The sequence of *Mus caroli* (diploid copy number 2) is used as an outgroup. A concatenated tree is shown in **Figure 4.1C**.

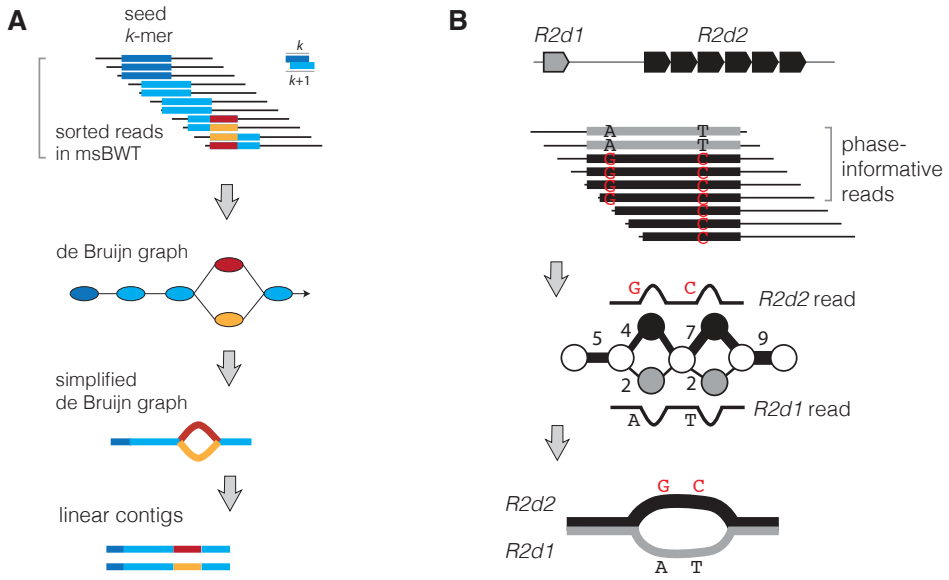

Figure 4.3: Targeted *de novo* assembly using the multi-string Burrows-Wheeler Transform (msBWT). (**A**) The msBWT and its associated FM-index implicitly represent a suffix array of sequencing reads, such that read suffixes sharing a k-mer prefix are adjacent in the data structure. This allows rapid construction of a local de Bruijn graph starting from a k-mer seed (dark blue) and extending by successive k-mers (light blue) containing the $(k-1)$-length suffix of the previous k-mer. A $(k-1)$-length prefix with more than one possible suffix (red and orange) creates a branch point. Adjacent nodes in the graph with in-degree and out-degree one can be collapsed into a single node, yielding a simplified graph, which can then be traversed to obtain linear contig(s). (**B**) Paralogs of *R2d* can be disentangled using the local de Bruijn graph by exploiting differences in copy number. Edges in the graph are weighted by read count, and linear contigs for the *R2d1* and *R2d2* paralogs obtained by traversing the graph in a manner that minimizes the variance in edge weights along possible paths. Phase-informative reads (those overlapping multiple paralogous variants) provide a second source of evidence.

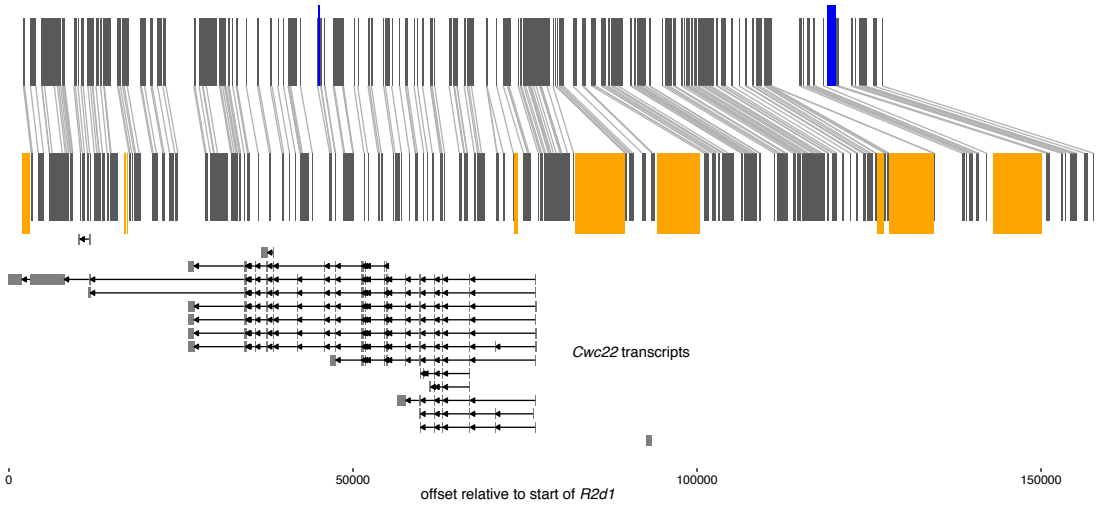

Figure 4.4: Pairwise alignment of *R2d2* contig (top) to the *R2d1* reference sequence (bottom). Dark boxes show position of repetitive elements present in both sequences; syntenic positions are connected by grey anchors, and blank space represents aligned bases in both sequences. Orange boxes indicate position of repetitive elements present in the *R2d1* sequence but not detected in *R2d2*; blue boxes indicate position of elements in *R2d2* but not *R2d1*. *Cwc22* transcripts are shown below the alignment.

Using 5.0 ± 1.0 million years before present (Mya) as the estimated divergence date for *M. caroli* and *M. musculus*[125,287], Bayesian phylogenetic analysis with BEAST v1.8[289] yields 1.6 Mya (95% HPD $0.7 - 5.1$ Mya) as the estimated age of the duplication event that gave rise to *R2d1* and *R2d2*. Although the assumption of a uniform molecular clock may not be strictly fulfilled for *R2d1* and *R2d2*, the totality of evidence — from presence/absence data across the mouse phylogeny, paralog-specific TE insertions, and sequence divergence between paralogs — strongly supports the conclusion that *R2d* was first duplicated within the last 2 My in the common ancestor of *M. musculus* and *M. spretus*.

For clarity, **Figure 4.1D** illustrates diploid copy number states that will be referenced in the remainder of the manuscript. Hereafter we refer to diploid copy numbers except when discussing inbred strains (which are effectively haploid).

4.2.2 Copy number polymorphism at *R2d2*

We previously demonstrated that haploid copy number of *R2d* ranges from 1 in the reference strain C57BL/6J and classical inbred strains A/J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HILtJ; to 2 in the wild-derived strain CAST/EiJ; to 34 in the wild-derived strain WSB/EiJ. Using linkage mapping in two multiparental genetic reference populations, the Collaborative Cross[?] and Diversity Outbred[222], we showed that, for the two strains with haploid copy number > 1, one of the copies maps to *R2d1* while all extra copies map to the *R2d2* locus at chr2: 83 Mbp[286]. *Cwc22* was recently reported to have diploid copy number as high as 83 in wild *M. m. domesticus*[268]. In whole-genome sequence data from more than 60 mice from both laboratory stocks and natural populations, we have observed no instances in which the *R2d* copy in *R2d1* is lost. We conclude that diploid copy number > 2 indicates at least one copy of *R2d* is present in *R2d2* (**Figure 4.1D**).

In order to understand the evolutionary dynamics of copy-number variation at *R2d2*, we investigated the relationship between copy number and the local phylogeny in the *R2d2* candidate region. In particular, we sought evidence for or against a single common origin for each of the copy-number states at *R2d2* which are derived with respect to the *M. spretus* – *M. musculus* common ancestor (**Figure 4.1D**). If a derived copy-number state has a single recent origin, it should be associated with a single haplotype at *R2d2*. If a derived copy-number state arises by recurrent mutation, the same copy number should be associated with multiple haplotype backgrounds and possibly in multiple populations.

The extent of *R2d* copy-number variation in *M. musculus*, as estimated on a continuous scale by qPCR, is shown in **Figure 4.5A**. (Note that the qPCR readout is proportional to copy number on the log scale. Extrapolation to integer copy number is increasingly noisy for copy numbers greater than > 6.) We confirmed that *R2d2* maps to chr2: 83 Mbp by performing association mapping between SNP genotypes from the MegaMUGA array[95] and the qPCR readout (**Figure 4.5B**).

We performed a similar analysis to test the hypothesis that *R2d2* alleles with high copy number (diploid copy number > 4, **Figure 4.1D**; hereafter "*R2d2*^{HC}") have a single origin. First we observed that *R2d2*^{HC} alleles are confined with few exceptions to *M. m. domesticus* (**Table 4.2**). The best-associated SNP on the MegaMUGA array (JAX00494952) only weakly tags copy number ($r^2 = 0.137$), but severe ascertainment bias on the MUGA platform[95] makes local LD patterns difficult to interpret. To examine further, we constructed a neighbor-joining phylogenetic tree for the region containing *R2d2* (chr2: 83 – 84 Mb) using genotypes from the $600,00$-SNP Mouse Diversity Array[92,157]. We restricted our attention to inbred strains or wild mice with homozygous, non-recombinant haplotypes in the target region. Twelve samples with *R2d2*^{HC} alleles, both wild mice and laboratory stocks, cluster in a single clade (**Figure 4.5C**). (A single *M. spretus* strain, SPRET/EiJ, also carries an *R2d2*^{HC} allele, but see **§ 4.3**).

Next we expanded the analysis to include an additional 11 samples with *R2d2*^{HC} alleles and evidence of heterozygosity around *R2d2*. The total set of 24 samples includes 7 wild-derived laboratory strains (DDO, RBA/DnJ, RBB/DnJ, RBF/DnJ, WSB/EiJ, ZALENDE/EiJ and SPRET/EiJ), 4 classical inbred strains (ALS/LtJ, ALR/LtJ, CHMU/LeJ and NU/J), a line derived from the ICR:HsD outbred stock (HR8[290]) and 12 wild-caught mice. All 24 samples with *R2d2*^{HC} alleles share an identical haplotype across a single 21 kbp interval, chr2: 83,896,447 – 83,917,565 (GRCm38/mm10 coordinates) (**Figure 4.5D**). These analyses support a single origin for *R2d2*^{HC} alleles within *M. m. domesticus*.

To test the hypothesis that losses of *R2d2* (diploid copy number < 4; at least one chromosome with zero copies in *R2d2*, **Figure 4.1D**) have a single origin, we examined their distribution across the three well-differentiated subspecies of *M. musculus*. Losses of *R2d2* occur in all subspecies of *M. musculus*, in populations that span its geographic range (**Table 4.2**). Based on this distribution and our previous observation that no common haplotype is shared in samples with low copy number in *M. m. domesticus*[286], we reject the hypothesis of single origin and conclude that *R2d2* has been

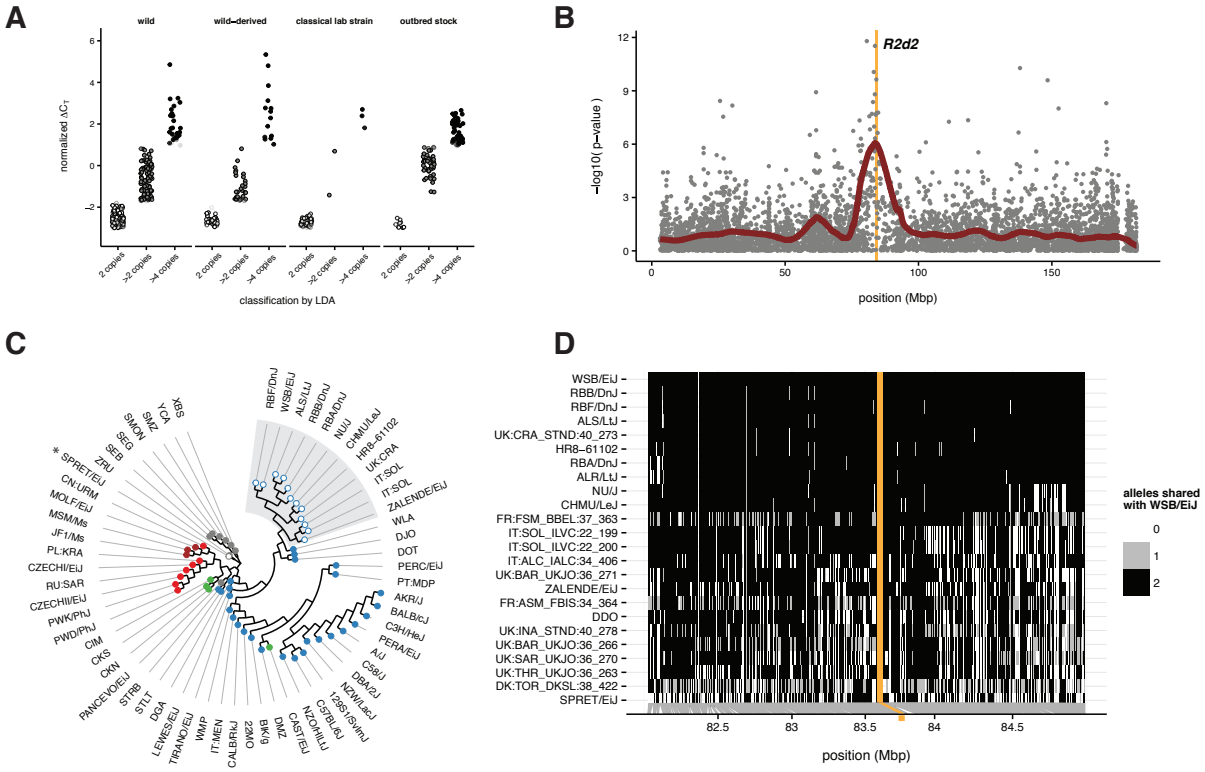

Figure 4.5: Copy-number variation of *R2d* in mouse populations worldwide. **A**) Copy-number variation as measured by quantitative PCR. The normalized ΔC_t value is proportional to $\log$ copy number. Samples are classified as having 2 diploid copies, > 2 copies or > 4 copies of *R2d* using linear discriminant analysis (LDA). (**B**) Fine-mapping the location of *R2d2* in 83 samples genotyped on the Mouse Diversity Array (MDA). Grey points give nominal p-values for association between *R2d* copy number and genotype; red points show a smoothed fit through the underlying points. The candidate interval for *R2d2* from[286], shown as an orange shaded box, coincides with the association peak. (**C**) Local phylogeny at chr2: 83-84 Mbp in 62 wild-caught mice and laboratory strains. Tips are colored by subspecies of origin: *M. m. domesticus*, blue; *M. m. musculus*, red; *M. m. castaneus*, green; *M. m. molossinus*, maroon; outgroup taxa, grey. Individuals with > 4 diploid copies of *R2d* are shown as open circles. (**D**) Haplotypes of laboratory strains and wild mice sharing a high-copy allele at *R2d2*. All samples share a haplotype over the region shaded in orange.

		R2d copy number		
Subspecies	**Region**	2 copies	> 2 copies	> 4 copies
domesticus	Americas	106	23	13
	Asia	7	2	1
	Europe/Med	131	59	54
		244	84	68
musculus	Americas	0	0	0
	Asia	0	1	0
	Europe/Med	3	4	0
		3	5	0
castaneus	Americas	0	0	0
	Asia	3	34	2
	Europe/Med	0	0	0
		3	34	2

Table 4.2: Frequency table of copy-number status by geographic origin for wild-caught and wild-derived *Mus musculus* individuals used in this study, stratified by subspecies. "Europe/Mediterranean" includes continental Europe, the United Kingdom and countries in the Mediterranean basin (Tunisia, Cyprus, Israel). "Asia" includes Asia, the Middle East and countries in the Indian Ocean basin (Madagascar).

Name	Generation	Copy number	Type
UNC_DO_G16_107F	G16	9.9 (1.7)	LOSS
UNC_DO_G13_044F	G13	8.3 (3.2)	LOSS
UNC_DO_G16_015F	G16	14.2 (6.7)	LOSS
UNC_DO_G16_125F	G16	17.4 (3.5)	LOSS
JDO-17	G19	18.8 (4.0)	LOSS
DO-G19-015	G19	21.2 (8.4)	LOSS
UNC_DO_G16_116F	G16	19.2 (2.4)	LOSS
UNC_DO_G16_096F	G16	37.8 (3.5)	GAIN

Table 4.3: Individuals from the DO population carrying *de novo* copy-number mutations at *R2d2*. Copy numbers were estimated by qPCR in progeny; standard errors shown in parentheses. Each was expected to be heterozygous for the WSB/EiJ allele (33 haploid copies).

lost multiple times on independent lineages in each subspecies.

Alternatively, we could posit that the *R2d* duplication never fixed in the ancestor of *M. musculus* and that both duplicated and un-duplicated alleles have bene maintained for 2 My as balanced polymorphisms in the major lineages in the Palearctic clade of *Mus*. We find this a less-likely scenario given current estimates of effective population size (N_e) in house mice ($50,000 - 250,000$ [125,291,292]) and the expected fixation time of a neutral allele ($\approx 4N_e$ [80]).

4.2.3 Sequence and structural diversity near *R2d2*

The extent of copy-number polymorphism involving *R2d2* suggests that it is prone to recurrent mutation. Consistent with these observations, we find that the rate of *de novo* copy-number changes at *R2d2* is extremely high in laboratory populations (**Figure 4.6**). In 183 mice sampled from the DO population we identified and confirmed through segregation analysis 8 new alleles, each with distinct copy number and each occurring in an unrelated haplotype (**Table 4.3**). Without complete pedigrees and genetic material from breeders a direct estimate of the mutation rate in the DO is not straightforward to obtain. However, since the population size is known, we can make an analogy to microsatellite loci [293] and estimate the mutation rate via the variance in allele sizes: 3.2 mutations per 100 transmissions (3.2%; 95% bootstrap CI $1.1\% - 6.0\%$).

Structural instability in this region of chromosome 2 extends outside the *R2d2* locus itself. Less than 200 kbp distal to *R2d2* is another duplication (**Figure 4.7B** grey shaded region) — containing a

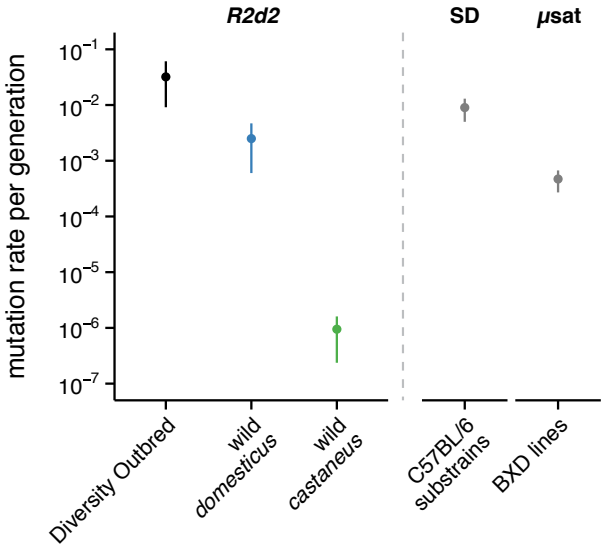

Figure 4.6: Estimates of per-generation mutation rate for CNVs at *R2d2* (± 1 bootstrap SE) in the DO population; among wild *M. m. domesticus;* and among wild *M. m. castaneus*. For comparison, mutation rates are shown for the CNV with the highest rate of recurrence in a C57BL/6J pedigree[294] and for a microsatellite whose mutation rate was estimated in the BXD panel[31].

retrotransposed copy of *Cwc22* — that is present in 7 tandem copies in the reference genome. That region, plus a further 80 kbp immediately distal to it, is copy-number polymorphic in wild *M. m. domesticus* and wild *M. m. castaneus* (**Figure 4.7B**). Instability of the region over a longer time-scale is demonstrated by the disruption, just distal to the aforementioned segmental duplication, of a syntenic block conserved across all other mammals (**Figure 4.3**).

Despite the high mutation rate for structural variants involving *R2d2* and nearby sequences, sequence diversity at the nucleotide level is modestly reduced relative to diversity in *R2d1* and relative to the genome-wide average in *M. m. domesticus*. In a 200 kbp region containing the *R2d2* insertion site at its proximal end, $\hat{\pi}$ (an estimator of average heterozygosity) in *M. m. domesticus* reduced by at least a factor of two from the local average of approximately 0.3% (which is comparable to previous reports in this subspecies[291]) (**Figure 4.7B**). Divergence between *M. musculus* and *M. caroli* is similar to its genome-wide average of $\sim 2.5\%$ over the same region.

Estimation of diversity *within* a duplicated sequence such as *R2d* is complicated by the difficulty of distinguishing allelic from paralogous variation. To circumvent this problem we split our sample of 26 wild *M. m. domesticus* into two groups: those having *R2d1* sequences only, and those having both *R2d1* and *R2d2* sequences. Within each group we counted the number of segregating sites among all *R2d2* copies, using nearby fixed differences between *R2d1* and *R2d2* to phase sites to *R2d2* (see **§ 4.5** for details), and used Watterson's estimator to calculate nucleotide diversity per site. Among *R2d1* sequences, $\hat{\theta} = 0.09\% \pm 0.03\%$ versus $\hat{\theta} = 0.04\% \pm 0.02\%$ among *R2d2* sequences (**Figure 4.7C**) and $\hat{\theta} = 0.13\% \pm 0.04\%$ among *R2d2* sequences in *M. m. castaneus*.

4.2.4 *R2d* contains the essential gene *Cwc22*

The *R2d* unit encompasses one protein-coding gene, *Cwc22*, which encodes an essential messenger RNA (mRNA) splicing factor[295]. The gene is conserved across eukaryotes and is present in a single copy in most non-rodent species represented in the TreeFam database (`http://www.treefam.org/family/TF300510`[296]). Five groups of *Cwc22* paralogs are present in mouse genomes: the copies in *R2d1* (*Cwc22*^{R2d1}) and *R2d2* (*Cwc22*^{R2d2}) plus retrotransposed copies in one locus at chr2: 83.9 Mbp and at two loci on the X chromosome (**Figure 4.8A**).

The three retrotransposed copies are located in regions with no sequence similarity to each other, indicating that each represents an independent retrotransposition event. The copy on chr2 was subsequently expanded by further duplication and now exists (in the reference genome) in 7

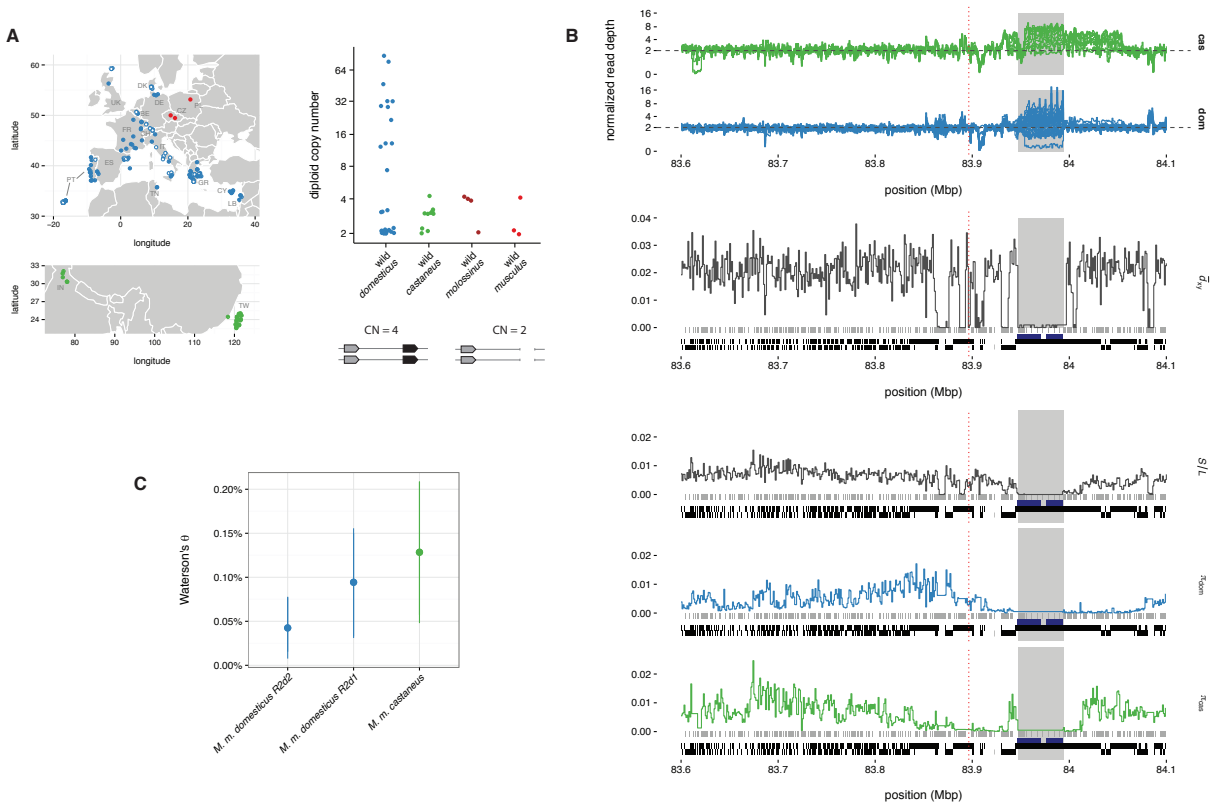

Figure 4.7: (**A**) Geographic origin of wild mice used in this study, color-coded by subspecies (blue, *M. m. domesticus*; red, *M. m. musculus*; green, *M. m. castaneus*). Diploid copy number of the *R2d* unit is shown for wild samples for which integer copy-number estimates are available: 26 *M. m. domesticus* and 10 *M. m. castaneus* with whole-genome sequencing data, and representatives from *M. m. molossinus* and *M. m. musculus* for comparison. Schematic shows the *R2d1*/*R2d2* configurations corresponding to diploid copy numbers of 2 and 4. (**B**) Profiles of read depth (first two panels), average sequence divergence to outgroup species *M. caroli* (d_{xy}, third panel), number of segregating sites per base (S/L, fourth panel) and within-population average heterozygosity (π, fifth and sixth panels). The region shown is 500 kbp in size; the insertion site of *R2d2* is indicated by the red dashed line. Grey boxes along baseline show positions of repetitive elements (from UCSC RepeatMasker track); black boxes show non-recombining haplotype blocks. Blue bars indicate the position of 7 tandem duplications in the mm10 reference sequence with $> 99\%$ mutual identity, each containing a copy of retro-*Cwc22*. Grey shaded region indicates duplicate sequence absent from *M. caroli*. (**C**) Estimated per-site nucleotide diversity within *M. m. domesticus R2d1*, *M. m. domesticus R2d2* and *M. m. castaneus R2d2*.

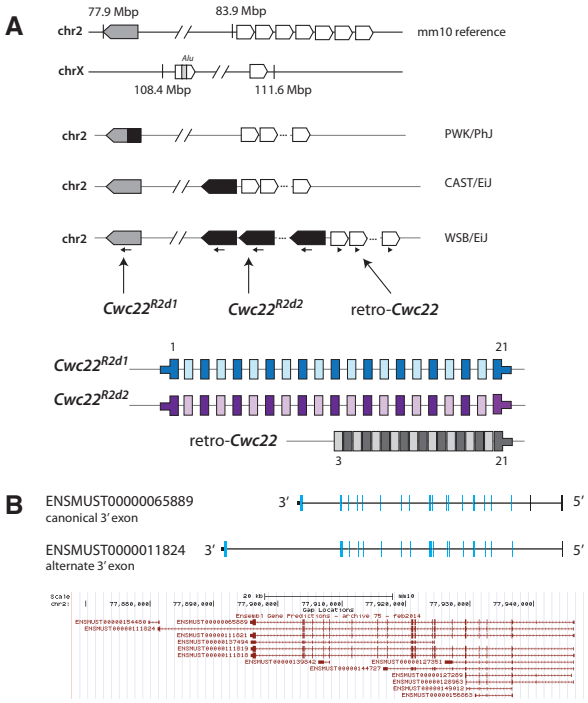

Figure 4.8: *Cwc22* paralogs in the mouse genome. (**A**) Location and organization of *Cwc22* gene copies present in mouse genomes. The intact coding sequence of *Cwc22* exists in in both *R2d1* (grey shapes) and *R2d2* (black shapes). Retrotransposed copies (empty shapes) exist in two loci on chrX and one locus on chr2, immediately adjacent *R2d2*. Among the retrotransposed copies, coding sequence is intact only in the copy on chr2. Exon numbers are shown in grey above transcript models. (**B**) Alternate transcript forms of *Cwc22*, using different 3′ exons. Coding exons shown in blue and untranslated regions in black. All Ensembl annotated transcripts are shown in the lower panel (from UCSC Genome Browser.)

copies with $> 99.9\%$ mutual similarity. The two retrotransposed copies on chrX are substantially diverged from the parent gene ($< 90\%$ sequence similarity), lack intact open reading frames (ORFs), have minimal evidence of expression among GenBank cDNAs, and are annotated as likely pseudogenes[297]. We therefore restricted our analyses to the remaining three groups of *Cwc22* sequences, all on chr2.

The canonical transcript of *Cwc22*^{R2d1} (Ensembl transcript ID ENSMUST00000065889[115]) is encoded by 21 exons on the negative strand. The coding sequence begins in the third exon and ends in the terminal exon (**Figure 4.8B**). Six of the seven protein-coding *Cwc22*^{R2d1} transcripts in Ensembl v83[115] use this terminal exon, while one transcript (ENSMUST0000011824) uses an alternative terminal exon. Alignment of the retrogene sequence (ENSMUST00000178960) to the reference genome demonstrates that the retrogene captures the last 19 exons of the canonical transcript — that is, the 19 exons corresponding to the full coding sequence of the parent gene.

4.2.5 Expression patterns of *Cwc22* paralogs

To identify the coding sequence of *Cwc22*^{R2d2} we first aligned the annotated transcript sequences of *Cwc22*^{R2d1} from Ensembl to our *R2d2* contig. All 21 exons present in *R2d1* are present in *R2d2*. We created a multiple sequence alignment and phylogenetic tree of *Cwc22* cDNAs and predicted amino acid sequences from *Cwc22*^{R2d1}, *Cwc22*^{R2d2}, retro-*Cwc22*, and CWC22 orthologs in 19 other placental mammals, plus opossum, platypus and finally chicken as an outgroup (**Figure 4.9**). An open reading frame (ORF) is maintained in all three *Cwc22* loci in mouse, including the retrogene. Information content of each column along the alignment (**Figure 4.10**) reveals that sequence is most conserved in two predicted conserved domains, MIF4G and MA3, required for *Cwc22*'s function in mRNA processing[295].

Next we examined public RNA-seq data from adult brain and testis in inbred strains with one or more copies of *R2d2* for evidence of transcription of each *Cwc22* family member. We identified several novel transcript isoforms specific to *R2d2* arising from two intron-retention events and one novel 3′ exon (**Figure 4.11A**). The 18$^{\text{th}}$ intron is frequently retained in *Cwc22*^{R2d2} transcripts, most likely due to an `A` $\rightarrow$ `G` mutation at the 5′ splice donor site of exon 17 in *Cwc22*^{R2d2}. The 12$^{\text{th}}$ intron is also frequently retained. While we could not identify any splice-region variants near this intron, it contains an ERV insertion that may interfere with splicing (**Figure 4.11A**). Both intron-retention events would create an early stop codon. Finally, we find evidence for a novel 3′ exon that extends

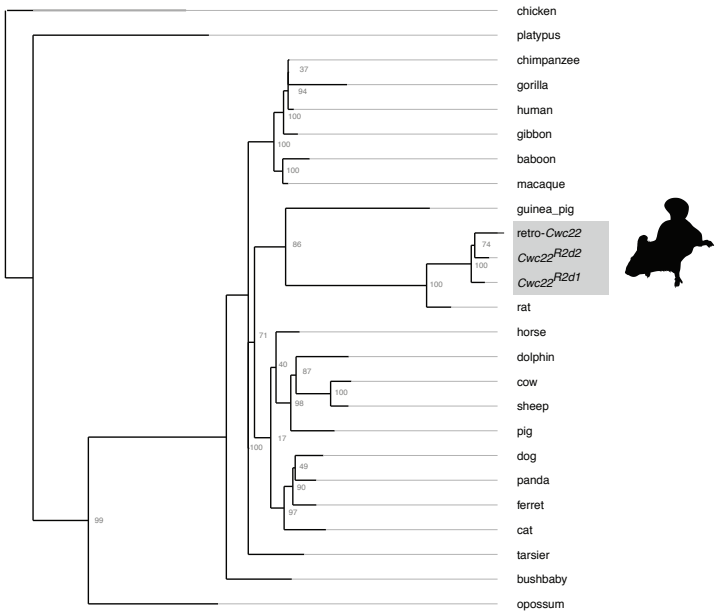

Figure 4.9: Phylogenetic tree constructed from amino acid sequences for mammalian CWC22 homologs (including all three mouse paralogs) with chicken as an outgroup. Node labels indicate support in 100 bootstrap replicates.

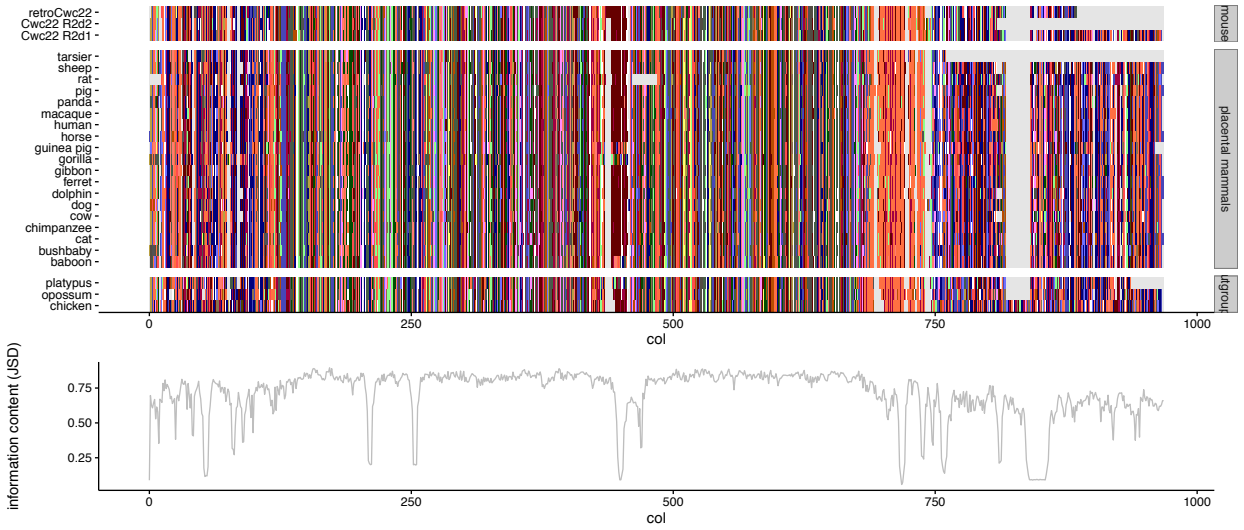

Figure 4.10: Alignment of amino acid sequences corresponding to mouse *Cwc22*^{R2d1}, *Cwc22*^{R2d2} and retro-*Cwc22*, plus CWC22 orthologs from 19 other placental mammals plus opossum, platypus and chicken as outgroups. Residues are colored according to biochemical properties and gaps are shown in grey. Information content of each column in the alignment, measured as the Jenson-Shannon divergence, is plotted in the lower panel.

to the boundary of the *R2d* unit and is used exclusively by *Cwc22*^{R2d2} (**Figure 4.11A**).

We estimated the expression of the various isoforms of *Cwc22*^{R2d1}, *Cwc22*^{R2d2} and retro-*Cwc22* in adult brain and testis. For brain we obtained reads from 8 replicates (representing both sexes) on 3 inbred strains[298], and for testis a single replicate on 23 inbred strains[299] and estimated transcript abundance using the `kallisto` package[300]. Briefly, `kallisto` uses an expectation-maximization (EM) algorithm to accurately estimate the abundance of a set of transcripts by distributing the "weight" of each read across all isoforms with whose sequence it is compatible. *Cwc22* is clearly expressed from all three paralogs in both brain and testis (**Figure 4.11B**). However, both the total expression and the pattern of isoform usage differ by tissue and copy number.

Maintenance of an ORF in all *Cwc22* paralogs for > 2 My is strong evidence of negative selection against disrupting mutations in the coding sequence, but long branches within the rodent clade in **Figure 4.9** suggest that *Cwc22* may also be under relaxed purifying selection or positive selection in rodents. The rate of evolution of *Cwc22* sequences in mouse is faster than in the rest of the tree ($\chi^2 = 4.33$, df = 1, $p = 0.037$ by likelihood ratio test).

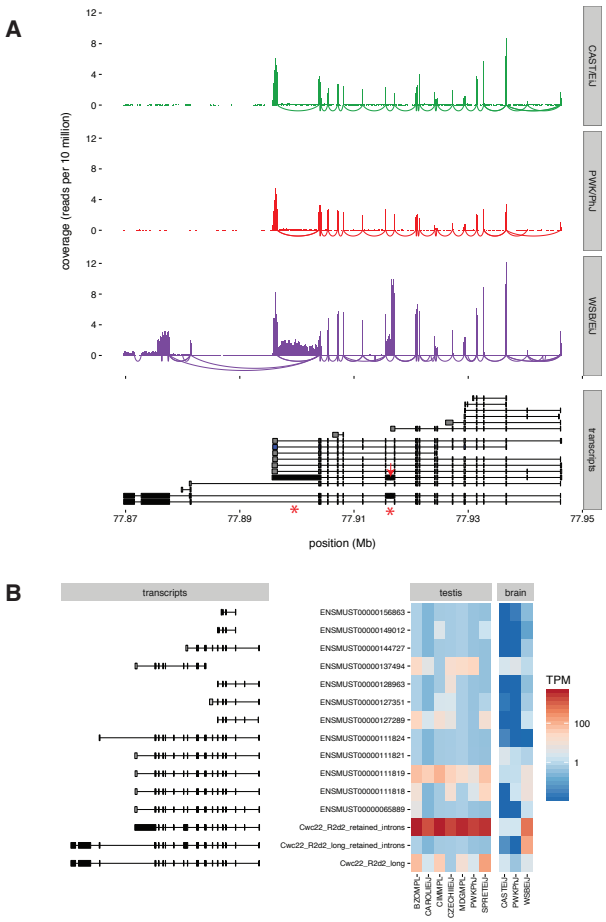

Figure 4.11: Expression of *Cwc22* isoforms. (**A**) Read coverage and splicing patterns in *Cwc22* in adult mouse brain from three wild-derived inbred strains. Swoops below x-axis indicate splicing events supported by 5 or more split-read alignments. Known transcripts of $Cwc22^{R2d1}$ (grey, from Ensembl), inferred transcripts from $Cwc22^{R2d2}$ (black) and the sequence of retro-*Cwc22* mapped back to the parent gene (blue) are shown in the lower panel. Red stars indicate retained introns; red arrow indicates insertion site of an ERV in *R2d2*. (**B**) Estimated relative expression of *Cwc22* isoforms (y-axis) in adult mouse brain and testis in wild-derived inbred strains (x-axis). TPM, transcripts per million, on log10 scale.

4.2.6 Non-allelic gene conversion between *R2d1* and *R2d2*

The topology of trees across *R2d* is generally consistent: a long branch separating the single *M. caroli* sequence from the *M. musculus* sequences, and two clades corresponding to *R2d1*- and *R2d2*-like sequences. However, we observed that the affinities of some *R2d* paralogs change along the sequence (**Figure 4.12A**), a signature of non-allelic (*i.e.* inter-locus) gene conversion. In this context, we use "gene conversion" to describe a non-reciprocal "copy-and-paste" transfer of sequence from one donor locus into a different, homologous receptor locus, without reference to a specific molecular mechanism[301].

To investigate further, we inspected patterns of sequence variation in whole-genome sequencing data from 15 wild-caught mice, 2 wild-derived inbred strains, and 22 classical inbred strains of mice with diploid *R2d* copy number 2. We first defined $1,411$ pairwise single-nucleotide differences (1 per 89 bp; Ti:Tv = 1.85) between *R2d2* and *R2d1* for which *R2d2* has the derived allele with respect to *M. caroli*. Then we tested for the presence of the derived allele, ancestral allele or both at each site in each sample. Finally we identified conversion tracts by manual inspection as clusters of derived variants shared with *R2d2*.

This analysis revealed non-allelic gene conversion tracts on at least 9 chromosomes out of the small sample of 54 chromosomes examined (**Figure 4.12B**). The conversion tracts range in size from approximately 1.2 kbp to 119 kbp. The boundaries of several tracts are shared within populations, suggesting that the tracts are shared by descent. We excluded the possibility of complementary losses from *R2d1* and *R2d2* — which would leave similar patterns of sequence variation — by finding read pairs spanning the boundary between *R2d1* and flanking sequence, and between *R2d1*-like and *R2d2*-like tracts on the same chromosome (examples shown in **Figure 4.13**).

The conversion tracts we detected are orders of magnitude longer than the 15 to 750 bp reported in recent studies of allelic gene conversion at recombination hotspots in mouse meiosis[210,302]. We require the presence of *R2d2*-diagnostic alleles at two or more consecutive variants to declare a conversion event, and these variants occur at a rate of approximately 1 per 100 bp, so the smallest conversion tracts we could theoretically detect are on the order of 200 bp in size. Even if we require only a single variant to define a conversion tract, all samples without a long conversion tract share fewer than 55 and most fewer than 10 derived alleles (of $1,411$ total sites) with *R2d2*, of which all are also shared by multiple other samples from different populations. This pattern indicates that

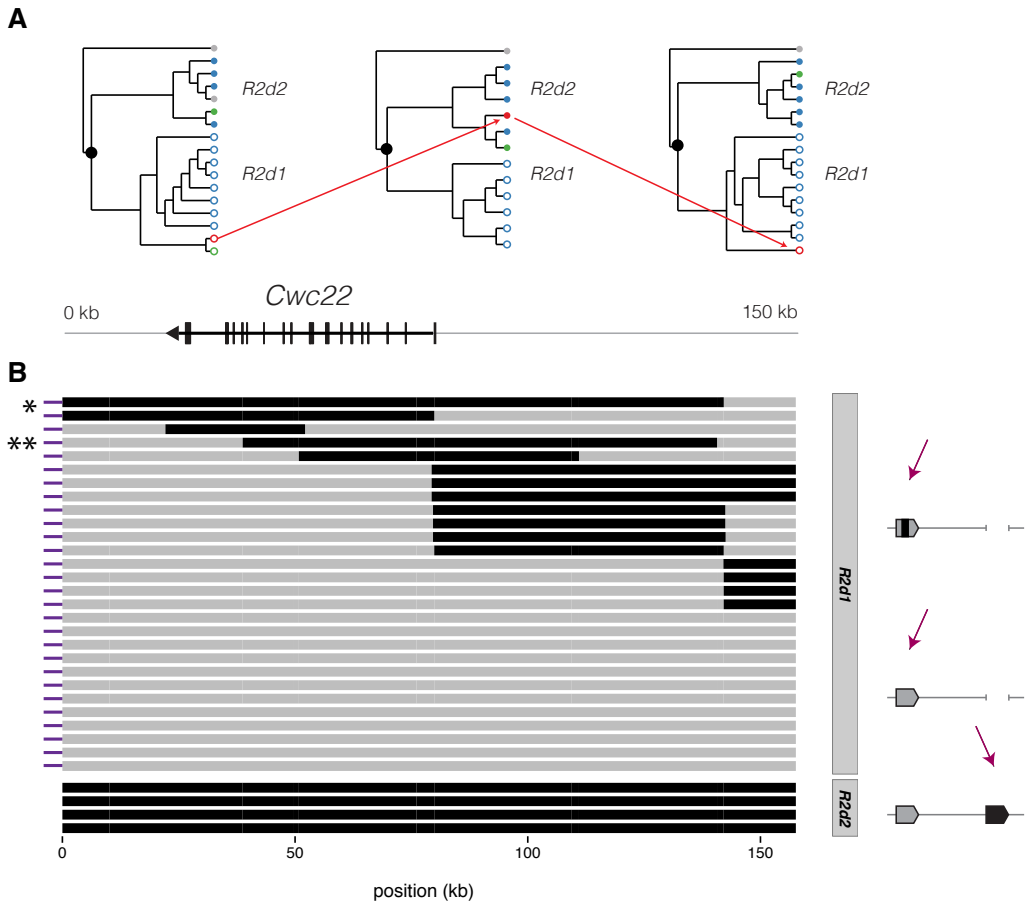

Figure 4.12: Signatures of non-allelic gene conversion between *R2d1* and *R2d2*. (**A**) Phylogenetic trees for three representative intervals across *R2d*. Sequences are labeled according to their subspecies of origin using the same color scheme as in **Figure ??**; open circles are *R2d1*-like sequences and closed circles are *R2d2*-like. Trees are drawn so that *M. caroli*, the outgroup species used to root the trees, is always positioned at the top. The changing affinities of PWK/PhJ (red) and CAST/EiJ (green) along *R2d* are evidence of non-allelic gene conversion. (**B**) *R2d* sequences from 20 wild-caught mice and 5 laboratory inbred strains. Each track represents a single chromosome; grey regions are classified as *R2d1*-like based on manual inspection of sequence variants, and black-regions *R2d2*-like. Upper panel shows sequences from samples with a single copy of *R2d*, residing in *R2d1*. Lower panel shows representative *R2d2* sequences for comparison. Asterisks indicate samples for which read alignments are shown in **Figure 4.13**.

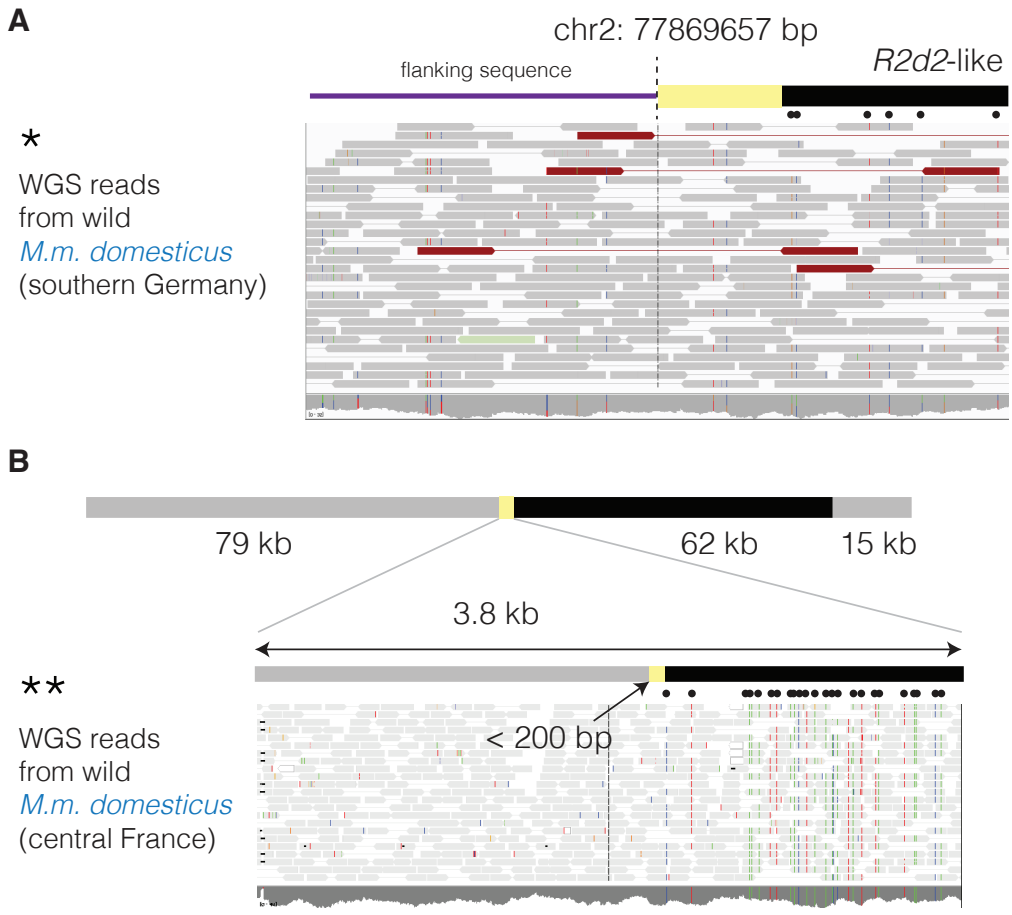

Figure 4.13: Physical linkage at boundaries of non-allelic gene conversion tracts. (**A**) Paired-end read alignments (visualized with IGV) across the proximal boundary (dashed line) of *R2d1* in a sample with a conversion tract extending to the boundary. Positions of derived variants shared with *R2d2* are indicated by black dots. (**B**) Read alignments across the boundary of a non-allelic gene conversion tract. *R2d1* sequence from a single chromosome is a mosaic of *R2d1*-like (grey) and *R2d2*-like (black) segments. A magnified view of read pairs in the 3.8 kbp surrounding the proximal boundary of the tract shows read pairs spanning the junction. Black dots indicate the position of derived alleles diagnostic for *R2d2*. The precise breakpoint lies somewhere in the yellow shaded region between the last *R2d1*-specific variant and the first *R2d2*-specific variant.

those sites in fact represent either artifacts (from mis-assignment of ancestral and derived alleles) or recurrent mutations rather than short gene conversions.

Four conversion tracts partially overlap the *Cwc22* gene to create a sequence that is a mosaic of *R2d1*- and *R2d2*-like exons (**Figure 4.12B**). Recovery of *Cwc22* mRNA in an inbred strain with a mosaic sequence (PWK/PhJ, see **§ 4.2.5**) indicates that its exons are intact, adjacent and properly oriented in *cis* to permit transcription. The presence of both *R2d1*- and *R2d2*-like sequence in extant *M. musculus* lineages with 2 diploid copies of *R2d* further reinforces our conclusion that the duplication is indeed ancestral to the divergence of *M. musculus*.

In addition to exchanges between *R2d1* and *R2d2*, we identified an instance of exchange between *R2d2* and the nearby retrotransposed copy of *Cwc22* in a single *M. m. domesticus* individual from Iran (IR:AHZ_STND:015; **Figure 4.14**). This individual carries a rearrangement that has inserted a 30 kbp fragment corresponding to the 3' half of $Cwc22^{R2d2}$ into the retro-*Cwc22* locus, apparently mediated by ~ 100 bp of homology between the exons of $Cwc22^{R2d2}$ and retro-*Cwc22*.

4.2.7 High copy number at *R2d2* suppresses meiotic recombination

The difficulty of fine-mapping *R2d2* in standard crosses[286] suggested that recombination is suppressed around *R2d2*. Based on our observation that recombination is suppressed around large structural variants (see Chapter 3), we tested whether the region around *R2d2* has lower recombination when an $R2d2^{HC}$ allele is present. Understanding patterns of recombination at *R2d2* is important for interpreting levels of sequence and haplotype diversity in the surrounding region.

First we analyzed local recombination rate in the DO population. **Figure 4.15A** shows the cumulative distribution of $2,917$ recombination events on central chromosome 2, stratified according to *R2d2* copy number of the participating haplotypes. The recombination map has a pronounced plateau in the region between *R2d1* and approximately 1 Mb distal to *R2d2* (dashed lines) for $R2d2^{HC}$ haplotypes, but not $R2d2^{LC}$ haplotypes. As a result, $R2d2^{HC}$ haplotype blocks overlapping *R2d2* are significantly longer than $R2d2^{LC}$ haplotype blocks ($p < 0.01$ by Wilcoxon rank-sum tests with Bonferroni correction) in 8 of the 10 generations sampled (**Figure 4.15B**). The difference arose early in the breeding of the DO and persists through the most recent generation for which the randomized breeding scheme was maintained[270].

Second we re-examined genotype data from 11 published crosses in which at least one parent was segregating for an $R2d2^{HC}$ allele. Whereas in the DO we used haplotype block length as

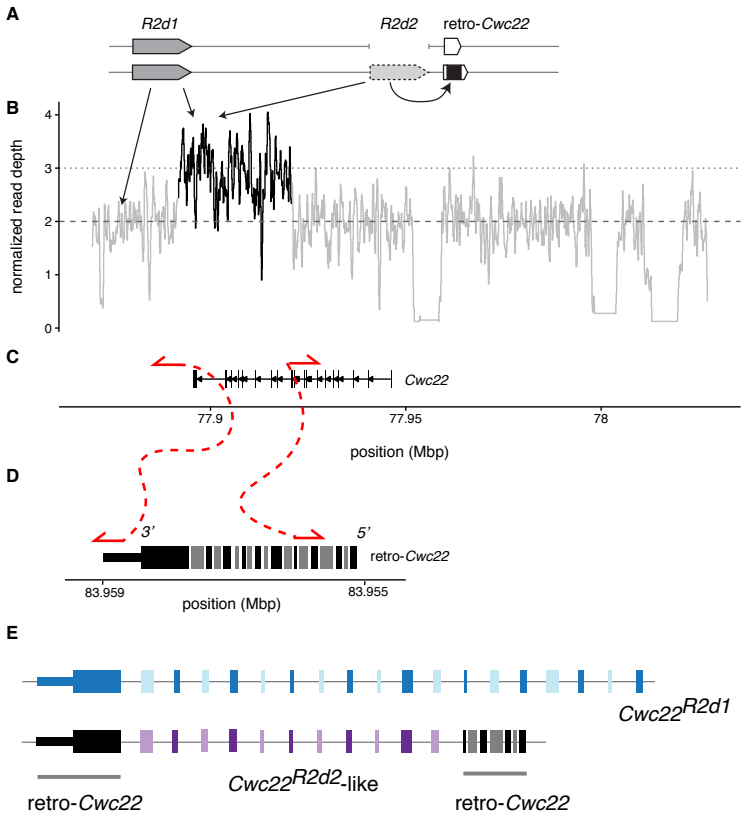

Figure 4.14: Partial loss of *R2d2* with structural rearrangement. (**A**) Inferred structure of the *R2d1-R2d2* region in IR:AHZ_STND:015, a wild *M. m. domesticus* individual from Iran. *R2d1* is present on both chromosomes but only a fragment of *R2d2* remains on one chromosome, and it has been transposed into the retro-*Cwc22* array. (**B**) Normalized depth of coverage (2 = normal diploid level) across *R2d*. Regions in grey represent reads from *R2d1* alone, while region in black captures reads from *R2d1* and *R2d2*, as shown by arrows from panel A. (**C**) Position of read pairs (red; not drawn to scale) with soft-clipped alignments to *R2d1*. The proximal read aligns in the 3′ UTR of *Cwc22*, and the distal read across an exon-intron boundary within the gene body. Note the "outward"-facing direction of the alignments. (**D**) Positions of the mates of the reads in panel C. Note that the x-axis is reversed so that the exons of retro-*Cwc22* (encoded on the plus strand) parallel those of *Cwc22* (encoded on the minus strand). The 3′ read maps across the boundary of th 3′ UTR of *Cwc22* and the ERV mediating the retrotransposition event. The 5′ read maps across two exon-exon boundaries in retro-*Cwc22*, so there is no ambiguity regarding its alignment to the retro-transposed copy. (**E**) Inferred structure of *Cwc22* paralogs in this sample. Note that one of the copies of retro-*Cwc22* is now a mosaic of retrotransposed and $Cwc22^{R2d2}$-derived sequence.

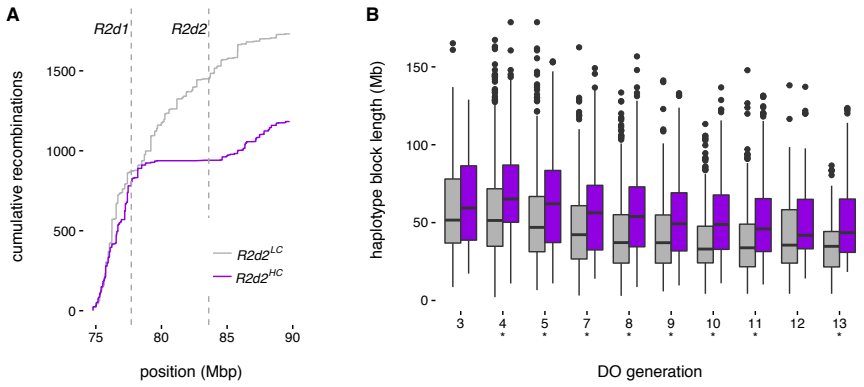

Figure 4.15: Suppression of crossing-over around *R2d2* in the DO. (**A**) Cumulative number of unique recombination events in the middle region of chr2 in genomes of 4,640 Diversity Outbred mice. Recombination events involving the high-copy-number WSB/EiJ haplotype are shown in purple and all other events in grey. Dashed vertical lines indicate the position of *R2d1* (left) and *R2d2* (right). (**B**) Distribution of haplotype block sizes at *R2d2* in selected generations of the DO, for $R2d2^{HC}$ (WSB/EiJ, purple) versus $R2d2^{LC}$ (the other seven founder haplotypes, grey). Asterisks indicate generations in which the length distributions are significantly different by Wilcoxon rank-sum test.

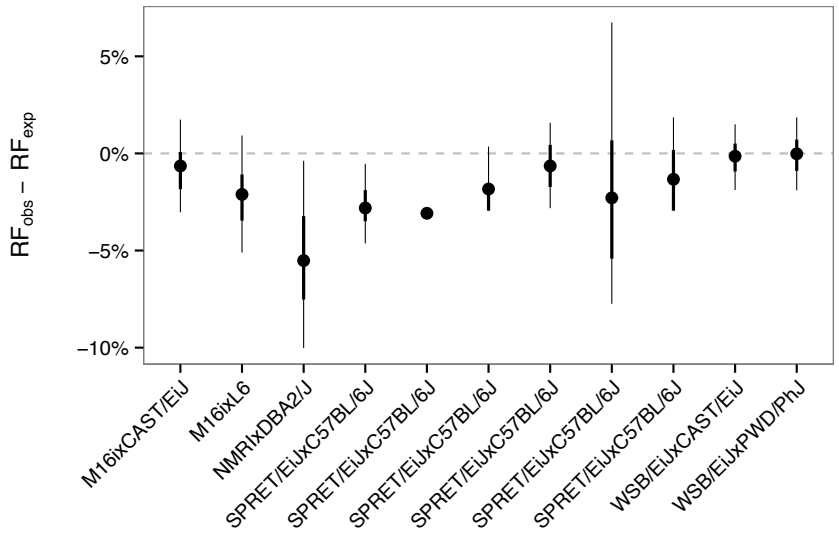

Figure 4.16: Difference between expected and observed recombination fraction between markers flanking *R2d2* in experimental crosses in which at least one parent is segregating for a high-copy allele of *R2d2*. Thick and thin vertical bars show 90% and 95% confidence bounds, respectively, obtained by non-parametric bootstrap.

a proxy for recombination rate, in these F_2 and backcross designs we can directly estimate the recombination fraction across *R2d2* and compare it to its expected value in the absence of an $R2d2^{HC}$ allele (**Figure 4.16**). In 9 of 11 crosses examined, the observed recombination fraction is lower than the expected ($p < 0.032$, one-sided binomial test).

4.3 Discussion

In this manuscript we have reconstructed in detail the evolution of a multi-megabase segmental duplication (SD) in mouse, *R2d2*. Our findings illustrate the challenges involved in accurately interpreting patterns of polymorphism and divergence within duplicated sequence.

SDs are among the most dynamic loci in mammalian genomes. They are foci for copy-number variation in populations, but the sequences of individual duplicates beyond those present in the reference genome are often poorly resolved. Obtaining the sequence of this "missing genome," as we have done for *R2d2*, is an important prerequisite to understanding the evolution of duplicated loci. Since each paralog follows a partially independent evolutionary trajectory, individuals in a population may vary both quantitatively (in the number of copies) and qualitatively (in which

copies are retained). Cycles of duplication and loss may furthermore lead to the fixation of different paralogs along different lineages. This "genomic revolving door"[303] leaves a signature of polymorphism far in excess of the genome-wide background, due to coalescence between alleles originating from distinct paralogs.

Accurate deconvolution of recent duplications remains a difficult task that requires painstaking manual effort. Clone-based and/or single-molecule long-read sequencing remain the gold standard techniques. But short reads at sufficient depth nonetheless contain a great deal of information. We exploited the specific properties of *R2d2* in the WSB/EiJ mouse strain — many highly-similar copies of *R2d2* relative to *R2d1*, with informative paralogous variants every ~ 100 bp — to obtain a nearly complete assembly of *R2d2* from short reads (**Figure 4.3**). With the sequence of both the *R2d1* and *R2d2* paralogs in hand, we were able to recognize several remarkable features of *R2d2* that are discussed in detail below.

4.3.1 Long-tract gene conversion

Previous studies of non-allelic gene conversion in mouse and human have focused either on relatively small (< 5 kbp) recombination hotspots within species, or have applied phylogenetic methods to multiple paralogs from a single reference genome[304]. This study is the first, to our knowledge, with the power to resolve large (> 5 kbp) non-allelic gene conversion events on an autosome in a population sample. We identify conversion tracts up to 119 kbp in length, orders of magnitude longer than tracts arising from allelic conversion events during meiosis. Gene conversion at this scale can rapidly and dramatically alter paralogous sequences, including — as shown in **Figure 4.12** — the sequences of essential protein-coding genes. This process has been implicated as a source of disease alleles in humans[301].

Importantly, we were able to identify non-allelic exchanges in *R2d1* as such only because we were aware of the existence of *R2d2* in other lineages. In this case the transfer of paralogous *R2d2* sequence into *R2d1* creates the appearance of deep coalescence among *R2d1* sequences. Ignoring the effect of gene conversion would cause us to overestimate the degree of polymorphism at *R2d1* by an order of magnitude, and would bias any related estimates of population-genetic parameters (for instance, of effective population size).

Our data are not sufficient to estimate the rate of non-allelic gene conversion between *R2d2* and homologous loci. At minimum we have observed two distinct events: one from *R2d2* into *R2d1*, and

a second from *R2d2* into retro-*Cwc22*. From a single conversion event replacing most of *R2d1* with *R2d2*-like sequence, the remaining shorter conversion tracts could be generated by recombination with *R2d1* sequences. Because we find converted haplotypes in both *M. m. musculus* and *M. m. domesticus*, the single conversion event would have had to occur prior to the divergence of the three *M. musculus* subspecies and subsequently remain polymorphic in the diverged populations. We note that all conversion tracts we observed are polarized: *R2d2* is always the donor.

The other possibility is that non-allelic gene conversion between *R2d* sequences is recurrent. Recurrent gene conversion homogenizes duplicate sequences, coupling their evolutionary trajectories (so-called *concerted evolution*[284]). The absolute sequence divergence ($\sim 2\%$) between *R2d1* and *R2d2* (**Figure 4.1B**) argues against the hypothesis that gene conversion has occurred at a uniformly high rate throughout their history. However, we cannot rule out a role for gene conversion in maintaining sequence identity between multiple copies of *R2d* located in *R2d2*. This would help explain the reduced diversity within *R2d2* versus *R2d1* (**Figure 4.7C**). There is some direct evidence that the rate of gene conversion is positively correlated with copy number and negatively correlated with physical distance between duplicates[305], so we might expect it to be highest for $R2d2^{HC}$ alleles.

In this respect *R2d2* may be similar to the male-specific region of the Y chromosome in mouse[306] and human[307]. The large palindromic repeats on chrY are homogenized by frequent non-allelic gene conversion[307,266] such that they have retained $> 99\%$ sequence identity to each other even after millions of years of evolution. Frequent non-allelic gene conversion has also been documented in arrays of U2 snRNA genes in human[308], and in rRNA gene clusters[309] and centromeric sequences[310,311] in several species.

4.3.2 Pervasive copy-number variation

Clusters of segmental duplications have long been known to be hotspots of copy-number variation in populations[239,43] and *de novo* mutations in pedigrees[294,47,34]. Recent large-scale sequencing efforts have revealed the existence of thousands of multiallelic CNVs segregating in human populations[39].

We have surveyed *R2d2* copy number in a large and diverse sample of laboratory and wild mice, and have shown that it varies from 0 to > 80 in certain *M. m. domesticus* populations (**Figure 4.7A**). In a cohort of outbred mice expected to be hemizygous for an $R2d2^{HC}$ allele from WSB/EiJ (33 diploid copies) we estimate that large deletions, > 2 Mbp in size, occur at a rate of 3.2% (95%

bootstrap CI $1.1\% - 6.0\%$) per generation. This estimate of the mutation rate for CNVs at *R2d2* should be regarded as a lower bound. The power of our copy-number assay to discriminate between copy numbers above ~ 25 is low, so that the assay is much more sensitive to losses than to gains. Even our lower-bound mutation rate exceeds that of the most common recurrent deletions in human (~ 1 per 7000 live births)[47] and is an order of magnitude higher than the most active CNV hotspots described to date in the mouse[294].

However, the structural mutation rate appears to depend strongly on the diplotype configuration at *R2d2*. As **Figure 4.1D** shows, individuals heterozygous for an *R2d2*^{HC} haplotype and an *R2d2*-null haplotype are in fact hemizygous for several megabases of DNA in *R2d2*. This has important consequences. High mutation rates are observed only in the context of populations in which hemizygosity for *R2d2*^{HC} is common (**Figure 4.6**): highest in the DO, and to a lesser extent in wild *M. m. domesticus* populations harboring both *R2d2*^{HC} and *R2d2*-null alleles. Homozygosity for *R2d2*^{HC} is not associated with mutability: in 8 recombinant inbred lines from the Collaborative Cross which are homozygous for an *R2d2*^{HC} haplotype, we observed zero new mutations in at least 400 meioses, through both the male and female germline (8 lines $\times 2$ meioses/generation $\times 25$ or more generations of inbreeding). Sex also appears to have a role in determining the mutation rate at *R2d2*: in a pedigree in which all females were hemizygous for *R2d2*^{HC}, zero new mutations were observed in 1256 meioses (data not shown).

Taken together, these observations hint at a common structural or epigenetic mechanism affecting the resolution of double-strand breaks in large tracts of unpaired (*i.e.* hemizygous) DNA during male meiosis. At least one other study in mouse has hinted that hemizygous SDs on the sex chromosomes are unstable in inter-subspecific hybrids[312]. Both the obligate-hemizygous sex chromosomes and large unpaired segments on autosomes are epigenetically marked for transcriptional silencing during male meiotic prophase[313,314], and are physically sequestered into the sex body[144]. Repair of double-strand breaks within the sex body is delayed relative to the autosomes[178] and involves a different suite of proteins[315]. We hypothesize that these male-specific pathway(s) are generally error-prone in the presence of non-allelic homologous sequences.

However, we cannot exclude the possibility that large-scale rearrangement (such as an inversion) associated with copy-number expansion at *R2d2* contributes to its instability. Physical mapping of

the *R2d2* locus in WSB/EiJ is in progress [2] and will shed light on this question.

4.3.3 Origin and distribution of an allele subject to meiotic drive

Females heterozygous for a high- and low-copy allele at *R2d2* preferentially transmit the high-copy allele to progeny via meiotic drive[286]. Meiotic drive can rapidly alter allele frequencies in laboratory and natural populations[316], and we show in **??** that$R2d2^{HC}$ alleles sweep through laboratory and natural populations despite reducing the fitness of heterozygous females. These "selfish sweeps" account, at least in part, for the marked reduction in within-population diversity in the vicinity of *R2d2* (**Figure 4.7B**).

The present study sheds additional light on the age, origins and fate of $R2d2^{HC}$ alleles. We find that $R2d2^{HC}$ alleles have a single origin in *M. m. domesticus*. They are present in several different "chromosomal races" — populations fixed for specific Robertsonian translocations between which gene flow is limited[317] — indicating that they were likely present at intermediate frequency prior to the origin of the chromosomal races within the past $6,000$ to $10,000$ years[318] and were dispersed through Europe as mice colonized the continent from the south and east[123]. The presence of $R2d2^{HC}$ in a non-*M. m. domesticus* sample (SPRET/EiJ, *M. spretus* from Cadiz, Spain) is best explained by recent introgression following secondary contact with *M. m. domesticus*[136,157].

4.3.4 Additional members of the CWC22 family

The duplication that gave rise to *R2d2* also created a new copy of *Cwc22*. Based on our assembly of the *R2d2* sequence, the open reading frame of $Cwc22^{R2d2}$ is intact and encodes a nearly full-length predicted protein that retains the two key functional domains characteristic of the *Cwc22* family. Inspection of RNA-seq data from samples with high copy number at *R2d2* reveals several novel transcript isoforms whose expression appears to be copy-number- and tissue-dependent. In testis, the most abundant isoform retains an intron containing an ERV insertion (red arrow in **Figure 4.11**), consistent with the well-known transcriptional promiscuity in this tissue. The most abundant isoforms in adult brain is unusual in that its stop codon is in an internal exon which is followed by a 7 kbp 3′ UTR in the terminal exon. Transcripts with a stop codon in an internal exon are generally subject to nonsense-mediated decay (NMD) triggered by the presence of exon-junction complexes

[2] Thomas Keane, personal communication

downstream the stop codon. Curiously, *Cwc22* is itself a member of the exon-junction complex[319].

That an essential gene involved in such a central biochemical pathway should both escape NMD and be overexpressed more than tenfold is surprising. Preliminary data from the Diversity Outbred population shows that the $R2d2^{HC}$ allele is associated with elevated levels of both *Cwc22* transcripts and protein in adult liver[320]. Further studies will be required to determine the distribution of transcription and translation of *Cwc22* across isoforms, tissues and developmental stages.

4.4 Conclusions and future directions

Our detailed analysis of the evolutionary trajectory of *R2d2* provides insight into the fate of duplicated sequences over short (within-species) timescales. The exceptionally high mutation rate and low recombination associated specifically with hemizygous $R2d2^{HC}$ alleles motivate hypotheses regarding the biochemical mechanisms which contribute to observed patterns of polymorphism at this and similar loci. Finally, the birth of a new member of the deeply conserved *Cwc22* gene family in *R2d2* provides an opportunity to test predictions regarding the evolution of young duplicate gene pairs.

4.5 Materials and methods

4.5.1 Mice

Wild *M. musculus* mice used in this study were trapped at a large number of sites across Europe, the United States, the Middle East, northern India and Taiwan. Trapping was carried out in accordance with local regulations and with the approval of all relevant regulatory bodies for each locality and institution.

Tissue samples from the progenitors of the wild-derived inbred strains ZALENDE/EiJ (*M. m. domesticus*), TIRANO/EiJ (*M. m. domesticus*) and SPRET/EiJ (*M. spretus*) were provided by Muriel Davisson (The Jackson Lab). Tissue samples from the high running (HR) selection and intercross lines were obtained from Ted Garland (University of California - Riverside). Further details regarding these samples are provided in **??**.

Female Diversity Outbred mice used for estimating mutation rates at *R2d2* were obtained from the Jackson Laboratory and housed with a single FVB/NJ male. Progeny were sacrificed at birth by cervical dislocation in order to obtain tissue for genotyping.

All live laboratory mice were handled in accordance with the IACUC protocols of the University

of North Carolina at Chapel Hill.

4.5.2 DNA preparation

High molecular weight DNA. High molecular weight DNA was obtained for samples genotyped with the Mouse Diversity Array or subject to whole-genome sequencing. Genomic DNA was extracted from tail, liver or spleen using a standard phenol-chloroform procedure[321]. High molecular weight DNA for most inbred strains was obtained from the Jackson Laboratory, and the remainder as a generous gift from Francois Bonhomme and the University of Montpellier Wild Mouse Genetic Repository.

Low molecular weight DNA. Low molecular weight DNA was obtained for samples to be genotyped on the MegaMUGA array (see below). Genomic DNA was isolated from tail, liver, muscle or spleen using Qiagen Gentra Puregene or DNeasy Blood & Tissue kits according to the manufacturer's instructions.

4.5.3 Whole-genome sequencing and variant discovery

Inbred strains. Sequencing data for inbred strains of mice except ZALENDE/EiJ and LEWES/EiJ was obtained from the Sanger Mouse Genomes Project website (`ftp://ftp-mouse.sanger.ac.uk/current_bams`) as aligned BAM files. Details of the sequencing pipeline are given in[29]. Coverage ranged from approximately $25\times$ to $50\times$ per sample.

The strains LEWES/EiJ and ZALENDE/EiJ were sequenced at the University of North Carolina High-Throughput Sequencing Facility. Libraries were prepared from high molecular weight DNA using the Illumina TruSeq kit and insert size approximately 250 bp, and 2×100 bp paired-end reads were generated on an Illumina HiSeq 2000 instrument. LEWES/EiJ was sequenced to approximately $12\times$ coverage and ZALENDE/EiJ to approximately $18\times$[322].

Wild mice. Whole-genome sequencing data from 26 wild *M. m. domesticus* individuals described in[268] was downloaded from ENA under accession #PRJEB9450. Coverage ranged from approximately $12\times$ to $25\times$ per sample. An additional two wild *M. m. domesticus* individuals, IT175 and ES446, were sequenced at the University of North Carolina to approximate coverage $8\times$ each. Raw reads from an additional 10 wild *M. m. castaneus* described in[292], sequenced to approximately $20\times$ each, were downloaded from ENA under accession #PRJEB2176. Reads for a single *Mus caroli* individual sequenced to approximately $40\times$ were obtained from ENA under accession #PRJEB2188. Reads for each sample were realigned to the mm10 reference using `bwa-mem` v0.7.12 with default

parameters[274]. Optical duplicates were removed with `samblaster`[275].

Variant discovery. Polymorphic sites on chromosome 2 in the vicinity of *R2d2* were called using `freebayes` v0.9.21-19-gc003c1e[323] with parameters `--standard-filters` using the Sanger Mouse Genomes Project VCF files as a list of known sites (parameter `-@`). Raw calls were filtered to have quality score > 30, root mean square mapping quality > 20 (for both reference and alternate allele calls) and at most 2 alternate alleles.

4.5.4 Copy-number estimation

R2d copy number was estimated using qPCR as described in[286]. Briefly, we used commercial TaqMan assays against intron-exon boundaries in *Cwc22* (Life Technologies assay numbers Mm00644079_cn and Mm00053048_cn) to determine copy number relative to reference genes *Tert* (cat. #4458368, for target Mm00644079_cn) or *Tfrc* (cat. #4458366, for target Mm00053048_cn). Cycle thresholds for *Cwc22* relative to the reference gene were normalized across assay batches using linear mixed models with batch and target-reference pair treated as random effects. Control samples with known haploid *R2d* copy numbers of 1 (C57BL/6J), 2 (CAST/EiJ), 17 (WSB/EiJ$\times$C57BL/6J)F_1 and 34 (WSB/EiJ) were included in each batch.

Samples were classified as having 1, $2 > 2$ haploid copies of *R2d* using linear discriminant analysis. The classifier was trained on the normalized cycle thresholds of the control samples from each plate, whose precise integer copy number is known, and applied to the remaining samples.

4.5.5 *De novo* assembly of *R2d2*

Raw whole-genome sequencing reads for WSB/EiJ from the Sanger Mouse Genomes Project were converted to a multi-string Burrows-Wheeler transform and associated FM-index (msBWT)[324] using the `msbwt` v0.1.4 Python package (`https://pypi.python.org/pypi/msbwt`). The msBWT and FM-index implicitly represent a suffix array of sequencing to provide efficient queries over arbitrarily large string sets. Given a seed k-mer present in that string set, this property can be exploited to rapidly construct a de Bruijn graph which can in turn be used for local *de novo* assembly of a target sequence (**Figure 4.3A**). The edges in that graph can be assigned a weight (corresponding to the number of reads containing the $(k+1)$-mer implied by the edge) which can be used to evaluate candidate paths when the graph branches (**Figure 4.3B**).

R2d2 was seeded with the 31 bp sequence (`TCTAGAGCATGAGCCTCATTTATCATGCCT`) at the proximal boundary of *R2d1* in the GRCm38/mm10 reference genome. A single linear contig was

assembled by "walking" through the local de Bruijn graph. Because WSB/EiJ has 33 copies of *R2d2* and a single copy of *R2d1*, any branch point in the graph which represents a paralogous variant should having outgoing edges with weights differing by a factor of approximately 33. Furthermore, when two (or more) branch points occur within less than the length of a read, it should be possible to "phase" the underlying variants by following single reads through both branch points (**Figure 4.3B**). We used these heuristics to assemble the sequence of *R2d2* (corresponding to the higher-weight path through the graph) specifically.

After assembling a chunk of approximately 500 bp the contig was checked for colinearity with the reference sequence (*R2d1*) using `BLAT` and `CLUSTAL-W2` (using the EMBL-EBI web server: `http://www.ebi.ac.uk/Tools/msa/clustalw2/`).

Repetitive elements such as retroviruses are refractory to assembly with our method. Upon traversing into a repetitive element, the total edge weight (total number of reads) and number of branch points (representing possible linear assembled sequences) in the graph become large. It was sometimes possible to assemble a fragment of a repetitive element at its junction with unique sequence but not to assemble unambiguously across the repeat. Regions of unassembleable sequence were marked with blocks of Ns, and assembly re-seeded using a nearby k-mer from the reference sequence.

The final contig was checked against its source msBWT by confirming that each 31-mer in the contig which did not contain an `N` was supported by at least 60 reads. A total of 16 additional haplotypes in 8 regions of *R2d* totaling 16.9 kbp (**Table 4.4**) were assembled in a similar fashion, using the WSB *R2d2* contig and the *R2d1* reference sequence as guides.

4.5.6 Sequence analysis of *R2d2* contig

Pairwise alignment of R2d *paralogs.* The reference *R2d1* sequence and our *R2d2* contig were aligned using `LASTZ` v1.03.54 (`http://www.bx.psu.edu/~rsharris/lastz/`) with parameters `–step=10 –seed=match12 –notransition –exact=20 –notrim –identity=95`.

Transposable element (TE) content. The *R2d2* contig was screened for TE insertions using the `RepeatMasker` web server (`http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker`) with species set to "mouse" and default settings otherwise. As noted previously, we could not assemble full-length repeats, but the fragments we could assemble at junctions with unique sequence allowed identification of some candidate TEs to the family level. *R2d1*-specific TEs were defined

Start	End	Label
77869657	77870807	A
77879768	77880997	B
77908401	77910062	C
77914268	77915817	D
77919218	77920466	E
77920467	77920996	F
77945777	77949523	G
77949564	77950925	H
77979000	77980657	I
77980658	77981322	J
78010326	78011760	K
78011761	78012421	L

Table 4.4: Regions of *R2d* targeted for *de novo* assembly in inbred strains. Coordinates are on chromosome 2 in the mm10 reference assembly.

as TEs annotated in the RepeatMasker track at the UCSC Genome Browser with no evidence (no homologous sequence, and no `N`s) at the corresponding position in the *R2d2* contig. Candidate *R2d2*-specific TEs were defined as gaps ≥ 100 bp in size in the alignment to *R2d1* for which the corresponding *R2d2* sequence was flagged by `RepeatMasker`.

Gene conversion tracts. To unambiguously define gene conversion events without confounding from paralogous sequence, we examined 15 wild *M. m. domesticus* samples and 37 laboratory strains with evidence of 2 diploid copies of *R2d*. We confirmed that these copies of *R2d* were located at *R2d1* by finding read pairs spanning the junction between *R2d1* and neighboring sequence. Gene conversion tracts were delineated as clusters of derived alleles shared with *R2d2*. Using a pairwise of alignment of *R2d2* and *R2d1* we identified single-nucleotide variants between the two sequences, and queried those sites in aligned reads for *Mus caroli*. If the *Mus caroli* and *R2d1* shared an allele, we recorded the site as a derived allele informative for the presence of *R2d2*. We used the resulting list of $1,411$ informative sites to query aligned reads for the samples of interest and recorded, for each site and each sample, whether the derived allele (*R2d2*), ancestral allele (*R2d1*) or both alleles were present. Conversion tracts were then identified by manual inspection. Boundaries of conversion tracts were defined at approximately the midpoint between the first *R2d1*- (or *R2d2*-)

specific variant and the last *R2d2*- (or *R2d1*-) specific variant.

Sequence diversity in R2d1 *and* R2d2. Assembling individual repeat units within *R2d2* is infeasible in high-copy samples. Instead we treated each *R2d* unit as an independent sequence and used the number of segregating sites to estimate sequence diversity. Segregating sites were defined as positions in a collection of alignments (BAM files) with evidence of an alternate allele. To identify segregating sites we used `freebayes` with parameters `-ui -Kp 20 -use-best-n-alleles 2 -m 8`. These parameters treat each sample as having ploidy up to 20, impose an uninformative prior on genotype frequencies, and limit the algorithm to the discovery of atomic variants (SNVs or short indels, not multinucleotide polymorphisms or other complex events) with at most 2 alleles at each segregating site. Sites in low-complexity sequence (defined as Shannon entropy < 1.6 in the 30 bp window centered on the site) or within 10 bp of another variant site were further masked, to minimize spurious calls due to ambiguous alignment of indels and microsatellite variants. To avoid confounding with the retrocopies of *Cwc22* outside *R2d*, coding exons of *Cwc22* were masked. Finally, sites corresponding to an unaligned or gap position in the pairwise alignment between *R2d1* and *R2d2* were masked.

To compute diversity in *R2d1* we counted segregating sites in 12 wild *M. m. domesticus* samples with 2 diploid copies of *R2d* (total of 24 sequences), confirmed to be in *R2d1* by the presence of read pairs spanning the junction between *R2d1* and neighboring sequence. To compute diversity in *R2d2*, we counted segregating sites in 14 wild *M. m. domesticus* samples with > 2 diploid copies of *R2d* (range $3-83$ per sample; total of 406 sequences) but excluded sites corresponding to variants among *R2d1* sequences. Remaining sites were phased to *R2d2* by checking for the presence of a 31-mer containing the site and the nearest *R2d1*-vs-*R2d2* difference in the raw reads for each sample using the corresponding msBWT. Sequence diversity was then computed using Watterson's estimator[83], dividing by the number of alignable bases ($128,973$) to yield a per-site estimate. Standard errors were estimated by 100 rounds of resampling over the columns in the *R2d1*-vs-*R2d2* alignment.

4.5.7 Microarray genotyping

Genome-wide genotyping was performed using MegaMUGA, the second version of the Mouse Universal Genotyping Array platform (Neogen/GeneSeek, Lincoln, NE)[95]. Genotypes were called using the `GenCall` algorithm implemented in the Illumina BeadStudio software (Illumina Inc, Carlsbad, CA). For quality control we computed, for each marker i on the array: $S_i = X_i + Y_i$,

where X_i and Y_i are the normalized hybridization intensities for the two alleles. The expected distribution of S_i was computed from a large set of reference samples. We excluded arrays for which the distribution of S_i was substantially shifted from this reference; in practice, failed arrays can be trivially identified in this manner[95]. Additional genotypes for inbred strains and wild mice from the Mouse Diversity Array were obtained from[157].

4.5.8 Analyses of *Cwc22* expression

RNA-seq read alignment. Expression of *Cwc22* was examined in adult whole brain using data from[298], SRA accession #SRP056236. Paired-end reads (2×100 bp) were obtained from 8 replicates each of 3 inbred strains: CAST/EiJ, PWK/PhJ and WSB/EiJ. Raw reads were aligned to the mm10 reference using `STAR` v2.4.2a[325] with default parameters for paired-end reads. Alignments were merged into a single file per strain for further analysis. Expression in adult testis was examined in 23 wild-derived inbred strains from[299], SRA accession #PRJNA252743. Single-end reads (76 bp) were aligned to the mm10 genome with `STAR` using default parameters for single-end, non-strand-specific reads.

Transcript assembly. Read alignments were manually inspected to assess support for *Cwc22* isoforms in the Ensembl v83 annotation. To identify novel isoforms in *R2d2*, we applied the `Trinity` v0.2.6 pipeline[326] to the subset of reads from WSB/EiJ which could be aligned to *R2d1* plus their mates (a set which represents a mixture of *Cwc22*^{R2d1} and *Cwc22*^{R2d2} reads). *De novo* transcripts were aligned both to the mm10 reference and to the *R2d2* contig using `BLAT`, and were assigned to *R2d1* or *R2d2* based on sequence similarity. Because expression from *R2d2* is high in WSB/EiJ, *R2d2*-derived transcripts dominated the assembled set. Both manual inspection and the Trinity assembly indicated the presence of retained introns and an extra 3′ exon, as described in the **Results**. To obtain a full set of *Cwc22* transcripts including those of both *R2d1* and *R2d2* origin, we supplemented the *Cwc22* transcripts in Ensembl v83 with their paralogs from *R2d2* as determined by a strict `BLAT` search against the *R2d2* contig. We manually created additional transcripts reflecting intron-retention and 3′ extension events described above, and obtained their sequence from the *R2d2* contig.

Abundance estimation. Relative abundance of *Cwc22* paralogs was estimated using `kallisto` v0.42.3[300] with parameters `-bias` (to estimate and correct library-specific sequence-composition biases). The transcript index used for pseudoalignment and quantification included only the *Cwc22*

targets.

4.5.9 Phylogenetic analyses

Tree for R2d. Multiple sequence alignments for sub-regions of *R2d* were generated using `MUSCLE`[327] with default parameters. The resulting alignments were manually trimmed and consecutive gaps removed. Phylogenetic trees were inferred with `RAxML` v8.1.9[328] using the `GTR+gamma` model with 4 rate categories and *M. caroli* as an outgroup. Uncertainty of tree topologies was evaluated using 100 bootstrap replicates.

Divergence time. The time of the split between *R2d1* and *R2d2* was estimated using the Bayesian method implemented in `BEAST` v1.8.1r6542[289]. We assumed a divergence time for *M. caroli* of 5 Mya and a strict molecular clock, and analyzed the concatenated alignment for our *de novo* assembled regions under the `GTR+gamma` model with 4 rate categories and allowance for a proportion of invariant sites. The chain was run for 1×10^7 iterations with trees sampled every 1000 iterations.

Local phylogeny around R2d2. Genotypes for 173 SNPs in the region surrounding *R2d2* (chr2: 83 — 84 Mb) were obtained for 90 individuals representing both laboratory and wild mice genotyped with the Mouse Diversity Array[157]. Individuals with evidence of heterozygosity (> 3 heterozygous calls) were excluded to avoid ambiguity in phylogenetic inference. A distance matrix for the remaining 62 samples was created by computing the proportion of alleles shared identical by state between each pair of samples. A neighbor-joining tree was inferred from the distance matrix and rooted at the most recent common ancestor of the *M. musculus*- and non-*M. musculus* samples.

Cwc22 *coding sequences*. To create the tree of *Cwc22* coding sequences, we first obtained the sequences of all its paralogs in mouse. The coding sequence of $Cwc22^{R2d1}$ (RefSeq transcript NM_030560.5) was obtained from the UCSC Genome Browser and aligned to our *R2d2* contig with `BLAT` to extract the exons of $Cwc22^{R2d2}$. The coding sequence of retro-*Cwc22* (genomic sequence corresponding to GenBank cDNA AK145290) was obtained from the UCSC Genome Browser. Coding and protein sequences of *Cwc22* homologs from non-*M. musculus* species were obtained from Ensembl[115]. The sequences were aligned with MUSCLE and manually trimmed, and a phylogenetic tree estimated as described above.

We observed that the branches in the rodent clade of the *Cwc22* tree appeared to be longer than branches for other taxa. We used `PAML`[115] to test the hypothesis that *Cwc22* is under relaxed purifying selection in rodents using the branch-site model (null model `model = 2, NSsites = 2,`

`fix_omega = 1`; alternative model `model = 2, NSsites = 2, omega = 1, fix_omega = 1`) as described in the `PAML` manual. This is a test of difference in evolutionary rate on a "foreground" branch (ω_1) — in our case, the rodent clade — relative to the tree-wide "background" rate (ω_0), with $H_0 : \omega_0 = \omega_1$ and $H_a : \omega_0 < \omega_1$. The distribution of the test statistic is an even mixture of a χ^2 distribution with 1 df and a point mass at zero; to obtain the p-value, we calculated the quantile of the χ^2 distribution with 1 df and divided by 2.

4.5.10 Analyses of recombination rate at *R2d2*

To test the effect of *R2d2* copy number on local recombination rate examined recombination events accumulated during the first 16 generations of breeding of the DO population, in which the high-copy *R2d2* allele from WSB/EiJ is segregating. Founder haplotype reconstructions were obtained for $4,640$ DO individuals (a subset of those in Chapter 3), and recombination events were identified as junctions between founder haplotypes. We compared the frequency of junctions involving a WSB/EiJ haplotype to junctions not involving a WSB/EiJ haplotype over the region chr2: 75-90 Mb. Within each generation we also tested for differences in the lengths of haplotype blocks overlapping *R2d2* using one-sided Wilcoxon rank-sum tests (alternative hypothesis: WSB/EiJ haplotypes longer than others). Resulting p-values were subject to Bonferroni correction: for nominal significance level $alpha = 0.01$, the corrected threshold is $p = \frac{0.01}{12} = 8.3 \times 10^{-4}$.

We also estimated the difference between observed and expected recombination fraction in 11 experimental crosses in which one of the parental lines was segregating for a high-copy allele at *R2d2*. We obtained expected recombination fractions from the standard mouse genetic map[190], which was constructed from crosses between strains lacking *R2d2*^{HC} alleles. Genotype data was obtained from The Jackson Laboratory's Mouse Phenome Database QTL Archive (`http://phenome.jax.org/db/q?rtn=qtl/home`). Recombination fractions were calculated using `R/qtl` (`http://rqtl.org/`). Confidence intervals for difference between observed and expected recombination fractions were calculated by 100 iterations of nonparametric bootstrapping over individuals in each dataset.

CHAPTER 5

Selfish selection on a structural variant

5.1 Introduction

[1] Population-level sequencing data have enabled analyses of positive selection in many species, including mice[329] and humans[330,331,332]. These studies seek to identify genetic elements, such as single nucleotide variants (SNVs) and copy number variants (CNVs), that are associated with phenotypic differences between populations that share a common origin[333,334]. A marked difference in local genetic diversity between closely related taxa might indicate that one lineage has undergone a sweep. During a sweep, a variant under strong positive selection rises in frequency and carries with it linked genetic variation ("genetic hitch-hiking"), thereby reducing local haplotype diversity[84,335]. In genomic scans for sweeps, it is typically assumed that the driving allele will have a strong positive effect on organismal fitness. Prominent examples of sweeps for which this assumption holds true (*i.e.* classic selective sweeps) include alleles at the *Vkorc1* locus, which confers rodenticide resistance in the brown rat[336], and enhancer polymorphisms conferring lactase persistence — the ability to digest milk into adulthood — in human beings[337]. However, we and others have suggested that selfish alleles that strongly promote their own transmission irrespective of their effects on overall fitness could give rise to genomic signatures indistinguishable from those of classic selective sweeps[338,339,340,341,342,246].

Suggestive evidence that sweeps may be driven by selfish alleles comes from studies in

[1] The results presented in this chapter are published in:

> Didion JP*, Morgan AP*, Yadgary L *et al.* (2016) *R2d2* drives selfish sweeps in the house mouse. *Mol Biol Evol* **33**:1381–1395. PMID 26882987.

This project was a joint effort of John Didion, Liran Yadgary and the author, with additional important contributions from Tim Bell, Lydia Ortiz de Solorzano, Rachel McMullan, and a large network of collaborators who contributed DNA samples from wild mice. Ted Garland contributed DNA samples and whole-genome sequence data from HR selection lines.

Drosophila. Incomplete sweeps have been identified at the *Segregation Distorter* (SD) locus[343] and in at least three X-chromosome systems[344,345,346,347], all of which drive through the male germline. In addition, genomic conflict has been proposed as a possible driver of two nearly complete sweeps in *D. mauritiana*[348]. Incomplete sweeps were also detected in natural populations of *Mimulus* (monkeyflower); the cause was identified as female meiotic drive of the centromeric *D* locus[349]. The fact that all evidence of selfish sweeps derives from two genera is to some extent reflective of a observational bias, but may also indicate a difference in the incidence or effect of selfish alleles between these taxa and equally well-studied mammalian species (*e.g.* humans and mice). Furthermore, the lack of completed selfish sweeps reported in the literature may be due to an unexpected strength of balancing selection, in which the deleterious effects of selfish alleles prevent them from driving to fixation, or due to insufficient methods of detection.

Here, we investigate whether a selfish allele can sweep in natural and laboratory populations of the house mouse, *M. m. domesticus*. The *R2d2* locus is introduced and described in detail in **Chapter 4**. Briefly, *R2d2* is a copy number gain of a 127 kb core element that contains a single protein-coding gene, *Cwc22* (a member of the mRNA-splicing complex.) Females heterozygous for *R2d2* preferentially transmit to their offspring an allele with high copy number ($R2d2^{HC}$) relative to an allele with low copy number ($R2d2^{LC}$), where "high copy number" is the minimum copy number with evidence of distorted transmission in existing experiments — approximately 7 units of the core element. In contrast to many meiotic drive systems, in which the component elements are tightly linked, the action of $R2d2^{HC}$ is dependent on unlinked modifier loci whose frequencies, modes of action, and effect sizes are unknown[286]. These modifier loci modulate the degree of transmission distortion; as a result, distorted transmission is present in some laboratory crosses segregating for $R2d2^{HC}$ alleles, but absent in others[350,351,290,352,353,354]. In this chapter we show that $R2d2^{HC}$ genotype is either uncorrelated or negatively correlated with litter size — a major component of absolute fitness in mice — depending on the presence of meiotic drive. $R2d2^{HC}$ therefore behaves as a selfish genetic element. In the We provide evidence of a recent "selfish selective sweep" at $R2d2^{HC}$ in wild *M. m. domesticus* mice and show that $R2d2^{HC}$ has repeatedly driven selfish sweeps in closed-breeding mouse populations.

Population	Freq	$2N$
BE	0.50	6
CH	0.32	28
CY	0.00	14
DE	0.67	6
DK	0.06	18
EC	0.00	24
ES	0.22	18
FR	0.15	26
GR	0.08	106
IT	0.09	34
LB	0.25	8
PT	0.13	54
TN	0.00	4
UK	0.00	6
USE	0.21	102
USW	0.00	24

Table 5.1: Table 1. *R2d2*^{HC} allele frequencies in wild *M. m. domesticus* populations. Populations are given as ISO country codes, except for USE (US East Coast - Maryland) and USW (US West Coast - Farallon Island).

5.2 Results

5.2.1 Evidence for a selfish sweep in wild mouse populations

As described in **Chapter 4**, copy number at *R2d2* varies from 0 to > 60 in wild mice. Here we analyzed copy number by population **Figure 5.1** and found that *R2d2*^{HC} alleles are segregating at a wide range of frequencies in natural populations ($0.00 - 0.67$; **Table 5.1**).

To test for a selfish sweep at *R2d2*^{HC}, we genotyped the wild-caught mice on the MegaMUGA array[95] and examined patterns of haplotype diversity on chromosome 2. In the case of strong positive selection, unrelated individuals are more likely to share extended segments that are identical by descent in the vicinity of the selected locus[355] compared with a population subject only to genetic drift. Consistent with this prediction, we observed an extreme excess of shared identity by descent (IBD) across populations around *R2d2* (**Figure 5.2A**): *R2d2* falls in the top 0.25% of

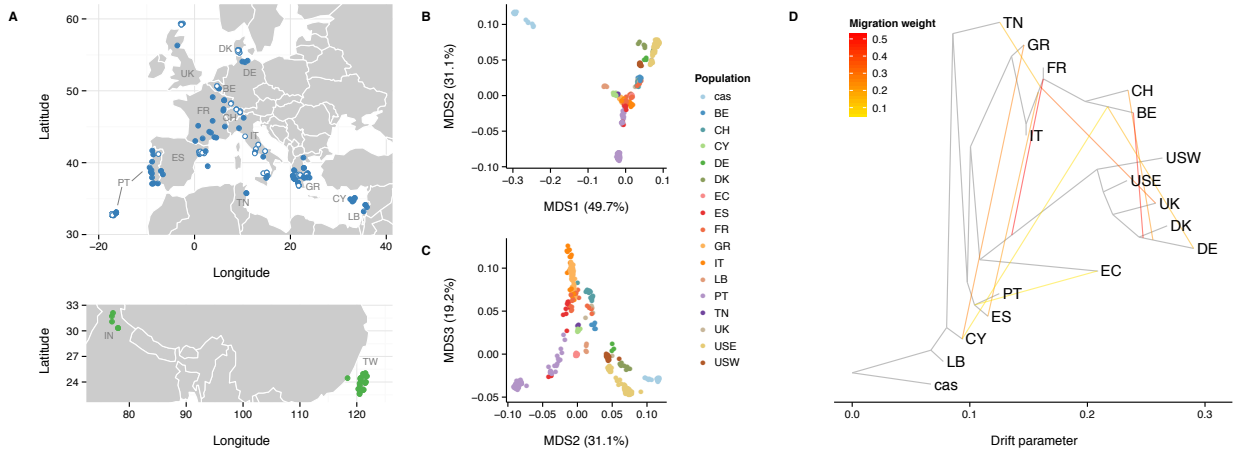

Figure 5.1: Wild mouse populations tested for *R2d2* status. (**A**) Geographic distribution of samples used in this study. Samples are colored by taxonomic origin: blue for *M. m. domesticus*, green for *M. m. castaneus*. Those with standard karyotype ($2n = 40$) are indicated by closed circles; samples with Robertsonian fusion karyotypes ($2n < 40$) are indicated by open circles. Populations from Floreana Island (Galapagos Islands, Ecuador; "EC"), Farallon Island (off the coast of San Francisco, California, United States; "USW"), and Maryland, United States ("USE") are not shown. **(B,C)** Multidimensional scaling (MDS) ($k = 3$ dimensions) reveals population stratification consistent with geography. *M. m. domesticus* populations are labeled by country of origin. Outgroup samples of *M. m. castaneus* origin cluster together ("cas"). **(D)** Population graph estimated from autosomal allele frequencies by TreeMix. Black edges indicate ancestry, while colored edges indicate gene flow by migration or admixture (with yellow to red indicating increasing probability of migration). Topography of the population graph is consistent with MDS result and with the geographic origins of the samples.

Chr	Start (Mb)	End (Mb)	Locus	Score
2	79.75	85.75	*R2d2*	0.108
4	3.25	7.75		0.051
4	149	149.5		0.045
5	113	113.5		0.045
7	35	36		0.049
7	132.75	137.25	*Vkorc1*	0.154 *
8	116.5	118		0.076
10	86.25	89		0.098
13	70	71.75		0.068
17	26.75	27.75	MHC; *t*-hap.	0.05 *
18	12.5	13.75		0.049
18	33	35.5		0.216

Table 5.2: The 12 loci above the 99^{th} percentile of IBD-sharing scores. Chromosome locations are given based on mouse genome build GRCm38/mm10. Loci identified as targets of positive selection are named and candidate targets of selection identified in wild mice in a previous study[329] are marked with an asterisk.

IBD-sharing scores across the autosomes. In all cases, the shared haplotype has high copy number and this haplotype appears to have a single origin in European mice (**Figure 5.4** and **Chapter 4**). Strong signatures of selection are also evident at a previously identified target of positive selection, the *Vkorc1* locus (distal chromosome 7)[356]. The 12 loci in the top 1% of IBD-sharing scores are shown in **Table 5.2**.

In principle, the strength and age of a sweep can be estimated from the extent of loss of genetic diversity around the locus under selection. From the SNP data, we identified a ~ 1 Mb haplotype with significantly greater identity between individuals with $R2d2^{HC}$ alleles compared to the surrounding sequence. We used published sequencing data from 26 wild mice[268] to measure local haplotype diversity around *R2d2* and found that the haplotypes associated with $R2d2^{HC}$ alleles are longer than those associated with $R2d2^{LC}$ (**Figure 5.2B-C**). This pattern of extended haplotype homozygosity is consistent with positive selection over an evolutionary timescale as short as 450 generations (see **§ 5.5**). However, due to the extremely low rate of recombination in the vicinity of *R2d2* (see **Chapter 4**), this is most likely an underestimate of the true age of the mutation. The fact

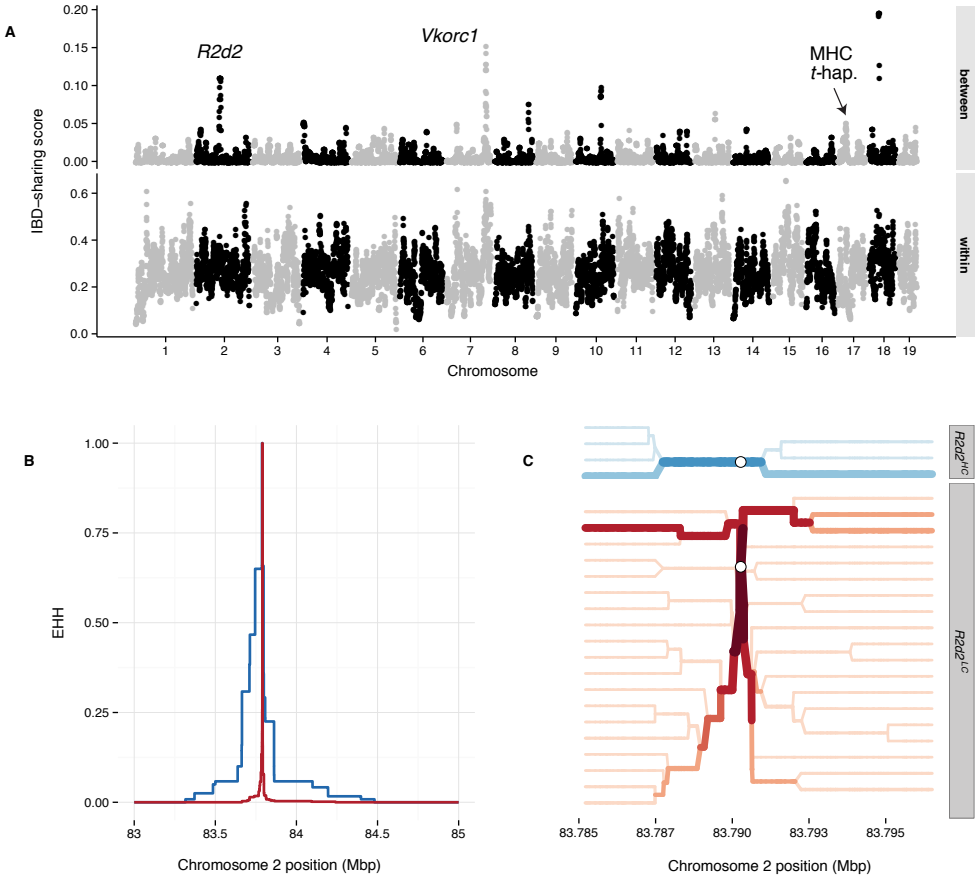

Figure 5.2: Haplotype-sharing at *R2d2* provides evidence of a selective sweep in wild mice of European origin. **(A)** Weighted haplotype-sharing score (see **§ 5.5**) computed in 500 kb bins across autosomes, within which individuals are drawn from the same population (lower panel) or different populations (upper panel). Peaks of interest overlay *R2d2* (chromosome 2; see **Figure 5.3** for zoomed-in view) and *Vkorc1* (distal chromosome 7). The position of the closely linked *t*-haplotype and major histocompatibility (MHC) loci is also marked. **(B)** Decay of extended haplotype homozygosity (EHH)[357] on the $R2d2^{HC}$-associated (blue) versus the $R2d2^{LC}$-associated (red) haplotype. EHH is measured outward from the index SNP at chr2:83,790,275 and is bounded between 0 and 1. **(C)** Haplotype bifurcation diagrams for the $R2d2^{HC}$ (top panel, blue) and $R2d2^{LC}$ (bottom panel, red) haplotypes at the index SNP (open circle). Darker colors and thicker lines indicate higher haplotype frequencies. Haplotypes are truncated 100 sites in each direction from the index SNP.

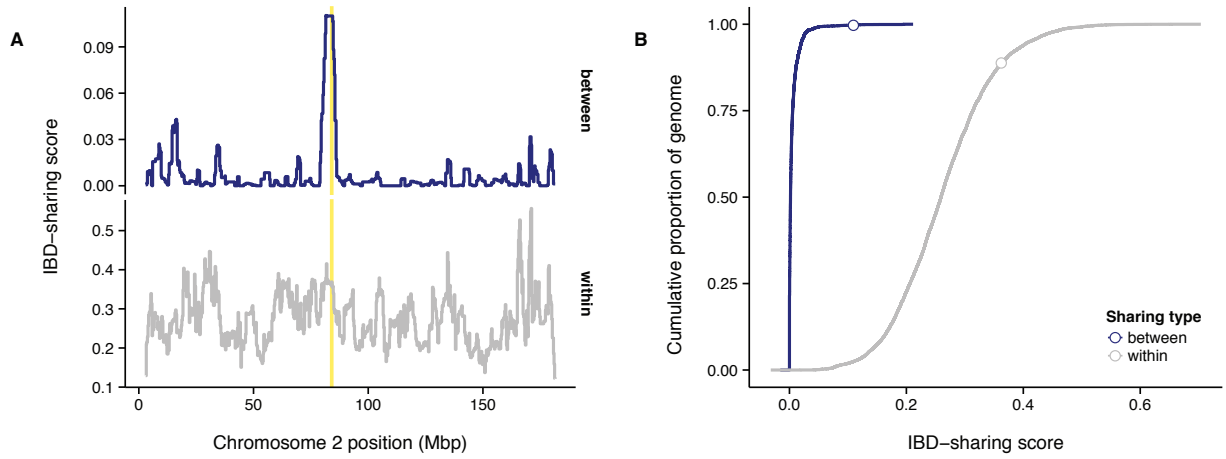

Figure 5.3: Haplotype-sharing on Chromosome 2 among wild mice of European origin. **(A)** Weighted haplotype-sharing score (see **§ 5.5**), computed in 500 kb bins across chromosome 2, when those individuals are drawn from the same population (grey line, lower panel) or different populations (blue line, upper panel). Interval containing *R2d2* is indicated by yellow shaded region. This panel is a magnified view of **Figure 5.2**. **(B)** Cumulative distribution of IBD-sharing probability across all autosomes either within (grey line) or between (blue line) populations. Open circles indicate value at *R2d2*.

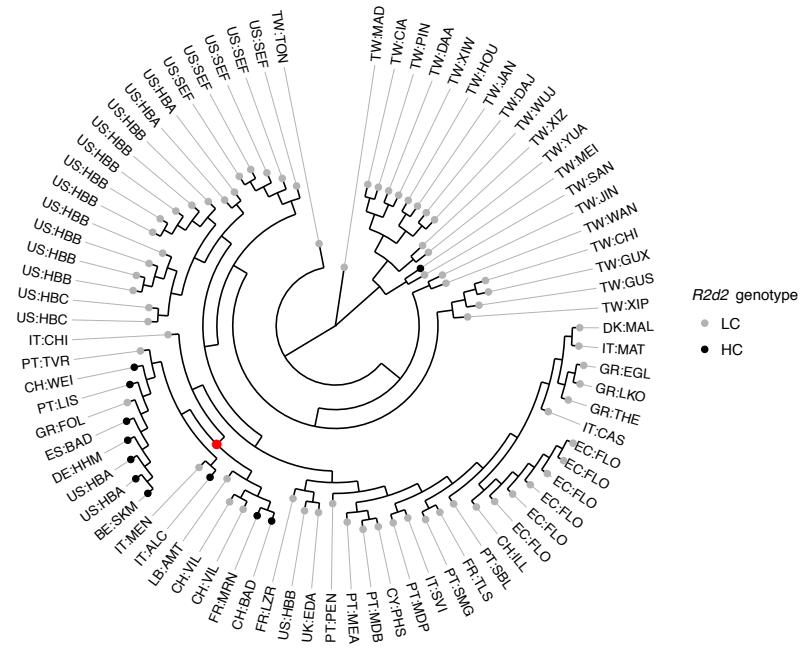

Figure 5.4: Local phylogeny at *R2d2*. Neighbor-joining tree constructed from genotypes at 39 SNPs in the interval chr2: 82 – 85 Mb. To avoid phase uncertainty, only the 80 individuals homozygous across the entire region are included. Names are abbreviated to country and locality code. The node marked in red is the most recent common ancestor of all $R2d2^{HC}$ haplotypes except one *M. m. castaneus* sample (TW:MEI) from an admixed population whose $R2d2^{HC}$ allele likely represents either genotyping error or an introgression event.

that the allele is widespread in Europe and the Americas as well as in mice from multiple locations in Iran argues in favor of less recent origin — at least $10,000$ years ago, before *M. m. domesticus* spread out of the near East[132].

It is important to note that the excess IBD we observe at *R2d2* (**Figure 5.2A**) arises from segments shared *between* geographically distinct populations (**Figure 5.1**). When considering sharing *within* populations only (**Figure 5.3**), *R2d2* is no longer an outlier. Therefore, it was unsurprising that we failed to detect a sweep around *R2d2* using statistics that are designed to identify population-specific differences in selection, like hapF_{LK}[358], or selection in aggregate, like iHS[359] (**Figure 5.5**).

5.2.2 A selfish sweep in the Diversity Outbred population

We validated the ability of *R2d2*^{HC} to drive a selfish sweep by examining *R2d2* allele frequencies in multiple closed-breeding laboratory populations for which we had access to samples from the founder populations. The Diversity Outbred (DO) is a randomized outbreeding population derived from eight inbred mouse strains that is maintained under conditions designed to minimize the effects of both selection and genetic drift[222]. Expected time to fixation or loss of an allele present in the founder generation (with initial frequency of $1/8$) is approximately 900 generations. The WSB/EiJ founder strain contributed an *R2d2*^{HC} allele which underwent more than a three-fold increase (from 0.18 to 0.62) in 13 generations ($p < 0.001$ by simulation with drift only; range $0.03-0.26$ after 13 generations in 1000 simulation runs) (**Figure 5.6A**), accompanied by significantly distorted allele frequencies ($p < 0.001$ by simulation) across a 100 Mb region linked to the allele (**Figure 5.6B-C**).

5.2.3 *R2d2*^{HC} has an underdominant effect on fitness

The fate of a selfish sweep depends on the fitness costs associated with the different genotypic classes at the selfish genetic element. For example, maintenance of intermediate frequencies of the *M. musculus t*-complex[360] and *Drosophila SD*[361] chromosomes in natural populations is thought to result from decreased fecundity associated with those selfish elements.

To assess the fitness consequences of *R2d2*^{HC}, we treated litter size as a proxy for absolute fitness (**Figure 5.7A**). We determined whether each female had distorted transmission of *R2d2* using a one-sided exact binomial test for deviation from the expected Mendelian genotype frequencies in her progeny. Average litter size among DO females homozygous for *R2d2*^{LC} ("LL" in **Figure 5.7A**: 8.1; 95% CI $7.8-8.3$; $N = 339$) is not different from females homozygous for *R2d2*^{HC} ("HH": 8.1;

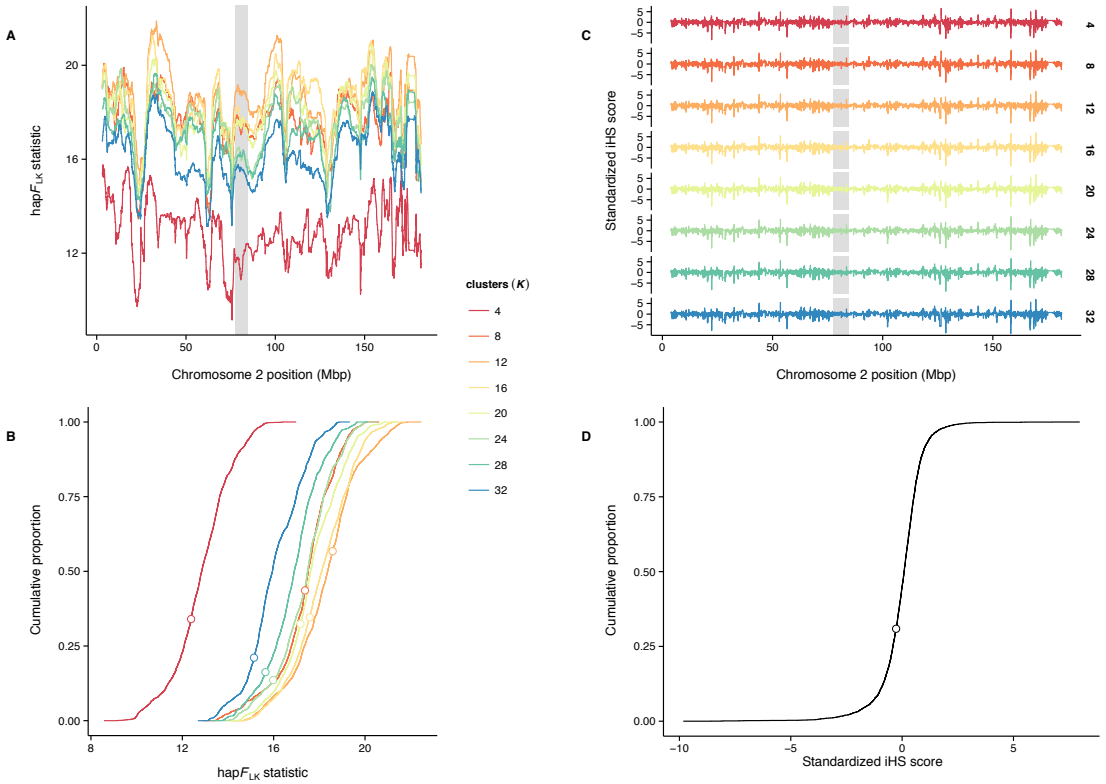

Figure 5.5: Tests for selection based on population differentiation and haplotype length do not detect sweeps at *R2d2*. **(A)** Plot of hapF_{LK} statistic along chromosome 2, for a range of values of the model parameter K (number of local haplotype clusters). **(B)** Cumulative distribution of hapF_{LK} across autosomes, for a range of values of K. Value of the statistic at *R2d2* is indicated by open circle. **(C)** Plot of standardized iHS score along chromosome 2 after phasing with `fastPHASE`, for a range of values of K. **(D)** Cumulative distribution of standardized iHS scores across autosomes after `fastPHASE` with $K = 12$. Value of the statistic at *R2d2* is indicated by open circle.

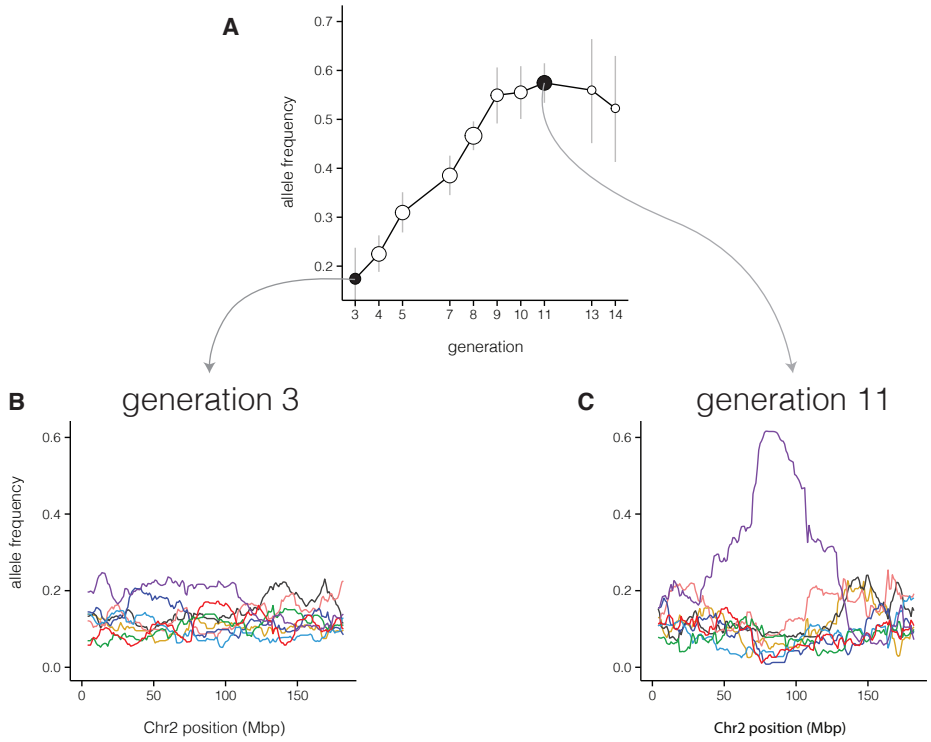

Figure 5.6: An $R2d2^{HC}$ allele rises to high frequency in the DO. **(A)** *R2d2* drives three-fold increase in WSB/EiJ allele frequency in 13 generations in the DO population. Circle sizes reflect number of chromosomes genotyped ($2N$); error bars are ± 2 SE. **(B,C)** Founder allele frequencies across chromosome 2 (averaged in 1 Mb bins) at generations 3 and 13 of the DO.

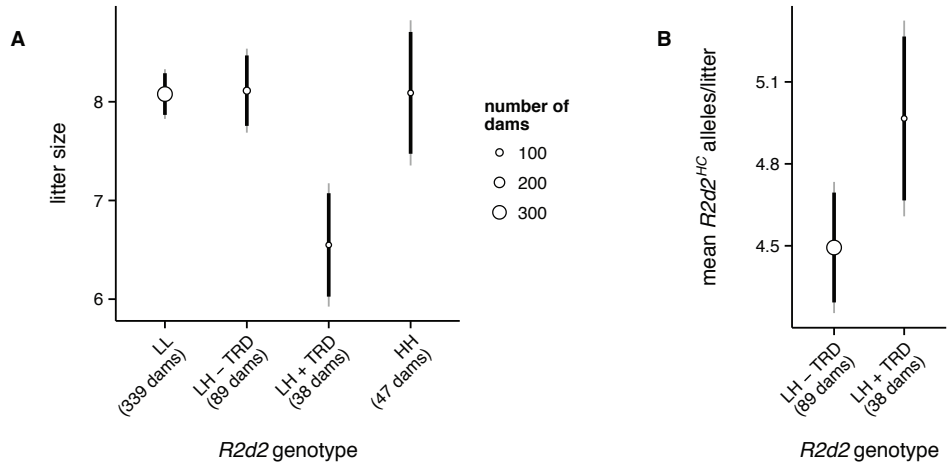

Figure 5.7: *R2d2*^{HC} has underdominant effect on fitness in females. **(A)** Mean litter size among DO females according to *R2d2* genotype: LL, *R2d2*$^{LC/LC}$; LH – TRD, *R2d2*$^{LC/HC}$ without transmission ratio distortion; LH + TRD, *R2d2*$^{LC/HC}$ with transmission ratio distortion; HH, *R2d2*$^{HC/HC}$. Circle sizes reflect number of females tested; error bars are 95% confidence intervals from a linear mixed model which accounts for parity and repeated measures on the same female (see **§ 5.5**.) **(B)** Mean absolute number of *R2d2*^{HC} alleles transmitted in each litter by heterozygous females with (LL + TRD) or without (LL – TRD) transmission ratio distortion. LL + TRD females transmit more *R2d2*^{HC} alleles despite their significantly reduced litter size.

95% CI $7.4 - 8.7$; $N = 47$) or heterozygous females without distorted transmission of *R2d2*^{HC} ("LH-TRD": 8.1; 95% CI $7.7 - 8.5$; $N = 89$). However, in the presence of meiotic drive, litter size is markedly reduced ("LH+TRD": 6.5; 95% CI $5.9 - 7.2$; $N = 38$; $p = 3.7 \times 10^{-5}$ for test of difference versus all other classes). The relative fitness of heterozygous females with distorted transmission is $w = 0.81$, resulting in a selection coefficient of $s = 1 - w = 0.19$ (95% CI $0.10 - 0.23$) against the heterozygote. Despite this underdominant effect, the absolute number of *R2d2*^{HC} alleles transmitted by heterozygous females in each litter is significantly higher in the presence of meiotic drive than its absence ($p = 0.032$; **Figure 5.7B**). The rising frequency of *R2d2*^{HC} in the DO thus represents a truly selfish sweep.

5.2.4 Selfish sweeps in other laboratory populations

We also observed selfish sweeps in selection lines derived from the ICR:Hsd outbred population[290] in which *R2d2*^{HC} alleles are segregating. Three of four lines selectively bred for high

voluntary wheel-running (HR lines) and two of four control lines (10 breeding pairs per line per generation in both conditions) went from starting *R2d2*^{HC} frequencies of ~ 0.75 to fixation in 60 generations or less — two lines were fixed by generation 20, and three more by generation 60 (**Figure 5.8A**). In simulations mimicking this breeding design and assuming normal Mendelian transmission (**Figure 5.8B**), median time to fixation was 46 generations (5^{th} percentile: 9 generations). Although the *R2d2*$^{\text{HC}}$ allele would be expected to eventually fix by drift in 6 of 8 lines given its high starting frequency, the observed rates of fixation were more rapid than expected ($p = 0.003$ in 1000 simulation runs). In a related advanced intercross segregating for high and low copy number alleles at *R2d2* (HR8×C57BL/6J[354]), we observed that *R2d2*^{HC} increased from a frequency of 0.5 to 0.85 in just 10 generations and fixed by 15 generations (**Figure 5.8C**) versus a median 184 generations in simulations ($p < 0.001$; **Figure 5.8D**). The increase in *R2d2*^{HC} allele frequency in the DO and advanced intercross populations occurred at least an order of magnitude faster than would have been predicted by drift alone.

Using archival tissue samples, we were able to determine *R2d2* allele frequencies in the original founder populations of 6 (out of ~ 60) wild-derived inbred strains currently available for laboratory use[118]. In four strains — WSB/EiJ, WSA/EiJ, ZALENDE/EiJ, and SPRET/EiJ — *R2d2*^{HC} alleles were segregating in the founders and are now fixed in the inbred populations. In the other two strains, LEWES/EiJ and TIRANO/EiJ, the founders were not segregating for *R2d2* copy number and the inbred populations are fixed, as expected, for *R2d2*^{LC} (**Figure 5.9**). This trend in wild-derived strains is additional evidence of the tendency for *R2d2*^{HC} to go to fixation in closed breeding populations.

5.3 Discussion

5.3.1 Why has the *R2d2*^{HC} allele not fixed?

Considering the degree of transmission distortion in favor of *R2d2*^{HC} (up to 95%[286]) and that *R2d2*^{HC} repeatedly goes to fixation in laboratory populations, the moderate frequency of *R2d2*^{HC} in the wild (0.14 worldwide, **5.1**) is initially surprising. We find no obvious association between geography and *R2d2*^{HC} allele frequency that might indicate the mutation's precise origin or its pattern of gene flow (**Table 5.1** and **Figure 5.1**).

Several observations may explain these results. First, our sampling was geographically sparse

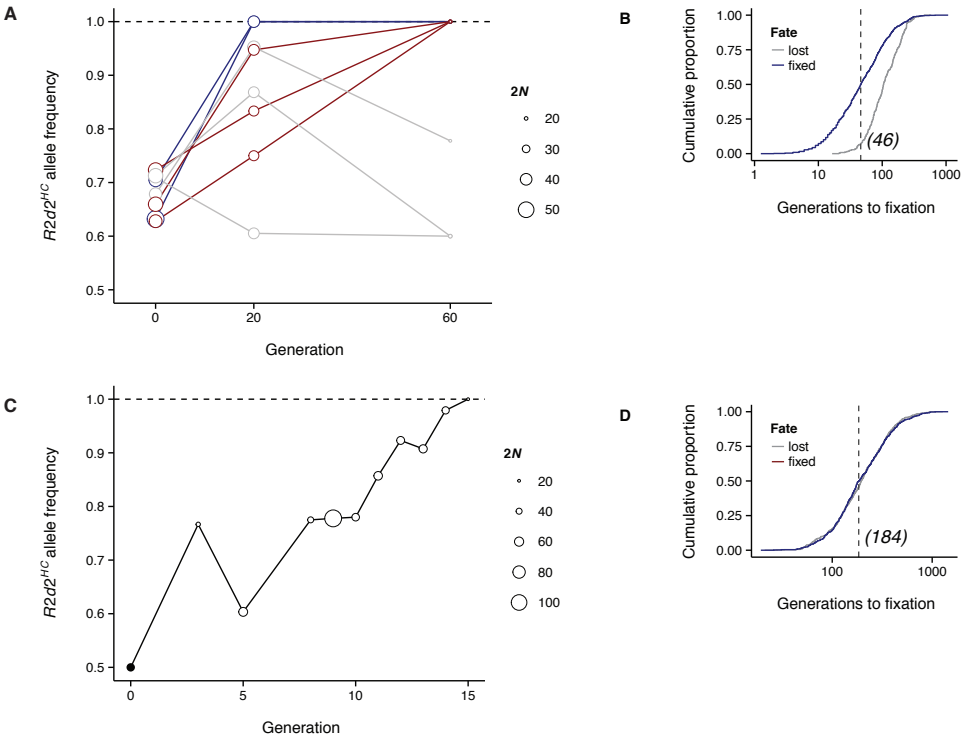

Figure 5.8: *R2d2*^{HC} alleles rapidly increase in frequency in ICR:Hsd-derived laboratory populations. **(A)** *R2d2*^{HC} allele frequency during breeding of 4 HR selection lines and 4 control lines. Trajectories are colored by their fate: blue, *R2d2*^{HC} fixed by generation 20; red, *R2d2*^{HC} fixed by generation 60; grey, *R2d2*^{HC} not fixed. Circle sizes reflect number of chromosomes ($2N$) genotyped. **(B)** Cumulative distribution of time to fixation (blue) or loss (grey) of the focal allele in 1,000 simulations of an intercross line mimicking the HR breeding scheme. Dotted line indicates median fixation time. **(C)** *R2d2*^{HC} allele frequency during breeding of an (HR8×C57BL/6J) advanced intercross line (AIL). Circle sizes reflect number of chromosomes ($2N$) genotyped. **(D)** Cumulative distribution of time to fixation (blue) or loss (grey) of the focal allele in 1,000 simulations of an advanced intercross line mimicking the AIL. Dotted line indicates median fixation time.

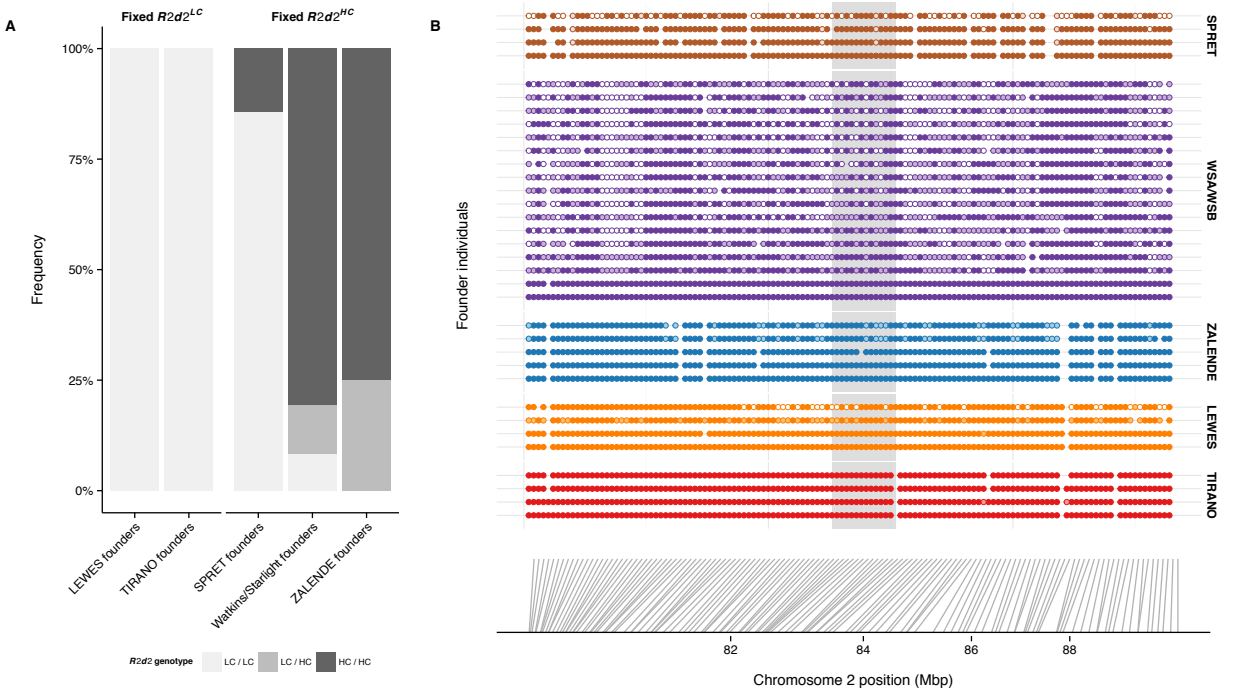

Figure 5.9: Multiple wild-derived inbred lines have fixed $R2d2^{HC}$ alleles that were segregating in founder populations. **(A)** *R2d2* genotype frequencies in available ancestors of wild-derived inbred lines, determined by qPCR (see **§ 5.5**). **(B)** Genotypes at markers on the MegaMUGA array in the region chromosome 2: 80 Mb – 90 Mb for founder individuals of the SPRE/EiJ (brown), ZALENDE/EiJ (blue), LEWES/EiJ (orange) or TIRANO/EiJ (red) inbred lines. For WSB/EiJ (purple), genotypes are from present-day wild individuals from the township of Centreville, Maryland. Genotypes are coded by identity-by-state (IBS) to the respective inbred line: dark circles, homozygous for allele fixed in inbred line; light circles, heterozygous; open circles, homozygous for alternative allele. Region containing *R2d2* indicated by grey shaded box. This panel demonstrates that $R2d2^{HC}$ was most likely not yet fixed in the early breeding generations of these lines.

and non-uniform; our allele frequency estimates may differ substantially from the true population allele frequencies at *R2d2*. Second, the reduction in litter size associated with *R2d2*^{HC} may have a greater impact on *R2d2* allele frequency in a natural population than in the controlled laboratory populations we studied. In our breeding schemes each mating pair contributes the same number of offspring to the next generation so that most fitness differences are effectively erased. Third, *R2d2*^{HC} alleles may be unstable and lose the ability to drive upon reverting to low copy number, as discussed in **Chapter 4**.

Fourth — and perhaps most importantly — meiotic drive at *R2d2* depends on at least two unlinked modifier loci whose effect sizes and allele frequencies are unknown.

5.3.2 Population dynamics of meiotic drive

In an infinitely-large and randomly-mating population, the dynamics of an underdominant meiotic drive allele are only dependent on the relationship between the degree of transmission distortion (m) and the strength of selection against heterozygotes (s)[362] [2]. This relationship can be expressed by the quantity q (see **??**), for which $q > 1$ implies eventual fixation of the driving allele, $q < 1$ implies that the allele will be purged, and $q \approx 1$ leads to maintenance of the allele at an (unstable) equilibrium frequency[362]. The fate of the driving allele in a finite population additionally depends on the population size — the smaller the population, the greater the likelihood that genetic drift will fix a mutation with $q < 1$ by chance (**Figure 5.10A-B**). We note that *R2d2*^{HC} appears to exist close to the $q \approx 1$ boundary ($s \approx 0.2$, $m \approx 0.7$, and thus $q \approx 0.96$).

The simplified model described in above ignores the fact that the action of *R2d2*^{HC} is dependent on unlinked modifier loci. Because the number and effect size of these modifiers is unknown, it is difficult to predict their influence on the fate of *R2d2*^{HC} alleles in the wild. To gain some qualitative insight on the problem, we used forward-in-time simulations to explore the effect of a single unlinked modifier locus on fixation probability of a driving allele. Under an additive model of drive ($m = 0.80$ for modifier genotype *AA*, 0.65 for genotype *Aa* and 0.50 for genotype *aa*), fixation probability is reduced and time to fixation is increased by the presence of the modifier locus (**Figure 5.10C-D**). As the modifier allele becomes more rare, fixation probability approaches

[2] While this is not the standard interpretation of the selection coefficient s, we chose it to be consistent with the notation of Hedrick Hedrick[362]

the neutral expectation ($\frac{1}{2N}$, where N is population size). Importantly, the driving allele tends to sweep until the modifier allele is lost, and then drifts either to fixation or loss (**Figure 5.10E**). Drift at modifier loci thus creates a situation akin to selection in a varying environment — one outcome of which is balancing selection[78]. This is consistent with the maintenance of *R2d2*^{HC} at intermediate frequencies in multiple populations separated by space and time, as we observe in wild mice.

5.4 Conclusions and future directions

Most analyses of positive selection in the literature assume that the likelihood of a newly arising mutation becoming established, increasing in frequency and even going to fixation within a population is positively correlated with its effect on organismal fitness. Here, we have shown that a selfish genetic element has repeatedly driven sweeps in which the change in allele frequency and the effect on organismal fitness are decoupled. Our results suggest that evolutionary studies should employ independent evidence to determine whether loci implicated as drivers of selective sweeps are adaptive or selfish.

Although a selfish sweep has clear implications for such experimental populations as the DO and the Collaborative Cross, the larger evolutionary implications of selfish sweeps are less obvious. On the one hand, sweeps may be relatively rare, as appears to be the case for classic selective sweeps in recent human history[363]. On the other hand, theory and comparative studies indicate that selfish genetic elements may be a potent force during speciation[339,362,342,340,246]. With the growing appreciation for the potential importance of non-Mendelian genetics in evolution and the increasing tractability of population-scale genetic analyses, we anticipate that the effects of selfish elements such as *R2d2* in natural populations, including their contributions to events of positive selection, will soon be elucidated.

5.5 Materials and methods

5.5.1 Mice

Wild *M. m. domesticus* were trapped at a large number of sites across Europe and the Americas (**5.1A**, upper panel) for a total sample size of 471. A set of 29 additional *M. m. castaneus* mice trapped in northern India and Taiwan (**Figure 5.1A**, lower panel) were included as an outgroup[157]. Trapping was carried out in concordance with local laws and either did not require approval or was carried out with the approval of the relevant regulatory bodies (depending on the locality and

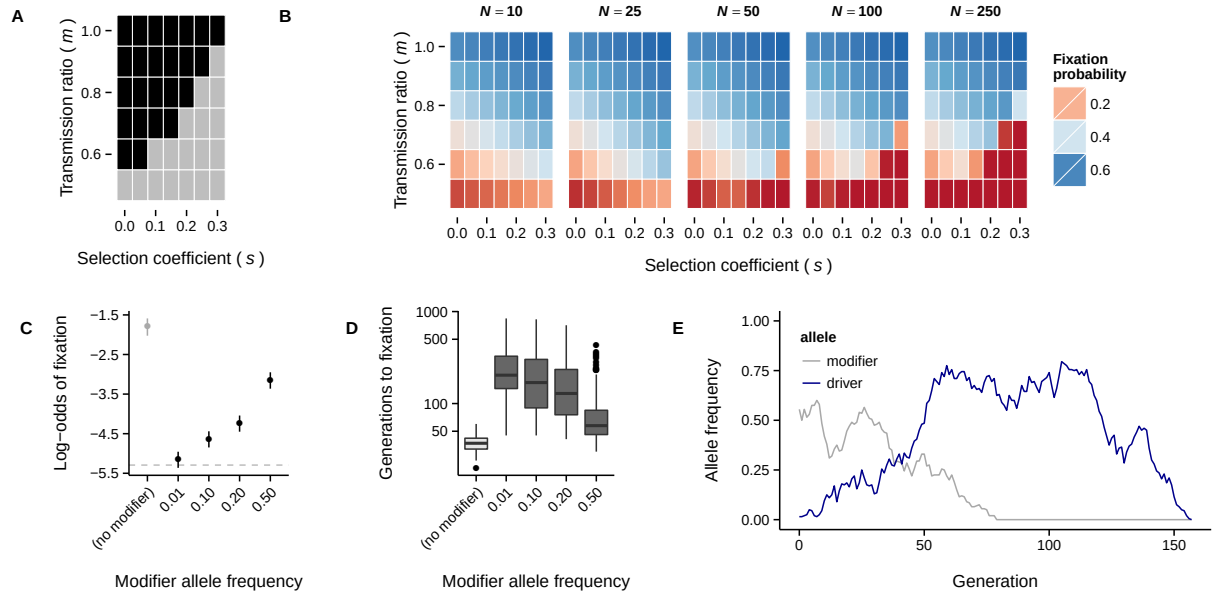

Figure 5.10: Population dynamics of a meiotic drive allele. (**A**) Phase diagram for a meiotic drive system like *R2d2* with respect to transmission ratio (m) and selection coefficient against the heterozygote (s). Regions of the parameter space for which there is directional selection for the driving allele are shown in black; regions in which there are unstable equilibria or directional selection against the driving allele are shown in grey. (**B**) Probability of fixing the driving allele as a function of m, s and population size (N). Notice that, in the area corresponding to the grey region of panel A, fixation probability declines rapidly as population size increases. (**C**) Probability of fixing the driving allele in simulations of meiotic drive dependent on no modifier (light gray) or a single modifier locus (dark gray) with varying allele frequency; $N = 100$, $s = 0.2$, maximum $m = 0.8$, initial driver frequency = $1/2N$. Estimates are given ± 2 SE. Grey dashed line corresponds to fixation probability for a neutral allele ($1/2N$). (**D**) Time to fixation of the driving allele. Values represent 100 fixation events in each condition. (**E**) Example allele-frequency trajectories from a "collapsed" selfish sweep. Although the modifier allele is present at intermediate frequency, the driving allele sweeps to a frequency of approximately 0.75. After the modifier allele is lost, the driver drifts out of the population as well.

institution).

All DO mice were bred at The Jackson Laboratory. Litter sizes were counted within 24 hours of birth. Individual investigators purchased mice for unrelated studies and contributed either tissue samples or genotype data to this study.

"High-running" (HR) selection lines and related advanced intercrosses were developed as previously described[290,353,354,364]. Mouse tails were archived from 3 generations of the HR selection lines (-2, $+22$, and $+61$) and from every generation of the HR8$\times$C57BL/6J advanced intercross.

Progenitors of wild-derived strains have various origins (see below), and were sent to Eva M. Eicher at The Jackson Laboratory for inbreeding in the early 1980s. Frozen tissues from animals in the founder populations were maintained at The Jackson Laboratory by Muriel Davidson until 2014, when they were transferred to the Pardo-Manuel de Villena laboratory at the University of North Carolina at Chapel Hill.

All laboratory mice were handled in accordance with the IACUC protocols of the investigators' respective institutions.

5.5.2 Progenitors of wild-derived inbred lines

Details of the origins of wild-derived inbred strains are taken from[155]. Founder mice for the strain Watkins Star Lines A and B (WSA and WSB, respectively) were trapped near the town of Centreville, Maryland by Michael Potter (working at the National Cancer Institute) in 1976. WSA and WSB were selected for dark agouti coat color with white head blaze. In 1986 breeders were sent to Eva M. Eicher at The Jackson Laboratory, where the lines have been maintained since as WSA/EiJ and WSB/EiJ. The LEWES/EiJ strain is descended from wild mice trapped by Potter near Lewes, Delaware in 1981. Breeders were sent to Eicher at the Jackson Laboratory in 1995, where the line has been maintained since. The ZALENDE/EiJ and TIRANO/EiJ inbred strains are descended from mice trapped by Richard D. Sage near the villages of Zalende, Switzerland and Tirano, Italy respectively, in the vicinity of the Poschiavo Valley at the Swiss-Italian border. Mice from SageÕs colony were transferred to Potter in 1981. A single breeding pair for each strain was transferred to Eicher at The Jackson Laboratory in 1982. The SPRET/EiJ inbred strain was derived from wild Mus spretus mice trapped near Puerto Real, Cadiz province, Spain by Sage in 1978. The Jackson LaboratoryÕs colony was initiated by Eicher from breeders transferred via Potter in 1983.

5.5.3 Microarray genotyping

Whole-genomic DNA was isolated from tail, liver, muscle or spleen using Qiagen Gentra Puregene or DNeasy Blood & Tissue kits according to the manufacturer's instructions. All genome-wide genotyping was performed using the Mouse Universal Genotyping Array (MUGA) and its successor, MegaMUGA (GeneSeek, Lincoln, NE)[95]. Genotype quality control was performed as described in **Chapter 4**.

5.5.4 PCR genotyping

Primers were design to amplify two regions predicted to be in close linkage ($< 0.1cM$) to *R2d2*. *Primer Set A* targets a 318 bp region with two distinct haplotypes in linkage with either the *R2d2*^{LC} allele or the *R2d2*^{HC} allele: 5′-`CCAGCAGTGATGAGTTGCCATCTTG`-3′ (forward) and 5′-`TGTCACCAAGGTTTTCTTCCAAAGGGAA`-3′ (reverse). *Primer Set B* amplifies a 518 bp region; the amplicon is predicted, based on whole-genome sequencing, to contain a 169 bp deletion in HR8 relative to the C57BL/6J reference genome: 5′-`GAGATTTGGATTTGCCATCAA`-3′ (forward) and 5′-`GGTCTACAAGGACTAGAAACAG`-3′ (reverse). Primers were designed using IDT PrimerQuest (`https://www.idtdna.com/Primerquest/Home/Index`).

Crude whole-genomic DNA for PCR reactions was extracted from mouse tails. The tissues were heated in 100 μl of 25 mM NaOH + 0.2 mM EDTA at 95°C for 60 minutes followed by the addition of 100 μl of 40 mM Tris-HCl. The mixture was then centrifuged at $2000 \times g$ for 10 minutes and the supernatant used as PCR template. PCR reactions were performed in a 10 μl volume and contained 0.25 mM dNTPs, 0.3 mM of each primer, and 0.5 units of GoTaq polymerase (Promega). Cycling conditions were 95°C, $2-5$ min; 35 cycles at 95°C, 55°C and 72°C for 30 sec each; with a final extension at 72°C for 7 min.

For *Primer Set A*, products were sequenced at the University of North Carolina Genome Analysis Facility on an Applied Biosystems 3730XL Genetic Analyzer. Chromatograms were analyzed with the Sequencher software package (Gene Codes Corporation, Ann Arbor, Michigan, United States). For *Primer Set B*, products were visualized and scored on 2% agarose gels. Assignment to haplotypes was validated by comparing the results to qPCR assays for the single protein-coding gene within *R2d2*, *Cwc22* (see below). For generation $+61$, haplotypes were assigned based on MegaMUGA genotypes and validated by the normalized per-base read depth from whole-genome sequencing (see below), calculated with `samtools mpileup`[365]. The concordance between qPCR, read depth,

and haplotypes assigned by MegaMUGA or Sanger sequencing is shown in **5.11**.

HR selection lines were genotyped at three generations, one before (-2) and two during ($+22$, $+61$) artificial selection was initiated. We genotyped 185 randomly selected individuals from generation -2 and 157 individuals from generation $+22$ using *Primer Set A*. An additional 80 individuals from generation $+61$ were genotyped with the MegaMUGA array, as noted above. The HR8×C57BL/6J advanced intercross line was genotyped with *Primer Set B* in tissues from breeding stock at generations 3, 5, 8, 9, 10, 11, 12, 13, 14 and 15.

5.5.5 Copy-number assays and assignment of *R2d2* status

Copy-number at *R2d2* was determined by qPCR for *Cwc22* as previously described (**Chapter 4**). Estimation of integer diploid copy numbers ≥ 3 by qPCR is infeasible without many technical and biological replicates, especially in the heterozygous state. We took advantage of *R2d2* diploid copy-number estimates from whole-genome sequencing for the inbred strains C57BL/6J (0), CAST/EiJ (2) and WSB/EiJ (66), and the (WSB/EiJ×C57BL/6J)F_1 (33) to establish a threshold for declaring a sample "high-copy." For each of the two target-reference pairs we calculated the sample mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of the normalized ΔC_t among CAST/EiJ controls and wild *M. m. castaneus* individuals together. We designated as "high-copy" any individual with normalized C_t greater than $\hat{\mu} + 2\hat{\sigma}$, *i.e.* any individual with approximately $> 95\%$ probability of having diploid copy number > 2 at *R2d2*. Individuals with high copy number and evidence of local heterozygosity (a heterozygous call at any of the 13 markers in the vicinity of *R2d2*) were declared heterozygous *R2d2*$^{HC/LC}$, and those with high copy number and no heterozygous calls in the candidate interval were declared homozygous *R2d2*$^{HC/HC}$.

5.5.6 Exploration of population structure in wild mice

Scans for signatures of positive selection based on patterns of haplotype-sharing assume that individuals are unrelated. We identified pairs of related individuals using the *IBS2** ratio[366], defined as *HETHET / (HOMHOM + HETHET)*, where *HETHET* and *HOMHOM* are the count of non-missing markers for which both individuals are heterozygous (share two alleles) and homozygous for opposite alleles (share zero alleles), respectively. Pairs with *IBS2** < 0.75 were considered unrelated. Among individuals who were a member of one or more unrelated pairs, we iteratively removed one sample at a time until no related pairs remained, and additionally excluded markers with minor-allele frequency < 0.05 or missingness > 0.10. The resulting dataset

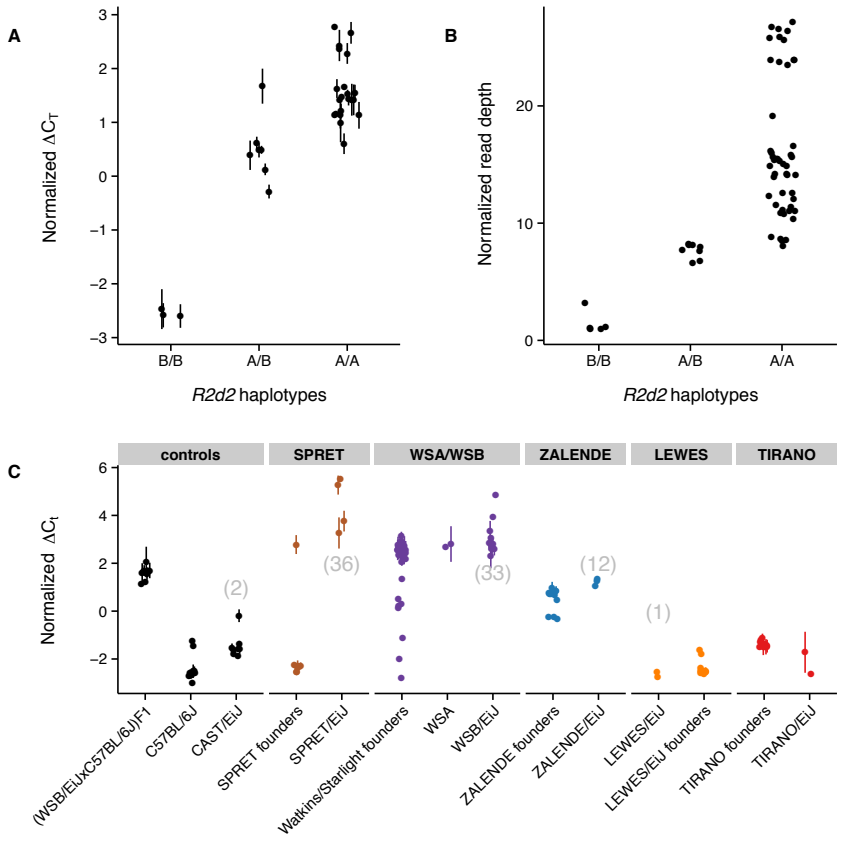

Figure 5.11: Characterization of Cwc22 qPCR assays. **(A)** Concordance between local haplotype and qPCR in HR lines. Normalized ΔC_t from qPCR assay against *Cwc22* versus local haplotype at chr2: 83 Mb (A = $R2d2^{LC}$, B = $R2d2^{HC}$) in HR generation $+61$ individuals. Error bars represent mean ± 1 SD over technical replicates, when present. **(B)** Normalized read depth at *R2d2* in whole-genome sequencing versus local haplotype. **(C)** *R2d2* copy number of wild-derived inbred mouse lines and available ancestors, estimated by qPCR. Samples listed as "control" are included as internal calibration points. For inbred strains that have been sequenced (CAST/EiJ, SPRET/EiJ, WSB/EiJ, ZALENDE/EiJ, LEWES/EiJ) copy numbers estimated from depth of coverage are indicated in parentheses.

contained genotypes for 396 mice at 58,283 markers.

Several of our analyses required that samples be assigned to populations. Because mice in the wild breed in localized demes and disperse only over short distances (on the order of hundreds of meters)[367], it seems reasonable to delineate populations on the basis of geography. We assigned samples to populations based on the country in which they were trapped. To confirm that these population labels correspond to genetically-differentiated clusters we performed two exploratory analyses of population structure. First, classical multidimensional scaling (MDS) of autosomal genotypes was performed with `PLINK` (options `--mdsplot --autosome`)[368]. The result is presented in **Figure 5.1B-C**, in which samples are colored by population. Second, we used `TreeMix`[369] to generate a population tree allowing for gene flow using the set of unrelated individuals. Autosomal markers were first pruned to reach a set in approximate linkage equilibrium (`plink --indep 25 1`). `TreeMix` was run on the resulting set using the *M. m. castaneus* samples as an outgroup and allowing up to 10 gene-flow edges (`treemix -root "cas" -k 10`) (**Figure 5.1D**). The clustering of samples by population evident by MDS and the absence of long-branch attraction in the population tree together indicate that our choices of population labels are biologically reasonable.

5.5.7 Scans for selection in wild mice

Two complementary statistics, hapF_{LK}[358] and allele-frequency-standardized iHS score[359], were used to examine wild mouse genotypes for signatures of selection surrounding *R2d2*. The hapF_{LK} statistic is a test of differentiation of local haplotype frequencies between hierarchically-structured populations. It can be interpreted as a generalization of Wright's F_{st} to a graph and additionally exploits local LD. Its model for haplotypes is that of `fastPHASE`[370] and requires a user-specified value for the parameter K, the number of local haplotype clusters. We computed hapF_{LK} in the set of unrelated individuals, using *M. m. castaneus* samples as an outgroup, for $K \in \{4, 8, 12, 16, 20, 24, 28, 32\}$ (`hapflk --outgroup "cas" -k` K) and default settings otherwise.

The iHS score (and its allele-frequency-standardized form) is a measure of extended homozygosity on a derived haplotype relative to an ancestral one. It requires phased genotypes. For consistency with the hapF_{LK} analysis, we used `fastPHASE` on the same genotypes over the same range of K with 10 random starts and 25 iterations of expectation-maximization (`fastphase -K`K `-T10 -C25`) to obtain phased genotypes. We then used `selscan`[371] to compute raw iHS scores (`selscan --ihs`) and standardized the scores in 25 equally-sized frequency bins

(`selscan-norm -bins 25`).

Values in the upper tail of the genome-wide distribution of hapF_{LK} or iHS represent candidates for regions under selection. We used percentile ranks directly and did not attempt to calculate approximate or empirical p-values.

5.5.8 Detection of identity-by-descent in wild mice

As an alternative test for selection, we computed density of IBD-sharing using the `RefinedIBD` algorithm of `BEAGLE` v4.0-r1399[372], applying it to the full set of 500 individuals. The haplotype model implemented in `BEAGLE` uses a tuning parameter (the "scale" parameter) to control model complexity — larger values enforce a model with fewer local haplotype clusters, increasing sensitivity for detecting IBD and decreasing computational cost at the expense of accuracy. The authors recommend a value of 2.0 for human data. We increased the scale parameter to 5.0 to increase detection power given our much sparser marker set and the relatively weaker local LD in mouse versus human populations[373]. We trimmed one marker from the ends of candidate IBD segments to reduce edge effects (`java -jar beagle.jar ibd=true ibdscale=5 ibdtrim=1`). We retained those IBD segments shared between individuals in the set of 396 unrelated mice. In order to limit noise from false-positive IBD segments, we further removed segments with LOD score < 5.0 or width < 0.5 cM.

An empirical IBD-sharing score was computed in 500 kb bins (with 250 kb overlap) as:

$$f_n = \frac{\sum_n s_{ij} p_{ij}}{w_{ij}}$$

where the sum in the numerator is taken over all IBD segments overlapping bin n and s_{ij} is an indicator variable which takes the value 1 if individuals i, j share a haplotype IBD in bin n and 0 otherwise. The weighting factor w_{ij} is defined as

$$w_{ij} = 0.001 \times \left(\frac{n_a n_b}{W}\right)^{\frac{1}{2}}$$

with

$$W = \max(n_a, n_b)$$

where n_a and n_b are the number of unrelated individuals in the population to which individuals i and j belong, respectively. This weighting scheme accounts for the fact that we oversample some geographic regions (for instance, Portugal and Maryland) relative to others. To explore differences in haplotype-sharing within versus between populations, we introduce an additional indicator p_{ij}. Within-population sharing is computed by setting $p_{ij} = 1$ if individuals i, j are drawn from the same population and $p_{ij} = 0$ otherwise. Between-population sharing is computed by reversing the values of p_{ij}. The result is displayed in **Figure 5.2**.

5.5.9 Analysis of local sequence diversity in whole-genome sequence

Whole-genome sequence data for wild mice was obtained and aligned as previously described (**Chapter 4**). Single-nucleotide variants (SNVs) relative to the reference sequence of chromosome 2 were called using `samtools mpileup` v0.1.19-44428cd with maximum per-sample depth of 200. Genotype calls with root-mean-square mapping quality < 30 or genotype quality < 20 were treated as missing. Sites were used for phasing if they had a minor-allele count > 2 and at most 2 missing calls. Phasing and imputation were was performed with `BEAGLE`, using 20 iterations for phasing and default settings otherwise (`java -jar beagle.jar phasing-its=20`). Sites were assigned a genetic position by linear interpolation the genetic map described in **Chapter 3**[247]. We note that, unlike for humans, a large panel of reference haplotypes does not exist for mice. Using sample haplotypes as templates for phasing results in higher rates of switching errors, especially when the sample size is small. Switching errors introduce bias towards the null hypothesis in EHH- and iHS-type tests, which compare the length of haplotypes linked to the derived versus the ancestral allele at a specific locus[359].

The *R2d2* critical interval spans positions 83,790,939 – 84,701,151 in the mm10 reference sequence. We used as the *R2d2*^{HC} index variant the site with strongest nominal association with *R2d2* copy number and located within 1 kb of the proximal boundary of the candidate interval. That variant is chr2:83,790,275`T`→`C`. The `C` allele is associated with high copy number and is therefore presumed to be the derived allele. We computed the extended haplotype homozygosity (EHH) statistic[357] in the phased dataset over a 1 Mb window on each side of the index site using `selscan` (`selscan --ehh --ehh-win 1000000`). The result is presented in **Figure 5.2B**. Decay of haplotypes away from the index variant was visualized as a bifurcation diagram (**Figure 5.2C**) using code adapted from the `R` package `rehh` (`https://cran.r-project.org/package=rehh`).

5.5.10 Estimation of age of *R2d2*^{HC} alleles

To obtain a lower bound for the age of *R2d2*^{HC} and its associated haplotype, we used the method from Stephens *et al.*[374]. Briefly, this method approximates the probability P that a haplotype is broken by recombination or mutation during the G generations since its origin as

$$P = e^{-G(-\mu+r)}$$

where μ and r are the per-generation rates of mutation and recombination, respectively. Assuming $\mu \ll r$ and, taking P' (the observed number of intact haplotypes) in a sample, as an estimator of P, obtain the following expression for G:

$$G \approx -\left(\log P'\right)/r$$

We enumerated haplotypes in our sample of 52 chromosomes at 3 SNPs spanning the *R2d2* critical interval. The most proximal SNP is the index variant for the EHH analyses (chr2:83,790,275$\mathtt{T}\rightarrow\mathtt{C}$); the most distal SNP is the SNP most associated with copy number within 1 kbp of the boundary of the candidate interval (chr2:84,668,280$\mathtt{T}\rightarrow\mathtt{C}$); and the middle SNP was randomly-chosen to fall approximately halfway between (chr2:84,079,970$\mathtt{C}\rightarrow\mathtt{T}$). The three SNPs span genetic distance 0.154 cM (corresponding to $r \approx 0.00154$). The most common haplotype among samples with high copy number was assumed to be the non-recombined haplotype. Among 52 chromosomes, 22 carried at least part of the *R2d2*^{HC}-associated haplotype; of those, 11 were ancestral and 11 recombinant. This gives an estimated age of 450 generations for *R2d2*^{HC}.

We note that the approximations underlying this model assume constant population size and the absence of selection. To the extent that haplotype homozygosity decays more slowly on a positively- (or selfishly-) selected haplotype, we will underestimate the true age of *R2d2*^{HC}. The matter is further complicated by the assumption that the recombination rate per meiosis (although not the population-scaled rate) is not genotype-dependent — which is clearly not the case for *R2d2* (**Chapter 4**).

5.5.11 Analyses of fitness effects of *R2d2*^{HC} in the DO

To assess the consequences of *R2d2*^{HC} for organismal fitness, we treated litter size as a proxy for absolute fitness. Using breeding data from 475 females from DO generations 13, 16, 18 and 19,

we estimated mean litter size in four genotype groups: *R2d2*$^{LC/LC}$ homozygous females; *R2d2*$^{HC/LC}$ heterozygous females with transmission ratio distortion (TRD) in favor of the *R2d2*^{HC} allele; *R2d2*$^{HC/LC}$ heterozygous females without TRD; and *R2d2*$^{HC/HC}$ homozygous females. Group means were estimated using a linear mixed model with parity and genotype as fixed effects and a random effect for each female using the `lme4` package for `R`. Confidence intervals were obtained by likelihood profiling and post-hoc comparisons were performed via *F*-tests, using the Kenward-Roger approximation for the effective degrees of freedom. The mean number of *R2d2*^{HC} alleles transmitted per litter by heterozygous females with and without TRD was estimated using a weighted linear model, with the total number of offspring per female as weights.

5.5.12 Whole-genome sequencing of HR selection lines

Ten individuals from generation $+61$ of each of the eight HR selection lines were subject to whole-genome sequencing. Briefly, high-molecular-weight genomic DNA was extracted using a standard phenol/chloroform procedure. Illumina TruSeq libraries were constructed using 0.5 μg starting material, with fragment sizes between 300 and 500 bp. Each library was sequenced on one lane of an Illumina HiSeq2000 flowcell in a single 2×100 bp paired-end run.

5.5.13 Null simulations of closed breeding populations

Widespread fixation of alleles due to drift is expected in small, closed populations such as the HR lines or the HR8xC57BL/6J advanced intercross line. But even in these scenarios, an allele under positive selection is expected to fix (1) more often than expected by drift alone in repeated breeding experiments using the same genetic backgrounds and (2) more rapidly than expected by drift alone. We used the `R` package `simcross` (`https://github.com/kbroman/simcross`) to obtain the null distribution of fixation times and fixation probabilities for an HR line under Mendelian transmission.

We assume that the artificial selection applied for voluntary exercise in the HR lines (described in[290]) was independent of *R2d2* genotype. This assumption is justified for two reasons. First, 3 of 4 selection lines and 2 of 4 control (unselected) lines fixed *R2d2*^{HC}. Second, at generations 4 and 10 of the HR8$\times$C57BL/6J advanced intercross, no quantitative trait loci (QTL) associated with the selection criteria (total distance run on days 5 and 6 of a 6-day trial) were found on chromosome 2. QTL for peak and average running speed were identified at positions linked to *R2d2*; however, HR8 alleles at those QTL were associated with decreased, not increased, running speed[354,364].

Without artificial selection, an HR line reduces to an advanced intercross line maintained by avoidance of sibling mating. We therefore simulated 100 replicates of an advanced intercross with 10 breeding pairs and initial focal allele frequency of 0.75. Trajectories were followed until the focal allele was fixed or lost. As a validation, we confirmed that the focal allele was fixed in 754 of 1000 runs, which is not different from the expected 750 ($p = 0.62$, binomial test). Simulated trajectories and the distribution of sojourn times are presented in **Figure 5.8**.

The HR8×C57BL/6J advanced intercross line was simulated as a standard biparental AIL with initial focal allele frequency of 0.5. Again, 1000 replicates of an AIL with 20 breeding pairs were simulated and trajectories were followed until the focal allele was fixed or lost. The result is presented in **Figure 5.8**.

5.5.14 Investigation of population dynamics of meiotic drive

We used two approaches to investigate the population dynamics of a female-limited meiotic drive system with selection against the heterozygote. First, we evaluated the fixation probability of a driving allele in relationship to transmission ratio (m), selection coefficient against the heterozygote (s) and population size (N) by modeling the population as a discrete-time Markov chain whose states are possible counts of the driving allele. Following[362]:

$$p_{t+1} = \frac{(1-s)(1+2m)p_t(1-p_t) + 2(1-p_t)^2}{2[1-2sp_t(1-p_t)]}$$

where p_{t+1} is the expected frequency of the driving allele in generation $t+1$ given its frequency in the previous generation (p_t). In an infinite population, the equilibrium behavior of the system is governed by the quantity q:

$$q = \frac{1}{2}(1-s)(1+2m)$$

When $q > 1$, the driving allele always increases in frequency. For values of $q \approx 1$ and smaller, the driving allele is either lost or reaches an unstable equilibrium frequency determined m and s.

Let M be the matrix of transition probabilities for the Markov chain with $2N+1$ states corresponding to possible counts of the driving allele in the population $(0, \ldots, 2N)$. The entries m_{ij} of M are

$$m_{ij} = \binom{2N}{i} (1 - p_{t+1})^{2N-i} (p_{t+1})^{i}$$

Given a vector p_0 of starting probabilities, the probability distribution at generation t is obtained by iteration:

$$p_t = p_0 M^t$$

We initiated the chain with a single copy of the driving allele. Since this Markov chain has absorbing states (namely allele counts 0 and $2N$), we approximated steady-state probabilities by iterating the chain until the change in probabilities between successive generations was $< 10^{-4}$. Fixation probability is given by the value of the entry $p_t[2N]$ at convergence. We evaluated all possible combinations of $0.5 \leq m \leq 1.0$ (in steps of 0.1) and $0 \leq s \leq 0.3$ (in steps of 0.05).

To investigate the effects of modifier loci on the frequency trajectory of a driving allele, we implemented in Python forward-in-time simulations under a Wright-Fisher model with selection. Simulations assumed a constant population size of $2N = 200$ chromosomes, each 100 cM long, with balanced sex ratio. At the beginning of each run a driving allele was introduced (at 50 cM) on a single, randomly chosen chromosome. Modifier alleles were introduced into the population independently at a specified frequency, at position 0.5 cM (*i.e.* unlinked to the driving allele). To draw the next generation, an equal number of male and female parents were selected (with replacement) from the previous generation according to their fitness. Among females heterozygous for the driving allele, transmission ratio (m) was calculated according to genotype at the modifier loci (if any). For males and homozygous females, $m = 0.5$. Individuals were assigned a relative fitness of $w = 1$ if $m = 0.5$ and $w = 0.8$ if $m > 0.5$. Recombination was simulated under the Haldane model (*i.e.* a Poisson process along chromosomes with no crossover interference). Finally, for each individual in the next generation, one chromosome was randomly chosen from each parent with probability m.

Simulation runs were restarted when the driving allele was fixed or lost, until 100 fixation events were observed in each condition of interest. Probability of fixation was estimated using the waiting time before each fixation event, assuming a geometric distribution of waiting times, using the `fitdistr()` function in the `R` package `MASS`.

CHAPTER 6

Sequence and structural diversity of mouse Y chromosomes

6.1 Introduction

The sex chromosomes are the only heteromorphic chromosome pair in mammals. In the vast majority of mammal species, one member of the pair — the Y chromosome — is sex-determining. Presence of the Y-encoded protein SRY is sufficient to initiate the male developmental program[6]. Since their divergence from the ancestral X chromosome approximately 180 million years ago (Mya)[5], mammal Y chromosomes have lost nearly all of their ancestral gene content. Although these losses have occurred independently along different lineages within the mammals, the small subset of genes that are retained in each linage tend to be dosage-sensitive and have housekeeping functions in core cellular processes such as transcription and protein degradation[375,376]. In addition, Y chromosomes have acquired — via transposition from autosomes — a small number of genes that are often present in many copies and are highly specialized for function in the male germline[377,306]. Several lines of evidence suggest that, at least in mouse, the evolution of the acquired genes is driven by intragenomic conflict with the X chromosome for transmission to progeny[378,379].

The repetitive content of mammal Y chromosomes makes them difficult to sequence, assemble and annotate accurately even with considerable manual effort. This has hampered efforts to understand Y-linked variation. Because the Y chromosome is passed only through the male germline and is obligately transmitted from fathers to sons without recombination, it provides a rich view into male-specific mutational, selective and demographic processes. We therefore took advantage of a recent high-quality assembly of the mouse Y[306] to perform a systematic survey of a diverse sample of Y chromosomes using published whole-genome sequencing datasets. In this chapter we characterize both sequence and structural variation in *Mus*, and use complementary gene expression data from testis to explore proximate functional consequences of this variation. We find that:

- Sequence diversity of Y chromosomes is $< 10\%$ that on autosomes and the site frequency spectrum is skewed towards low-frequency alleles. These patterns are best explained by a recent population bottleneck.

- Copy number of Y-acquired genes is extremely variable in *Mus*: wild *M. m. domesticus* have, on average, three times as many copies as wild *M. spretus*. X-linked homologs of Y-acquired genes are also variable in copy number, but only one family — *Slx/Slxl1* and *Sly* — have correlated copy number between the sex chromosomes.

- The expresssion pattern of Y-linked genes in the testis is differentiated between Y-chromosome lineages. In hybrids between subspecies, the expression of Y-acquired genes is governed by Y chromosome genotype independent of the X chromosome. The direction of the effect is opposite of copy number.

- Both the population-genetic and functional evidence provide limited support for the hypothesis that intragenomic conflict between the sex chromosomes has a major role in shaping diversity on either the X or the Y chromosome within *Mus musculus*.

In the remainder of this section we introduce important themes in the evolution of sex chromosomes, restricting our attention to mammals unless otherwise stated.

6.1.1 Origins of sex chromosomes

Sex chromosomes have emerged many times in independent plant and animal lineages. Although the morphology and content of extant sex chromosomes is extremely diverse, the evolution of a new sex-chromosome pair generally follows a recognizable pattern (**Figure 6.1**). First, a sex-determining allele arises on an autosome. Theory predicts that sexually-antagonistic and male-advantageous mutations — those conferring advantage to only one sex — should be favored to the extent that they are in linkage disequilibrium with the sex-determining allele. This in turn favors mutations that suppress recombination between the proto-sex chromosomes[380,381]. Once recombination has ceased, the proto-sex chromosomes begin to diverge by independent accumulation of mutations. The sex-limited chromosome (in mammals, the Y) loses most of its functional content and the expression patterns of those genes that remain become specialized for function in the germline of the heterogametic sex (in mammals, the male)[382]. The other chromosome (the X)

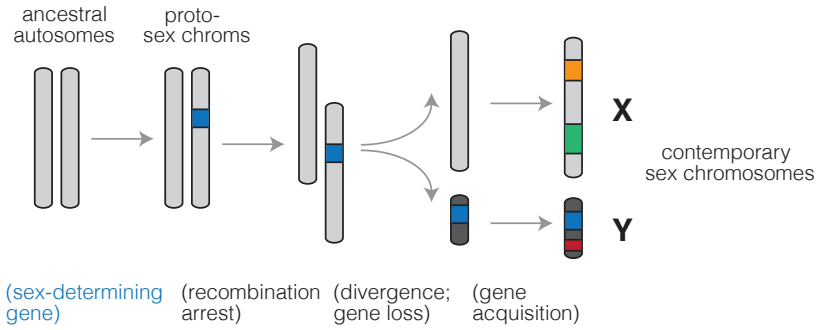

Figure 6.1: Evolution of heteromorphic sex chromosomes.

can still recombine in the homogametic sex (in mammals, the female) and so retains more of its ancestral identity.

Contrary to rather dramatic predictions that the mammalian Y chromosome is bound for extinction[382], empirical studies Y chromosomes have demonstrated that most gene loss occurs in early proto-sex chromosomes, and that older sex chromosomes like those of mammals are more stable[375]. The evolutionary diversity of Y chromosomes in mammals arises from the set of Y-acquired genes, which make up a small fraction of some Y chromosomes and a much larger fraction in others — from 5% in rhesus to 45% in human[5] (**Figure 6.2**).

6.1.2 The mouse Y chromosome

Early molecular studies of the mouse Y chromosome hinted that it consisted of repetitive sequences, with copy number in the hundreds, and that it was evolving rapidly[383,384]. Unlike other mammalian Y chromosomes, which are dominated by large blocks of heterochromatin[5], the mouse Y was also known to be large and almost entirely euchromatic[?]. Spontaneous mutations in laboratory stocks allowed the mapping of male-specific tissue antigens and the sex-determining factor *Sry* to the short arm of the chromosome (Yp)[385], while lesions on the long arm (Yq) were associated with infertility and defects in spermatogenesis[386,387].

Sequencing of the mouse Y in the reference strain C57BL/6J was finally completed in 2014 after more than a decade of painstaking effort[306]. The long arm of the chromosome was shown to consist of approximately 200 copies of a 500 kb unit containing the acquired genes *Sly*, *Ssty1*, *Ssty2* and *Srsy* (**Figure 6.3**). The copies retain $98 - 99.99\%$ mutual sequence identity. Ancestral genes are restricted to the short arm. Analysis of BAC clones three additional strains (AKR/J, CAST/EiJ

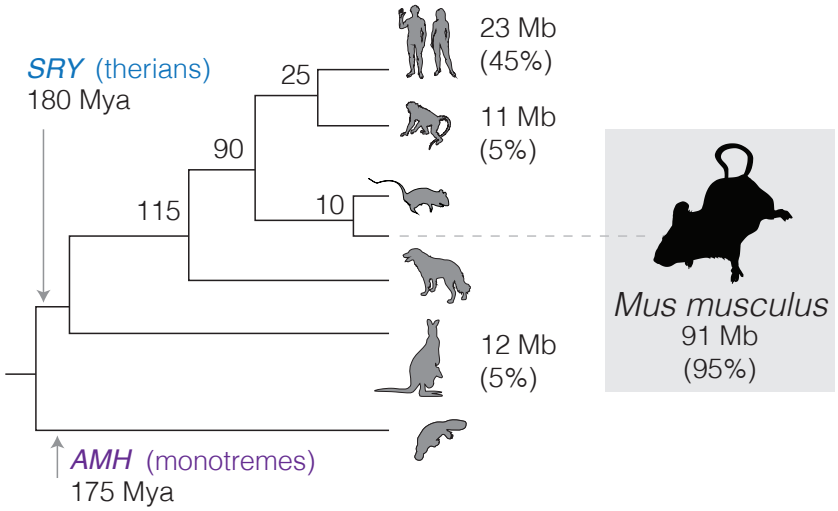

Figure 6.2: Y chromosomes of mammals. The Y chromosome of therian mammals, characterized by the sex-determining factor *SRY*, diverged from the mammal X approximately 180 Mya. (The monotremata have a different sex-determining factor, *AMH*, and an idiosyncratic five-pair sex chromosome system.) Y chromosome sizes and the fraction of sequence occupied by multicopy, Y-acquired genes are shown at the tips of the tree.

and SPRET/EiJ) showed that the other Ys have similar sequence content but possibly different organization and copy number.

6.1.3 Intragenomic conflict between the sex chromosomes

The dramatic co-amplification of genes on the X and Y chromosomes is thought to be a byproduct of competition between the X and Y for transmission to the next generation. The current consensus favors an unidentified X-linked sex-ratio distorter whose action is suppressed by one or more Y-linked factors[378]. Consistent with this hypothesis, the *Sly* and *Slx* families act in opposing directions to maintain or relieve transcriptional silencing of the sex chromosomes during and after meiosis (meiotic sex chromosome inactivation, MSCI)[388,379]. Overexpression of *Sly* (via knockdown of *Sly*) in the testis results in sex ratio distortion in favor of males; the reverse is true for overexertion of *Slx*. Disruption of MSCI is also associated with male sterility in inter-subspecific hybrids between *M. m. domesticus* and *M. m .musculus*[146], and sperm morphology defects map to the Y chromosome in some crosses[260]. Together these observations suggest that the intragenomic conflict between the sex chromosomes in mouse is played out in post-meiotic spermatids and may have mechanistic overlap with hybrid male sterility.

6.2 Results

6.2.1 A catalog of Y-linked sequence variation in mouse

Whole-genome sequence data for 68 male mice was collected from published sources. The final set consisted of 42 wild-caught mice; 20 classical inbred strains; 1 laboratory mouse derived from an outbred stocks; and 5 wild-derived inbred strains (**Table 6.1**). All three cardinal subspecies of *M. musculus* (*domesticus*, *musculus* and *castaneus*) are represented. *Mus spretus* and *Mus spicilegus* served as close outgroups for analyses of the Y chromosome, and a female *Mus caroli* individual was used as a more distant outgroup in analyses of the mitochondrial genome. We restricted our attention to 1.6 Mb of sequence on the short arm of the Y accessible to alignment of short reads.

SNVs and small indels were ascertained in jointly in all samples and assigned ancestral or derived status based on the consensus call among the *M. spretus* samples. We identified $27,715$ high-confidence SNVs (transitions:transversions $= 1.72$) and $3,009$ high-confidence indels segregating in *M. musculus* after applying stringent filters for genotype quality (see **§ 6.5**). Of these 286 (0.9%) fall in protein-coding genes, and only 161 are predicted to impact protein function.

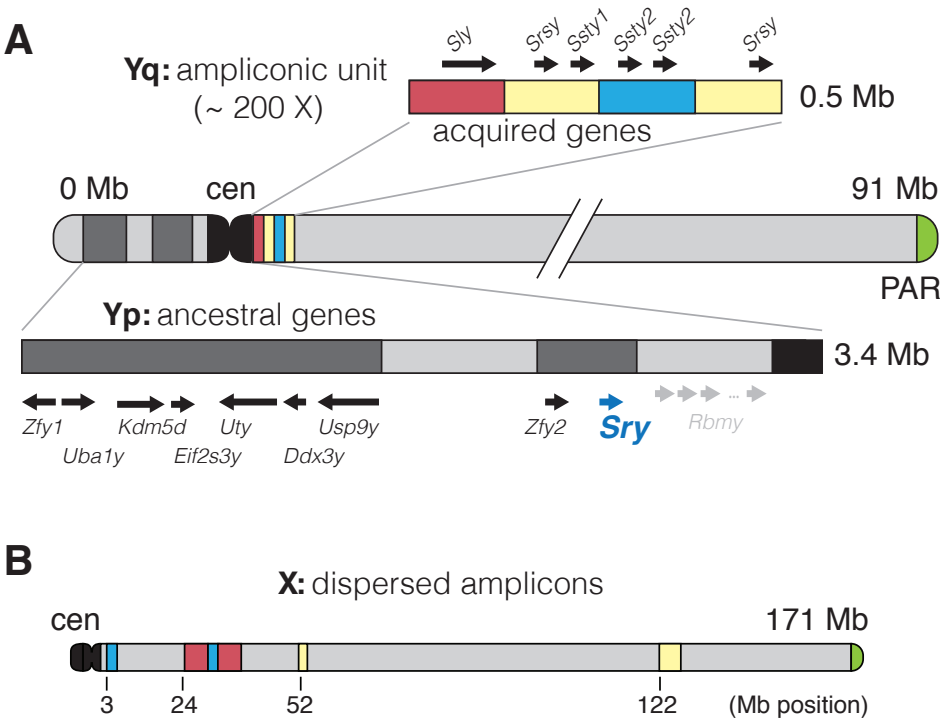

Figure 6.3: Structure of the mouse X and Y chromosomes in the C57BL/6J reference strain. (**A**) The short arm of the Y (Yq) consists primarily of genes shared the X and retained since the X and Y diverged from the ancestral autosome pair. These genes are interspersed with blocks of segmental duplications (light grey). The sex-determining factor *Sry* is encoded on the short arm. The long arm (Yq) consists of approximately 200 copies of a 500 kb repeating unit containing the acquired genes *Sly*, *Ssty1*, *Ssty2* and *Srsy*. The sequence in the repeat unit can be roughly divided into three families "red," "yellow" and "blue" following[306]. (**B**) The X choromosome, unlike the Y, is acrocentric. Homologs of the acquired genes from the Y (*Slx*, *Slxl1*, *Sstx* and *Srsx*; shown above using colored blocks as on the Y) are present in high copy number but are arranged in tandem chunks, rather than intermingled as on the Y.

Type	Population	Locality	N	
wild	*M. m. domesticus*	DE	8	
		FR	8	
		IR	8	
	M. m. musculus	CZ	2	
		KZ	3	
		AF	5	
	M. m. castaneus	IN	3	
	M. spretus	ES	4	
	M. spicilegus	HU	1	
wild-derived	*M. m. domesticus*	CH	1	ZALENDE/EiJ
		US	1	LEWES/EiJ
	M. m. musculus	CZ	1	PWK/PhJ
	M. m. castaneus	TH	1	CAST/EiJ
	M. spretus	ES	1	SPRET/EiJ
lab	-	-	21	

Table 6.1: Wild and laboratory mice used for Y chromosome analyses.

One group of inbred strains in our dataset — C57BL/6J, C57BL/10J, C57L/J and C57BR/cdJ — have a known common ancestor in the year 1929. We used this fact to obtain a direct estimate of the male-specific point mutation rate: $5.4 \times 10^{-9} - 8.1 \times 10^{-9}$ bp^{-1} generation^{-1}, assuming an average of three generations per year. This interval just contains the sex-averaged autosomal rate of 5.4×10^{-9} bp^{-1} generation^{-1} recently estimated from whole-genome sequencing of mutation-accumulation lines[389]. Using the ratio between paternal to maternal mutations in mouse (2.78) estimated in the classic studies from Russell and colleagues[10], we obtain a male-specific autosomal rate of 7.9×10^{-9} bp^{-1} generation^{-1}, in good agreement with our estimate from the Y chromosome.

6.2.2 Phylogeography of Y chromosomes

A phylogenetic tree for the Y chromosome and mitochondrial genome were constructed with `BEAST` (**Figure 6.4**). The approximate time to most recent common ancestor (MRCA) of *M. m. musculus* Y chromosomes is $275,000$ (95% highest posterior density interval [HPDI] $267,000 - 282,000$) years ago. Within *M. musculus*, the *musculus* subspecies diverges first, although the internal branch separating it from the MRCA of *domesticus* and *castaneus* is very short. Consistent

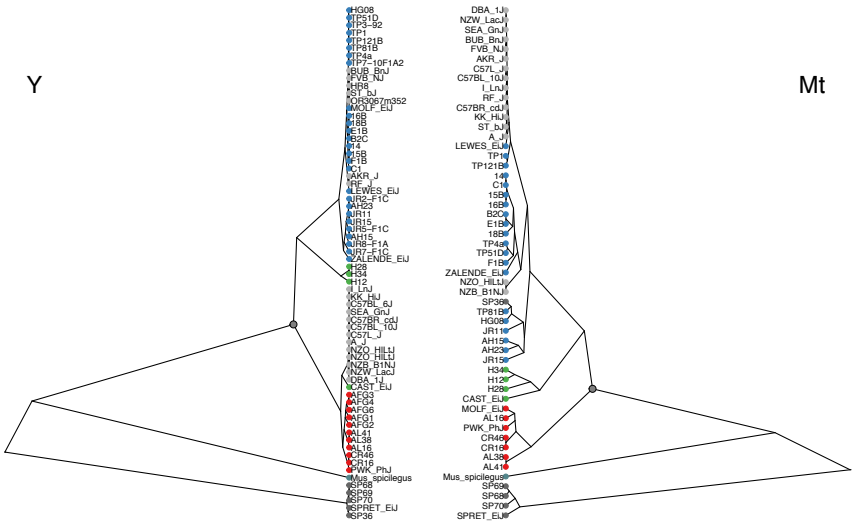

Figure 6.4: Phylogenetic trees for Y chromosomes (left) and complete mitochondrial genomes (right) of wild and laboratory mice. Wild mice are colored according to their taxonomic origin: blue, *M. m. domesticus*; red, *M. m. musculus*; green, *M. m. castaneus*; dark grey, *M. m. spretus*; and light grey, classical laboratory strains.

with several previous studies, we find that the "old" classical inbred strains share a single Y haplogroup within *M. m. musculus*. This haplogroup is distinct from that of European and central Asian wild mice and is probably of east Asian origin[137,390]. Strains related to "Swiss" outbred stocks (FVB/NJ, NOD/ShiLtJ, HR8) and those of less certain American origin (AKR/J, BUB/BnJ)[155] have Y chromosomes with affinity to western European populations. *M. m. castaneus* harbors two distinct paraphyletic lineages: one corresponding to the Indian subcontinent and another represented only by the wild-derived inbred strain CAST/EiJ (from Thailand.) The latter haplogroup probably corresponds to the southeast Asian lineage identified in previous reports[130,157].

The Y-chromosome tree otherwise shows perfect concordance between clades and geographic locations. Within the *M. m. domesticus* lineage we can recognize two distinct haplogroups corresponding roughly to western Europe and Iran and the Mediterranean basin, respectively. Similarly, within *M. m. musculus*, the eastern European mice (from Bavaria, Czech Republic) are well-separated from the central Asian mice (Kazakhstan and Afghanistan). Relationships between geographic origins and phylogenetic affinity are considerably looser for the mitochondrial genome.

We even found evidence for inter-subspecific introgression: one nominally *M. spretus* individual from central Spain (SP36) carries a *M. spretus* Y but a *M. m. domesticus* mitochondrial genome (arrowhead in **Figure 6.4**).

6.2.3 Sequence diversity and tests for selection

We estimated nucleotide diversity within subspecies directly from genotype likelihoods[391], rather than from called genotypes at variable sites. The rank ordering of subspecies by Y chromosome diversity parallels what has previously been shown for autosomes: *castaneus* >> *domesticus* > *musculus* (**Table 6.2**). Our estimates of diversity at Y-linked sites ($\pi_{\text{dom}} = 0.029\% \pm 0.001\%$, $\pi_{\text{mus}} = 0.037\% \pm 0.001\%$, $\pi_{\text{cas}} = 0.177\% \pm 0.003\%$) are in line with previous reports[291,130]. To provide context for observed levels of Y-linked variation, we compared relative diversity in pairwise combinations of the autosomes, X and Y chromosomes within subspecies to neutral expectations (**Table 6.3**). We found a deficit of variation on both sex chromosomes relative to the autosomes. The effect is stronger on the X (approximately 80% lower nucleotide diversity than expected) than the Y chromosome (40%), and is stronger in *domesticus* and *musculus* than in *castaneus*.

Levels of population differentiation, measured by F_{st}, are also increased on the sex chromosomes relative to autosomal loci. Here the effect is strongest for the Y chromosome (**Table 6.4**), with F_{st} values ranging from 0.62 (*musculus-castaneus*) to 0.71 (*domesticus-castaneus*).

To investigate possible causes of reduction in diversity on the Y chromosome, we used two complementary families of tests: the Hudson-Kreitman-Aguade (HKA) test[392] and variations on Tajima's D statistic[393]. The HKA test compares the ratio of polymorphisms to fixed substitutions at two or more unlinked loci. Under the null hypothesis, the rate of fixation of neutral alleles is equal across loci even if locus-specific mutation rates are not. We compared the Y chromosomes to the mitochondria and to X chromosome separately in *domesticus*, *musculus* and *castaneus*. The null hypothesis is rejected for *domesticus* ($p = 8.9 \times 10^{-5}$) and *musculus* ($p = 0.04$) but not for *castaneus* ($p = 0.76$) in the Y-mitochondria comparison. In both *musculus* and *domesticus*, the Y chromosome shows a deficit of polymorphism relative to the mitochondria (**Table 6.5**). No excess of divergence relative to polymorphism was detected in the Y-X comparison.

The second family of tests is typified by Tajima's D statistic. These statistics represent standardized differences between two different estimators of the population-scaled mutation rate and capture departures from neutrality in different portions of the site frequency spectrum (SFS).

Locus	Pop	L	N	S	θ_w	θ_π	D	D_{FL}	F	Y
A	*dom*	235019	52	3961	0.373 (0.006)	0.339 (0.006)	−0.33	2.35	1.47	−0.42
	mus		20	3875	0.465 (0.008)	0.409 (0.008)	−0.51	1.72	1.16	−0.40
	cas		6	4496	0.838 (0.013)	0.786 (0.013)	−0.40	1.87	1.57	−0.15
X	*dom*	77654	26	349	0.118 (0.008)	0.056 (0.005)	−2.09	−5.11	−4.92	−0.89
	mus		10	263	0.120 (0.009)	0.073 (0.005)	−1.97	−3.51	−3.77	−0.53
	cas		4	485	0.341 (0.017)	0.315 (0.015)	−0.80	−1.80	−1.91	−0.073
Y	*dom*	995467	26	2199	0.058 (0.001)	0.029 (0.001)	−1.97	−4.86	−4.66	−0.84
	mus		10	1613	0.057 (0.001)	0.037 (0.001)	−1.78	−3.12	−3.35	−0.49
	cas		4	3493	0.191 (0.003)	0.177 (0.003)	−0.79	−1.24	−1.37	−0.07
M	*dom*	979	26	18	0.482 (0.177)	0.142 (0.052)	−2.45	−4.82	−4.87	−1.21
	mus		10	9	0.335 (0.169)	0.190 (0.096)	−1.83	−1.51	−1.91	−0.65
	cas		4	3	0.141 (0.111)	0.130 (0.101)	−0.63	−1.14	−1.21	−0.09

Table 6.2: Sequence diversity statistics for autosomes, X and Y chromosomes and mitochondrial genome, by population. L, total sizes; S, segregating sites; θ_w, Watterson's θ; θ_π, Tajima's pairwise θ; D, Tajima's D; D_{FL}, Fu and Li's D; F, Fu and Li's F; Y, Achaz's Y. Both estimators of θ are expressed as percentages with bootstrap standard errors in parentheses.

		Population		
Comparison	**Expected**	*dom*	*mus*	*cas*
X:A	3/4	0.163 (0.012)	0.176 (0.013)	0.399 (0.026)
Y:A	1/4	0.0868 (0.0028)	0.0909 (0.0030)	0.225 (0.005)
Y:X	1/3	0.531 (0.047)	0.5120 (0.0392)	0.5640 (0.0294)

Table 6.3: Diversity ratios between pairs of chromosome types. Bootstrap standard errors are shown in parentheses. Rightmost column shows expected values under neutral model with equal sex ratios.

		Population		
Locus		*dom*	*mus*	*cas*
A	*dom*	-	0.275 (0.0031)	0.388 (0.0033)
	mus	0.822 (0.013)	-	0.284 (0.0028)
	cas	0.91 (0.015)	0.847 (0.011)	-
X	*dom*	-	0.622 (0.012)	0.681 (0.0094)
	mus	0.165 (0.011)	-	0.567 (0.0098)
	cas	0.231 (0.011)	0.238 (0.01)	-
Y	*dom*	-	0.68 (0.0034)	0.71 (0.0026)
	mus	0.165 (0.0031)	-	0.616 (0.0033)
	cas	0.191 (0.0029)	0.171 (0.0029)	-
M	*dom*	-	0.65 (0.12)	0.87 (0.054)
	mus	0.52 (0.19)	-	0.529 (0.14)
	cas	0.566 (0.21)	0.289 (0.11)	-

Table 6.4: Population differentiation (F_{st}, above diagonal) and divergence per site (d_{xy} as percentage, below diagonal) for autosomes and sex chromosomes. Bootstrap standard errors in parentheses.

	Population		
Loci	*dom*	*mus*	*cas*
Y vs X	0.79 (0.374)	0.45 (0.502)	0.03 (0.853)
Y vs M	*15.35 (8.9×10^{-5})	*4.42 (0.0356)	0.10 (0.755)
X vs M	*8.36 (3.8×10^{-3})	2.10 (0.148)	0.20 (0.653)

Table 6.5: Hudson-Kreitman-Aguadé (HKA) tests for neutral evolution of Y chromosomes compared to X-linked or mitochondrial loci. Entries in the table are the χ^2 statistic from the HKA test with p-values in parentheses. Comparisons for which the null hypothesis is rejected are marked with asterisks (*).

Negative values indicate a skew in the SFS towards low-frequency alleles; significance thresholds are established with coalescent simulations. Tajima's D, Fu and Li's D and F[394] and Achaz's Y[395] all take significantly negative values on the Y chromosome in *domesticus* and *musculus* but not *castaneus* (**Table 6.2**).

6.2.4 Demography of male lineages

Because it is inherited only through the male line and does not undergo recombination, the Y chromosome is a sensitive marker for the male-specific demographic history of populations. We used an approximate Bayesian computation (ABC)[396,397] strategy to evaluate models for patrilineal demography against our Y chromosome dataset. Neutral coalescent simulations were carried out under several demographic scenarios (**Figure 6.5A**). Simulated and observed polymorphism data at putatively neutral sites were compared using summaries over the joint SFS across *domesticus*, *musculus* and *castaneus* (see **Methods**). In the ABC scheme, a subset of simulations yielding summary statistics "close" to the observed values are treated as a sample from the marginal posterior distribution over demographic model parameters[398].

We evaluated eight families of demographic models of increasing complexity and used Bayes factors for model selection (**Figure 6.5**). Models with gene flow (I – IV) generally provided better fit to the data than models without gene flow (V – VIII). The best-fitting model (model VII) includes a bottleneck shared by *domesticus* and *musculus* but not *castaneus*. It captures the key features of the observed data: reduced diversity in *domesticus* and *musculus*; excess of low-frequency alleles in *domesticus* and *musculus*; and approximately equal F_{st} between all population pairs (**Figure ??**). Under this model, N_e for castaneus is approximately 1.5-fold higher than in *domesticus* or *musculus* and the three Y chromosome lineages began to diverge $636,000$ generations in the past (**Figure 6.6** and **Table 6.6**). The inferred bottleneck is sharp, reducing N_e by 89% (50% HPDI $2-13\%$). Its timing ($19,700$ generations in the past) is consistent with fossil and genetic evidence that Eurasian mammal populations experienced a sharp contraction around the time of the last glacial maximum $10,000-25,000$ years ago[399,400,132]. The demographic parameters we estimate from our relatively small sample of Y chromosomes is in good agreement with previous estimates from a much larger sample but under a more restrictive family of models[130]. Importantly, they offer a plausible alternative to selection to explain the high differentiation and low diversity of Y chromosomes in *M. musculus*.

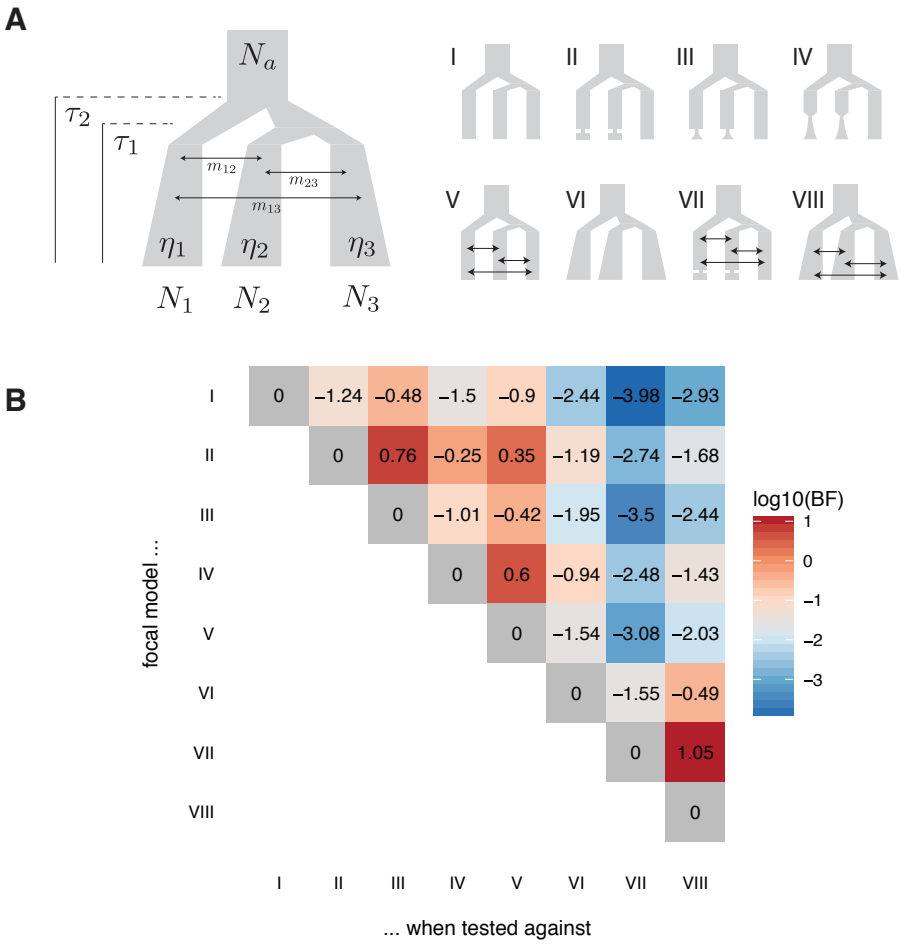

Figure 6.5: Definition and selection of patrilineal demographic history. (**A**) Schematic representation of eight three-population scenarios used in the simulation step of the ABC procedure. (**B**) Pairwise comparisons of relative goodness-of-fit using Bayes factors (BF). Each cell shows the $\log_{10}$ BF of the fitted model on the y axis against the fitted model on the x-axis. One log-unit represents a 10-fold relative increase in posterior model probability.

	Parameter						
Model	N_0	N_{cas}	N_{dom}	N_{mus}	τ_1	τ_2	β
I	61.2 (23.5, 68.3)	44.2 (22.1, 53.9)	39.3 (17.9, 46.5)	40.4 (15, 44.9)	93 (39.4, 116)	201 (109, 259)	-
II	80.6 (45.7, 91.9)	52.4 (31.7, 62)	59 (36.3, 70.8)	54.5 (31.2, 66.8)	138 (60.6, 151)	260 (147, 297)	0.281 (0.262, 0.463)
III	66.2 (47.9, 72.5)	44.9 (24, 50)	43.4 (30.6, 47.1)	41.7 (26.1, 50.1)	83.5 (36.7, 94.6)	231 (133, 260)	0.242 (0.198, 0.419)
IV	82 (43.5, 98.8)	60.8 (29.8, 70.5)	54.6 (33.6, 74.1)	55.4 (22.3, 66)	133 (37.9, 145)	277 (114, 315)	0.132 (0.0682, 0.215)
V	78.7 (50.8, 85.4)	56.8 (35.1, 63.5)	42.4 (23, 47.2)	56.1 (37.9, 63.2)	185 (96.6, 203)	334 (207, 377)	-
VI	69.9 (39.1, 82.3)	41.8 (16.9, 48.6)	33.2 (11.8, 38.8)	36.3 (16.2, 44.9)	221 (84.9, 270)	456 (256, 577)	-
VII	98.5 (65.4, 116)	64.2 (35.6, 73.6)	54.7 (35.7, 69)	46.9 (13.2, 53.9)	317 (74.2, 352)	636 (269, 725)	0.107 (0.02, 0.128)
VIII	63.2 (42, 68.8)	42.2 (24, 46.7)	20.5 (12.5, 28.9)	27.8 (12.8, 33.6)	226 (137, 261)	427 (247, 458)	-

	Parameter					
Model	η_2	η_3	η_1	m_{12}	m_{13}	m_{23}
III	1.03 (0.494, 1.27)	0.988 (0.393, 1.11)	-	-	-	-
IV	0.941 (0.282, 1.21)	0.973 (0.833, 1.69)	-	-	-	-
V	1.04 (0.291, 1.13)	1.01 (0.453, 1.2)	0.629 (0.237, 0.959)	-	-	-
VI	-	-	-	0.897 (0.309, 1.3)	1.15 (0.598, 1.4)	0.886 (0.0498, 0.981)
VII	-	-	-	0.84 (0.457, 1.29)	0.694 (0.319, 0.924)	0.829 (0.192, 1.15)
VIII	1.01 (0.877, 1.7)	1.13 (0.758, 1.6)	1.03 (0.664, 1.5)	1.27 (0.984, 1.86)	0.904 (0.423, 1.13)	0.77 (0.0893, 0.817)

Table 6.6: Parameter estimates for demographic models fit to Y chromosome SFS by ABC. Values shown are posterior medians with 50% HPDIs. Models refer to **Figure 6.5**. Units are as follows: for population sizes, thousands of individuals; for split times, thousands of generations; for growth rates, unitless exponential rate constants; for migration rates, units of $4N_e m$, the effective number of migrants per generation.

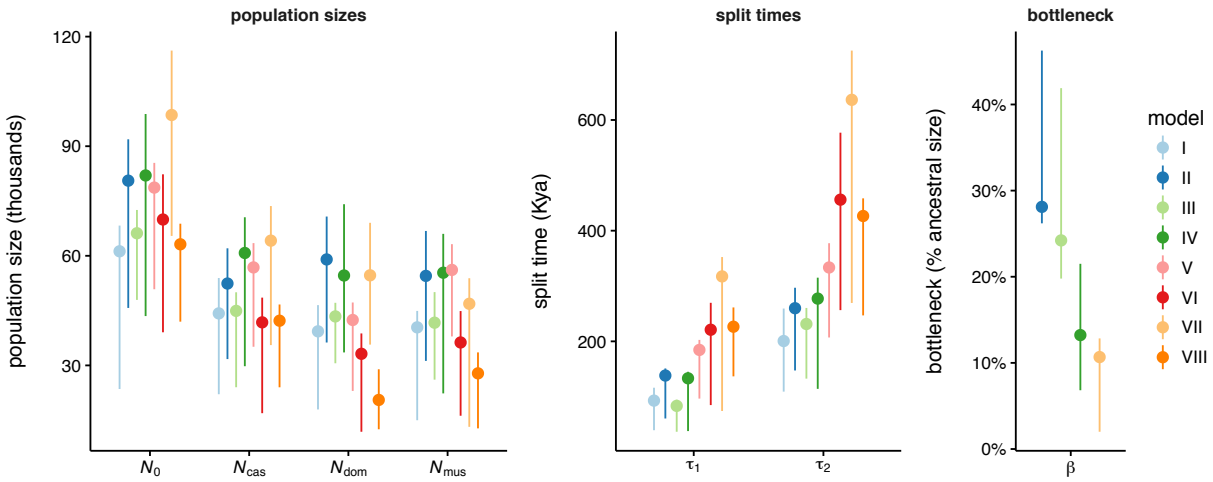

Figure 6.6: Marginal posterior distributions of key demographic parameters, shown as posterior median and 50% HPDI. Notation follows **Figure 6.5**.

6.2.5 Modes of copy-number variation on the Y

We examined copy number along Yp using depth of coverage. Approximately 779 kb (24%) of Yq consists of segmental duplications or gaps in the reference assembly (**Figure 6.3**); for these regions we scaled the normalized read depth by the genomic copy number in the reference sequence to arrive at a final copy-number estimate for each individual. All of the known duplications on Yp are polymorphic in laboratory and natural populations (**Figure 6.7**). The distribution of CNV alleles follows the SNV-based phylogenetic tree. Only one region, at the centromeric end of Yq, contains a known protein-coding gene (*Rbmy*). Consistent with a previous report[378], we find that *musculus* Y chromosomes have more copies of *Rbmy* than *domesticus* or castaneus chromosomes. We identified one additional CNV overlapping a gene: the wild-derived inbred strain LEWES/EiJ (from Delaware; *M. m. domesticus* ancestry) carries an 82 kb duplication containing *Eif2s3y*.

The highly repetitive content of Yq precludes a similarly detailed characterization of copy-number variation on this chromosome arm. However, by aggregating the reads aligning to the fundamental ampliconic units it is possible to assess the total size and proportional composition of Yq. We counted the total number of reads mapping to the interdigitated "red", "blue" and "yellow" sequence families (as defined in[306] and shown in **Figure 6.3**), divided by the total number of mapped reads, to estimate the composition and size of Yq in each sample. Consistent with the hypothesis that Yq expands and contracts by gain or loss of copies of the ampliconic unit, we find

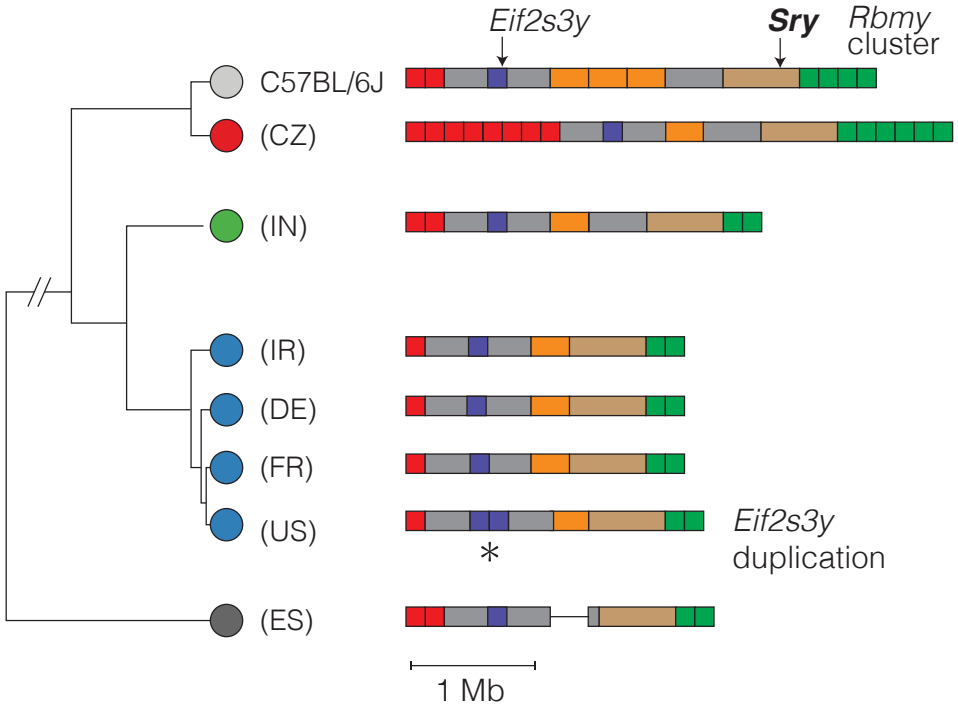

Figure 6.7: Schematic view of structural variation on the Y chromosome short arm (Yq), superposed on SNV-based phylogenetic tree. Copy-number variable regions are indicated with colored blocks and invariant regions with grey blocks. All CNVs shown overlap a segmental duplication in the reference sequence. Only two CNVs overlap protein-coding genes: a duplication in North American mice encompassing *Eif2s3y* (purple) and an expansion of the ampliconic *Rbmy* cluster (green) in *M. m. musculus*. Color scheme for *Mus* taxa follows **Figure 6.4**.

that the proportional composition of Yq is very similar across taxa (**Figure 6.8A**). However, the total size of Yq varies dramatically within *Mus* (**Figure 6.8B**: from a median 19 Mb in *M. spretus* to 61 Mb in *M. m. domesticus*.

The hypothesis of X-Y intragenomic conflict predicts that, if expression levels are at least roughly proportional to copy number, amplification of gene families on Yq should be countered by amplification of their antagonistic homologs on the X. We tested this hypothesis by comparing the copy number of X- and Y-linked homologs of the *Slx/y*, *Sstx/y* and *Srsx/y* families in wild mice. **Figure 6.9** shows that copy number on X and Y are correlated only for *Slx/y*; within that family, the *M. m. musculus* form an outlying clusetr. The relationship between *Slx*-family and *Sly*-family copy number is almost exactly linear if *M. m. musculus* samples are excluded (slope = 1.1 [95% CI $0.9 - 1.2$]; $R^2 = 0.84$). This supports previous evidence that conflict between X and Y, if it exists, is mediated primarily through expression of *Slx* and *Sly*[379].

The intragenomic conflict hypothesis also predicts selection for copy number of co-amplified regions on the X chromosome. This should reduce nucleotide diversity at sites closely linked to the co-amplified regions relative to sites further away. We calculated nucleotide diversity (θ_π) and Tajima's D in 100 kb windows across the X chromosome in same samples for which we estimated copy number on Yq. Notwithstanding the X-chromosome-wide deficit in nucleotide diversity relative to autosomes, we observed neither additional reduction in diversity in the vicinity of co-amplified regions nor a skew towards low-frequency variants (**Figure 6.10**). Tests for a linear relationship between diversity and distance from the nearest co-amplified region, or for an ordinal trend across bins of distance, were not significant in any population.

6.2.6 Differentiation of Y-linked gene expression during spermatogenesis

The gene complement of the Y chromosome is specialized for function in the male germline. To understand functional differences between Y chromosome haplogroups we therefore focused on expression of Y-linked genes in testis. To isolate the effect of the Y chromosome in *cis* from *trans* effects of genetic background (and in particular the X chromosome), we re-analyzed expression data from two published studies in intersubspecific hybrids between *M. m. domesticus* and *M. m. domesticus*.

In the first study[151], gene expression was measured by microarray in whole testes of 175 age-matched, laboratory-reared males that were first-generation offspring of parents trapped in the

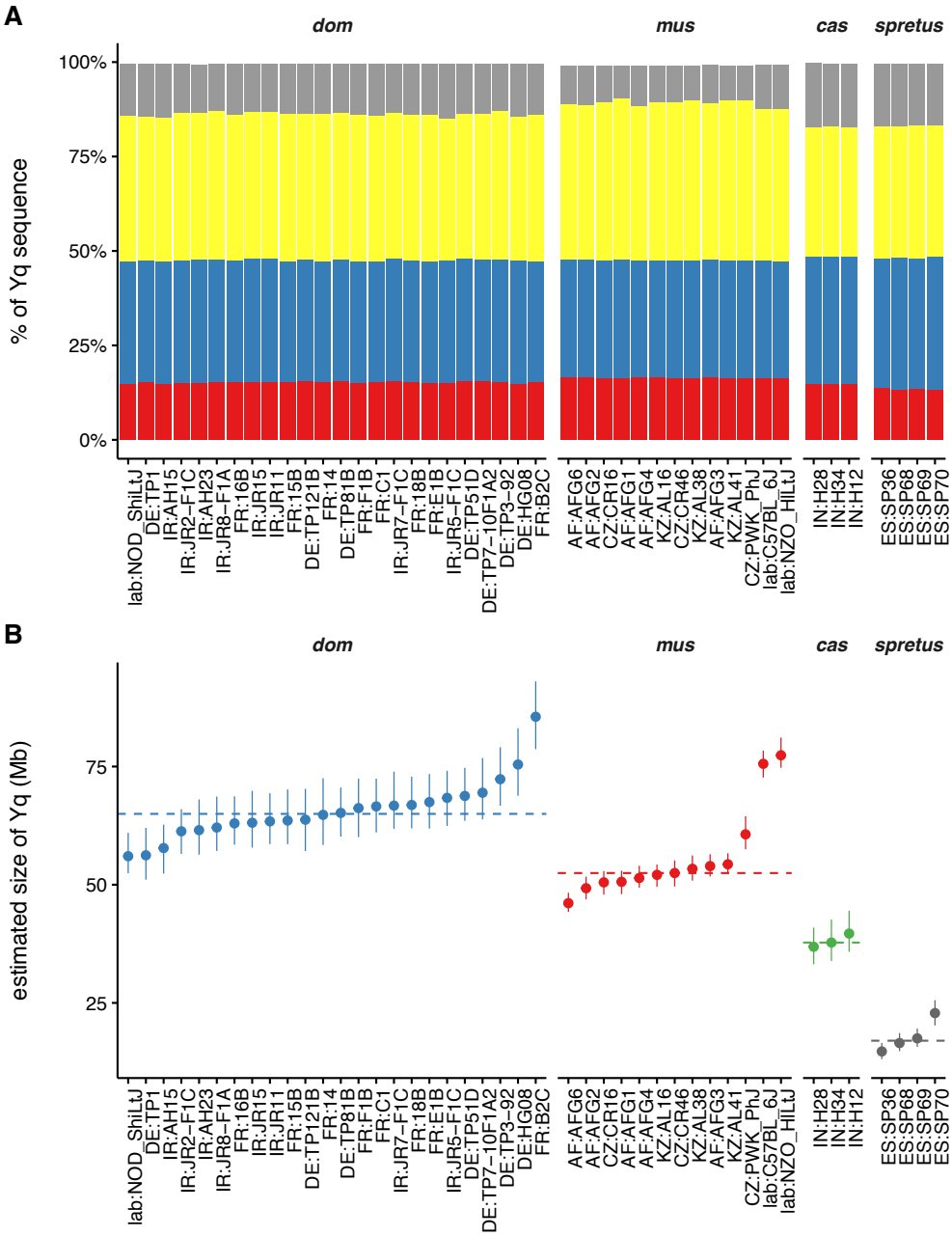

Figure 6.8: Structural variation on the Y chromosome long arm (Yq). (**A**) Proportional composition of Yq from wild mice and selected laboratory strains of all three subspecies plus *M. spretus*, according to "red," "yellow" and "blue" and "other" sequence families defined in[306]. Each column corresponds to a single sample; sample names are prefaced by country of origin. (**B**) Estimated total size of Yq for the samples shown in panel A. Dashed lines indicate within-subspecies median.

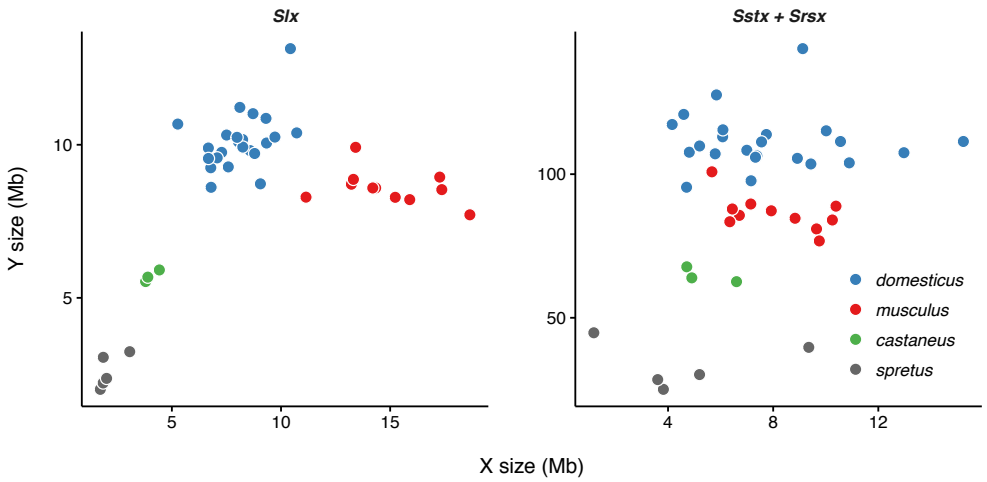

Figure 6.9: Genomic size of co-amplified gene families on X versus Y. Each point corresponds to a single individual.

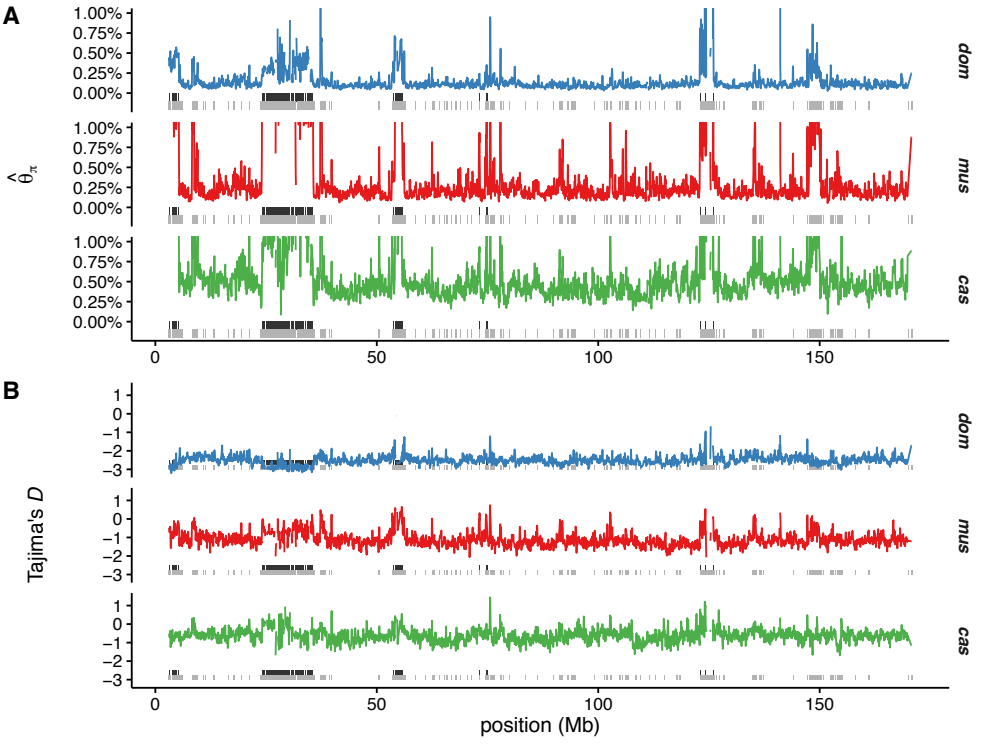

Figure 6.10: Sequence diversity across the X chromosome. (**A**) Within-population sequence diversity across the X chromosome, measured by Tajima's pairwise estimator θ_π. Dark grey boxes below the x-axis show locations of co-amplified regions; light grey boxes show all segmental duplications > 1 kb in size. (**B**) As above, but showing Tajima's D.

domesticus-musculus hybrid zone in southern Germany. This experimental design allowed us to evaluate the marginal differences in expression between the *domesticus* and *musuclus* Y chromosomes against autosomal and X-chromosome backgrounds with varying degrees of intersubspecific admixture. We first assessed global patterns of Y-linked expression variation using principal components analysis (PCA) (**Figure 6.11A**). The first principal component, accounting for 32% of expression variance, clearly separates individuals by Y-chromosome haplogroup. The association between PC1 and Y haplogroup remains strong even after accounting for possible effects of admixture on the autosomes and X-chromosome ($F_{1,173} = 269.3$, $p < 10^{-10}$). Surprisingly, the effect is magnified by *domesticus* ancestry on the X chromosome ($F_{1,173} = 5.4$, $p = 0.02$), implicating X-Y interactions in the regulation of Y-linked genes.

Of $21,139$ genes assayed by the microarray platform, $9,559$ (45%) were differentially-expressed according to Y chromosome haplogroup at false discovery rate (FDR) < 0.05; of those, 373 (2%) had an expression difference of two-fold or greater. Among 19 Y-linked genes 16 (84%) were differentially-expressed (**Figure 6.11B**), but only one — *Sly* — had fold-difference greater than two. *Sly* has 2.4-fold (95% CI $2.1 - 2.8$) higher expression from *musculus* than *domesticus* Y chromosomes.

A second study[401] measured gene expression by RNA-seq in the testes of reciprocal F_1 hybrids between the wild-derived inbred strains LEWES/EiJ (from Delaware; *M. m. domesticus* ancestry) and PWK/PhJ (from the Czech Republic; primarily *M. m. musculus* ancestry). We re-analyzed the RNA-seq data using an improved transcript annotation which includes a comprehensive set of transcript models for co-amplified genes on Yq and the X chromosome in addition to transcript models in the public Ensembl annotation (see **§ 6.5**). Expression was estimated at the transcript level, and these estimates were aggreaged to gene level for analysis. PCA on expression of Y-linked genes separates F_1s according to the paternal strain (**Figure 6.12A**). Of 19 Y-linked genes with coding potential, 9 (47%) are differentially-expressed according to Y chromosome genotype (**Figure 6.12B**). Two genes on Yp show expression differences consistent in direction with a difference in DNA copy number. *Eif2s3y* (copy gain in LEWES/EiJ) has 2.3-fold (95% CI $1.4 - 3.6$) increased expression from the LEWES/EiJ Y chromosome, and *Rbmy* (copy gain in PWK/PhJ) has 2.0-fold (95% CI $1.3 - 2.8$) increased expression from the PWK/PhJ Y chromosome. As in mice from the hybrid zone, *Sly* is 2.2-fold (95% CI $1.2 - 4.0$) higher expressed from the *musculus* Y chromosome.

Because dosage imbalance between X- and Y-linked co-amplified genes causes infertility and

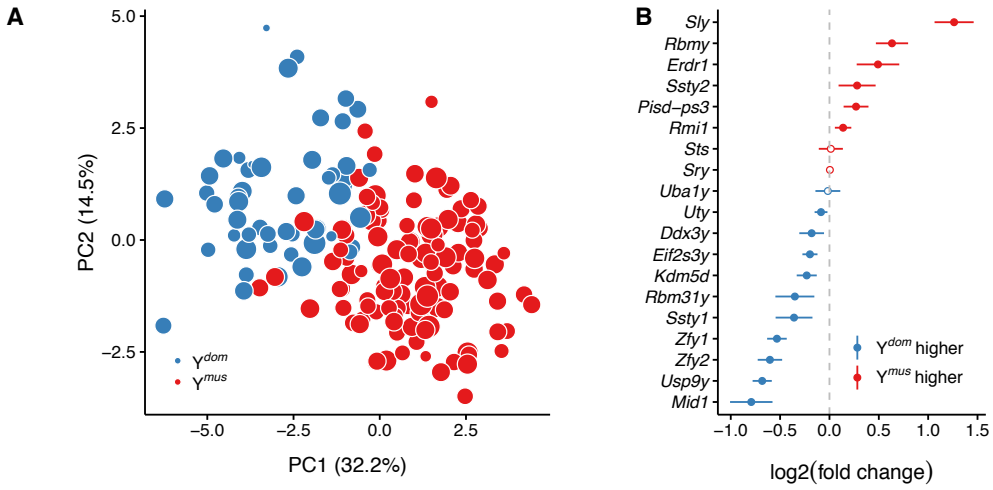

Figure 6.11: Y-linked gene expression in testes of mice from the *domesticus-musculus* hybrid zone in Bavaria. (**A**) Principal components analysis (PCA) on expression of Y-linked genes in 175 male mice. Points are colored according to Y chromosome haplogroup and sized according to testis weight, a surrogate phenotype for hybrid sterility. (**B**) Test for differential expression of Y-linked genes by Y chromosome haplogroup. Closed circles, significant difference at FDR < 0.05; open circles, no significant difference. Effect size is represented as $\log_2$ (fold difference) with corresponding 95% CI.

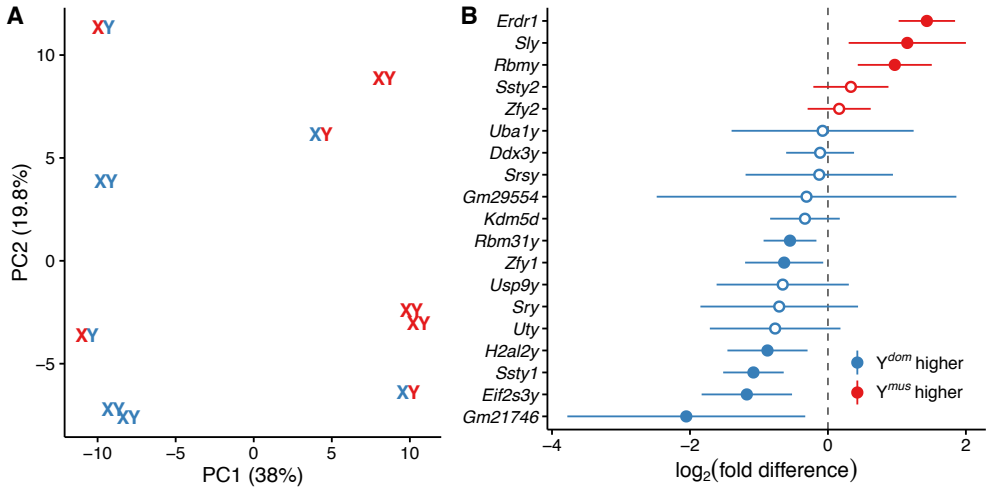

Figure 6.12: Y-linked gene expression in testes of intra-subspecific and reciprocal inter-subspecific F_1 hybrids between wild-derived strains of *M. m. domesticus* and *M. m. musculus* origin. (**A**) Principal components analysis (PCA) on expression of Y-linked genes in 10 male mice. Points are colored according to ancestry of X- and Y-chromosomes (blue, *M. m. domesticus*; red, *M. m. musculus*.) (**B**) Test for differential expression of Y-linked genes by Y chromosome haplogroup. Closed circles, significant difference at FDR < 0.05; open circles, no significant difference. Effect size is represented as $\log_2$ (fold difference) with corresponding 95% CI.

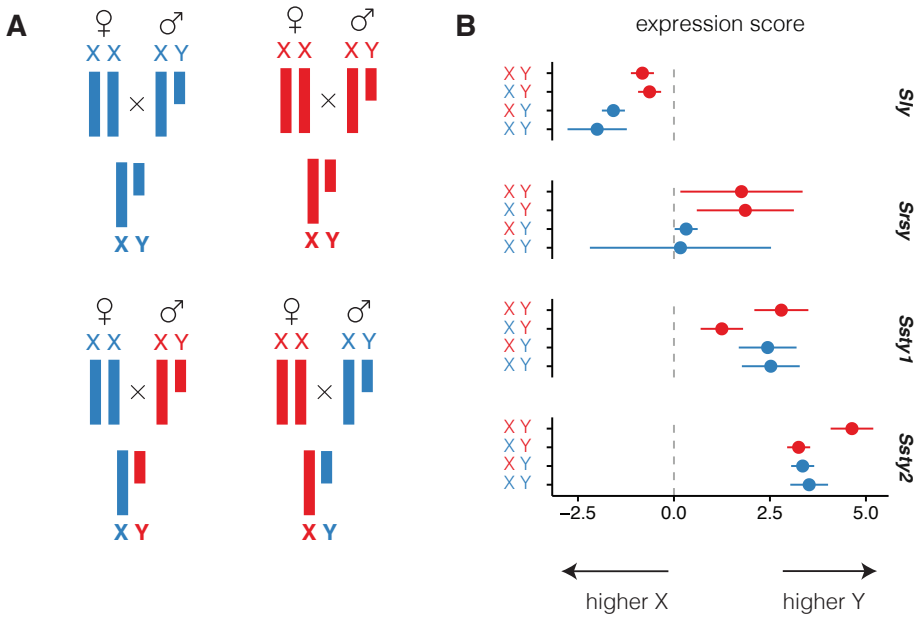

Figure 6.13: X vs Y expression balance in testis for co-amplified gene families. (**A**) Experimental design: intra-subspecific (top panel) and reciprocal inter-subspecific (bottom panel) F_1 hybrids between wild-derived strains. *M. m. domesticus* shown in blue and *M. m. musculus* in red. (**B**) Expression score (see **§ 6.5**), measuring relative expression from X- and Y-linked members of co-amplified gene families in testes of males from the crosses shown in panel A.

sperm abnormalities, we also examined the X:Y expression ratio for each gene family (**Figure 6.13**). Expression of Y-linked copies (versus X-linked homologs) of *Sly* and *Srsy* is lower from the *domesticus* than the *musculus* Y chromosome, independent of X chromosome origin. The pattern for *Ssty1* and *Ssty2* relative to *Sstx* is more complex: expression balance for these genes appears to depend on an X-Y interaction.

6.3 Discussion

6.3.1 Phylogeography of mouse Y chromosomes

We confirm the long-standing observation that at least two Y haplogroups are present in classical laboratory strains and related outbred stocks[137,390]. One haplogroup falls within the *M. m. musculus* clade almost certainly originated in Japan and represents part of the *M. m. molossinus* contribution to classical inbred strains[162,157]. The last common ancestor of Y chromosomes in this haplogroup

was recent: within the last 550 years. The other two haplogroups are of *M. m. domesticus* origin. One is present in "Swiss" mice such as NOD/ShiLtJ and FVB/NJ and has closest affinity to Y chromosomes found in present-day northern and central Germany, while the other is found in American strains and is not clearly associated with a sampled European lineage (**Figure 6.4**).

Among Y chromosomes from wild mice, phylogenetic affinity mirrors geography. The same cannot be said for the mitochondria, which are both more genetically diverse within populations and less differentiated between them: see **Figure 6.14**. The correlation between geographic origin and phylogenetic distance is $\rho = 0.24$ (95% CI $0.21 - 0.27$) for the Y chromosome but only $\rho = 0.10$ (95% CI $0.08 - 0.13$) for the mitochondrial genome. We found one case of inter-specific introgression involving a *M. spretus* female and a *M. m. domesticus* male. Taken together, these observations indicate that the degree of genetic mixing is greater for female than male lineages. Several explanations are possible. First, dispersal behavior may differ between sexes. There is little evidence to support the conjecture that female mice disperse more readily than males; if anything, the opposite is true[402,403,367]. But females are generally more successful at integrating into a new group than males[404,367]. Second, genetic incompatibilities may accumulate more rapidly on the Y chromosome and serve as a barrier to gene flow. Studies of the *domesticus-musculus* hybrid zone in eastern Europe have consistently shown that allele-frequency clines are narrower and steeper for Y-linked than autosomal loci[138,405], and hybrid male sterility constitutes the primary reproductive barrier between mouse subspecies[141]. However, it is seems unlikely that genetic incompatibilies would arise within $5,000 - 10,000$ generations between local populations of the same subspecies (*e.g* in France and Germany). Finally, the lack of apparent geographic mixing between male lineages may simply be a consequence of the low diversity of Y chromosomes. If one or two lineages dominate in each locality, it is not surprising that we should observe little sharing between localities.

6.3.2 What explains the deficit of Y-linked sequence variation?

In the absence of selection and assuming equal mutation rates at all loci, genetic diversity is proportional to the (effective) number of chromosomes in the population[81]. Expected diversity on the Y chromosome, which is hemizygous and only passed through the male germline, is therefore only one-fourth that of the autosomes, which are diploid and passed through both sexes, in a population with sex ratio at parity[256]. Departures from these ratios can be a signal of (1) unequal sex

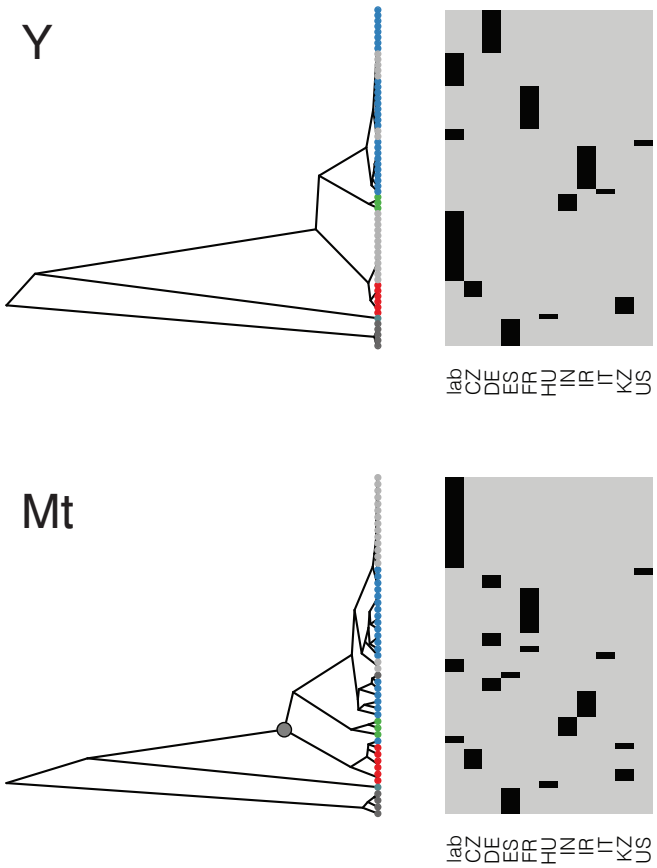

Figure 6.14: Phylogenetic versus geographical relationships for Y chromosomes (top) compared to mitochondria (bottom). Phylogenetic trees and color scheme follow **Figure 6.4**. Heatmaps at right are incidence matrices of samples onto countries (shown as two-letter country codes). More block structure indicates greater clustering of samples by geography.

ratio; (2) sex differences in mutation rate; (3) population size changes; or (4) selection[406]. Although the selection imposed by intragenomic conflict on the sex chromosomes would reduce genetic diversity relative to the autosomes, we sought to rule out other more pedestrian explanations.

An excess of males versus females in the population increases Y:A and decreases X:A relative to the neutral expectation; an excess of females has the reverse effect[256]. This pattern is not consistent with our data: we observe marked reduction in both X:A and Y:A (**Table 6.3**). We observed Y:X diversity ratios of approximately 0.53 ± 0.05 in *M. m. domesticus* and *M. m. domesticus*, significantly greater than the expected value of $\frac{1}{3}$. This discrepancy can be explained in part by the quite strong reduction in diversity across the entire X chromosome relative to the autosomes (**Table 6.3**). Differences in germline mutation rate also likely contribute: in mammals, the mutation rate is generally higher in males than females (although the details of the relationship α depend on life history[26]). The Y:X diversity ratio we observe is consistent with $\alpha \approx 3$[406], in good agreement with the empirical estimate of $\alpha = 2.78$[10].

Sex chromosomes, because of their smaller effective population sizes, are more sensitive to the effects of population growth and contraction than autosomes. Both X:A and Y:A decline during a bottleneck[407]. In addition to decreasing nucleotide diversity, bottlenecks are also predicted to leave behind an excess of low-frequency alleles that can be detected by statistics of the form of Tajima's D[393]. This is exactly what we observe in *M. m. domesticus* and *M. m. musculus* for both sex chromosomes (**Table 6.2**). The demographic models we fit to Y chromosome SFS via ABC support a strong bottleneck in *M. m. domesticus* and *M. m. musculus*, the populations with the greatest reductions in X:A and Y:A diversity (**Figure 6.5**). A bottleneck is thus a parsimonious, neutral explanation for the deficit of nucleotide diversity we observe on both sex chromosomes.

Nonetheless, several caveats apply. First, defining a single population for each subspecies obscures further substructure which is certainly present in the data. Sample size outside *M. m. domesticus* was not sufficient to divide the populations further but this deserves attention in future work. Second, model choice for ABC is subjective. We chose a panel of population-genetic scenarios that represent plausible histories of house mouse populations based on the literature. Many other scenarios are possible and our power to discriminate between them is limited given our quite modest sample size. Finally, we cannot exclude a role for background selection — a decrease in diversity at sites linked to targets of purifying selection[408] — on sex chromosome diversity,

particularly on the Y chromosome.

6.3.3 Mutational mechanisms on the Y chromsome

The Y chromosome provides a direct view of the mutational spectrum of the male germline. We exploited this fact to estimate the male-specific point mutation rate in mouse ($5.4 \times 10^{-9} - 8.1 \times 10^{-9}$ bp^{-1} generation^{-1}). The mutation rate for large structural variants, especially on Yq, must be much higher: the Yq has more than tripled in size in less than 2 My between the divergnce of *M. spretus* and *M. m. domesticus* (**Figure 6.8**). Clusters of duplicated sequences are often assumed to be especially mutable because they are prone to non-allelic homologous recombination[44]. But this is trivially not the case for the male-specific portion of the Y chromosome, which has no homologous partner with which to pair or recombine. Structural variation on the Y must therefore arise via errors of replication during mitosis or by intrachromosomal recombination. In humans and other great apes, exchange between duplicated sequences on opposite arms of the metacentric primate Y chromosome appears to be common[307,409,266]. We propose that a similar process drives the expansion and contraction of Yq amplicons in mouse. This provides further evidence for an idea discussed at length elsewhere in this thesis: unpaired sequences are prone to structural mutation in male meiosis. If that is true, we predict that expansion of X-linked ampliconic families also occurs primarily via mutation in the male germline.

6.3.4 Equivocal support for hypothesis of X-Y intragenomic conflict

The sequence content of all mammal Y chromosomes studied to date can be divided into two classes: ancestral genes with X-linked homologs and broad tissue expression patterns, and acquired genes expressed only during spermatogenesis[5]. The acquired genes are lineage-specific and often have X-linked homologs, and both the X- and Y-linked members of the pair exist in many copies[410]. This has led several authors to conclude that the evolution of mammalian sex chromosomes is driven by recurrent intragenomic conflicts whose principal actors are members of the "co-amplified" gene families. Over long evolutionary timescales — between species — intragenomic conflict should result in correlated evolution of X- and Y-linked gene families. Previous studies comparative analyses of the mouse and rat X[25] and Y[306] supported this idea: multicopy, lineage-specific genes on the X all have Y-linked homologs. One pair of co-amplified families, *Slx* and *Sly*, have opposing actions in post-meiotic spermatids and appear to promote the transmission of their own chromosome[388,411,379].

The hypothesis of X-Y intragenomic conflict in mouse has not been tested over short (within-species) evolutionary timescales. In this chapter we use newly-available whole-genome sequence data from wild mice to and published gene expression data from experimental crosses to test for signatures of intragenomic conflict. We found a linear and almost exactly one-to-one relationship between the (estimated) copy number of *Slx* and *Sly* across 2 My of evolution in *M. spretus*, *M. m. castaneus* and *M. m. domesticus*, consistent with the conflict hypothesis (**Figure 6.9**). *M m. musculus* is an outlier due to copy number gains on the X rather than the Y chromosome. However, in inter-subspecific crosses between strains of *M. m. domesticus* and *M. m. musculus* origin, *Sly* is more highly expressed from the *musculus* than the *domesticus* Y chromosome in the testis — opposite from the predicted direction of effect based on copy number — and the relative expression of *Slx* versus *Sly* is independent of X chromosome genotype (**Figure 6.13**). These observations demonstrate that the relationship between DNA copy number and transcript abundance is not so simple as predicted by the conflict hypothesis. Nor do we detect any local reduction in X-linked diversity in the vicinity of co-amplified gene families as would be predicted if they are targets of recent strong selection.

Several biases are possible in these analyses. Quantification of DNA copy number is based on alignment to a reference sequence assembled from C57L/6J (*M. m. musculus* Y chromosome and predominantly *M. m. domesticus* X chromosome). To the extent that sequence differences between reads and reference may hamper alignment, we might underestimate copy number in samples more diverged from the reference sequence. At least for the *Slx* and *Sly* families, this bias is minimal: we obtain a tight linear relationship from *M. spretus* (average divergence $1-2\%$) through *M. m. domesticus* (average divergence $< 0.3\%$), and none of the samples under consideration are phylogenetically close to classical inbred strains (see **Chapter 2**). Moreover, our results are qualitatively similar to direct cytological observations in wild *M. spretus* and *M. musculus*[412]. It is equally unlikely that estimates of gene expression are biased by sequence divergence: members of the *Slx* and *Sly* have pairwise identity approximately 90% on average[306], and our results of all expression analyses were robust to the choice of k-mer size in the quantification algorithm (not shown, but see **§ 6.5**.)

Taken together, our population-genetic and functional analyses provide at best weak support for the idea that the X and Y chromosomes in *Mus* are engaged in ongoing intragenomic conflict. Others have reported deviations in the census sex ratio in favor of males in areas of the *M. m. domesticus*-

M. m. musculus hybrid zone where *musculus* Y chromosomes have introgressed into *domesticus* territory[405] but not in analogous laboratory crosses[143]. We cannot exclude the possibility that some level of X-Y conflict persists — in fact, this seems likely to have shaped the sex chromosomes along the rodent lineage — but it must be mediated by factors other than simply copy number in house mice. Neutral demographic factors are sufficient to explain reductions in sex chromosome diversity relative to autosomes. Nonetheless, the correlation in copy number of X- and Y-linked families (excepting *M. m. musculus*) is difficult to explain by neutral processes and remains mysterious.

6.4 Conclusions and future directions

In this chapter we have presented the first comprehensive survey of Y-linked sequence variation in the house mouse. We find that sequence diversity is markedly reduced not only on the Y but also on the X chromosome relative to autosomes and detect a skew in the site-frequency spectrum towards rare alleles. Demographic modelling and theory point to a strong population bottleneck as the likely cause. Mouse sex chromosomes vary widely in copy number of ampliconic genes with roles in spermiogenesis, but we find limited evidence for intragenomic conflict at the level of copy number or gene expression in the testis.

Although we have documented large variation in copy number of ampliconic, mouse-specific gene families on Yq in natural mouse populations, we can say little about their higher-order organization. Nor can we determine how many gene copies in each family retain coding potential. Addressing these questions will require alternative sequencing technologies that provide longer reads and long-range physical linkage information. Which of the many copies of ampliconic gene families on X and Y are functionally equivalent, and the consequences of sequence and structural variation of particular copies for male reproductive traits, remain open questions.

The deeper evolutionary origins of ampliconic gene families in mouse also remain to be investigated. The oldest of the ampliconic families, *Sstx/y*, was present as a gametologous pair in the common ancestor of mouse and rat[378], but the *Slx/y* and *Srsx/y* arose since the divergence from rat. A broader survey of rodent Y chromosomes would provide valuable context for the evolutionary trajectory of the mouse Y. Finding multiple additional examples of co-amplification of X- and Y-linked sequence would bolster the argument that intragenomic conflict has a prominent role in the evolution of mammalian sex chromosomes.

6.5 Materials and methods

6.5.1 Alignment and variant-calling

Whole-genome sequencing reads were aligned to the mm10 reference sequence using `bwa mem` v0.7.15-r1140[274] with default parameters. Optical duplicates were marked using `samblaster` and excluded from downstream analyses. Regions of the Y chromosome accessible for variant calling were identified using the `CallableLoci` tool in the `GATK` v3.3-0-g37228af[413]. To be declared "callable" within a single sample, sites were required to have depth consistent with a single haploid copy ($3 < \text{depth} < 50$) and $< 25\%$ of overlapping reads having mapping quality (MQ) zero. The analysis was restricted to Yp. The final set of callable sites was defined as any site counted as callable within > 10 samples. In total, $2\,289\,336$ bp (77% of the non-gap length of Yp) were deemed callable.

SNVs and short indels in callable regions were identified using `freebayes` v1.0.2[323]. Variants were called in all samples jointly. Reads with MQ < 10, basecall quality < 13 and > 9 mismatches (to the reference sequence) were excluded. Candidate variant sites were required to have read depth > 3 and at most 3 alleles. The raw call set was filtered to have quality score > 30 and per-sample depth > 3, all heterozygous genotypes were treated as missing to reflect the haploid nature of the Y.

Filtered variants were normalized to their atomic SNV or indel representation using `vcflib`. Functional consequences were predicted using `SnpEff`[414] using the most recent annotation database available (GRCm38/Ensembl 82).

6.5.2 Size estimation of co-amplified regions of Yq and X

The tandem repeats of Yq remain incompletely represented in the mm10 reference genome. To obtain the best possible quantification of sequencing coverage on Yq, all unmapped reads and reads mapping to mm10 Y were re-aligned to the Y chromosome contig of[306] using `bwa mem` with default parameters. Coverage was estimated over all reads, regardless of mapping quality, in each of the "red", "yellow", "blue" and "grey" blocks in Figure 3 and Table S4 of[306]. Read counts were normalized against a region of the X chromosome (chrX: $68.6 - 68.7$ Mb, containing the gene *Fmr1*) known to be present in a single haploid copy in all samples in the study. (This normalization implicitly accounts for mapping biases due to divergence between the target sample

and the reference genome, provided the X and Y chromosomes diverge at roughly equal rates.) To estimate the total size of co-amplified regions of Yq we simply calculated the weighted sum of normalized coverage in the "red", "yellow" and "blue" blocks.

Copy number for X-linked ampliconic genes was estimated in similar fashion. The regions of the X chromosome for each sequence family was taken from[25] (lifted over to mm10) and boundaries were trimmed by manual inspection against segmental duplications identified in the mm10 X sequence. Using BLAST searches we identified the *Spin2* family – with members in several clusters on the proximal X chromosome – as *Sstx*, and included X-linked *Spin2* paralogs in our abundance estimates for *Sstx*.

6.5.3 Estimation of site frequency spectra

Site frequency spectra (SFS) for the Y chromosome were calculated from genotype likelihoods at callable sites using `ANGSD` v0.910-133-g68dd0f2[391]. Genotype likelihoods for the Y chromosome were calculated under the `GATK` haploid model after applying base alignment quality (BAQ) recalibration with the recommended settings for bwa alignments (`-baq 1 -c 50`). Only reads with MQ > 20 and bases with call quality > 13 were considered. Sites were filtered to have per-individual coverage consistent with the presence of a single haploid copy ($3 <$ depth < 40), and to be non-missing in at least 3 individuals per population. Site-wise allele frequencies were computed within each population separately, and the joint SFS across the three populations was estimated from these frequencies. The consensus genotype over 5 *M. spretus* males was used as the ancestral sequence to polarize alleles as ancestral or derived. For estimating uncertainties in diversity statistics, 100 bootstrap replicates were obtained for the joint SFS.

SFS for the X chromosome were estimated using the same parameters but with the consensus haploid genotype from a single *M. caroli* female as the ancestral sequence. For the mitochondria, different filtering criteria were used ($10 <$ depth < 1000) to reflect differences in expected coverage for this organellar genome. *M. caroli* was again used as the ancestral sequence. For estimating the autosomal SFS we used sequence from chromosome 1 and used a diploid rather than haploid model for genotype likelihoods.

Some inconsistencies may arise due to the use of different outgroup species, at different evolutionary distances, for the autosomes, X, Y and mitochondria. We unfortunately did not have access to whole-genome sequence from a male more divergent than *M. spretus* to use as an outgroup for

the Y. However, because hybrid offspring of a *M. musculus* dam and a *M. spretus* sire are generally sterile[132], there is little change of introgression of a *M. spretus* Y chromosome into *M. musculus*. Nor did we find evidence for incomplete lineage sorting of Y chromosomes between *M. spretus* and *M. musculus* in our dataset.

6.5.4 Diversity statistics

Diversity statistics and neutrality tests were calculated from joint SFS using the `R` package `sfsr` (`http://github.com/andrewparkermorgan/sfsr`)[1]. Hudson-Kreitman-Aguade (HKA) tests were performed with `sfsr` and p-values obtained from the χ^2 distribution with a single degree of freedom as suggested in[392]. (Results were checked against the `HKA` software from Jody Hey, in which significance thresholds are set via coalescent simulations; all significant tests were significant under both methods.)

6.5.5 Demographic inference

Possible demographic scenarios for male lineages in *M. musculus* were explored using approximate Bayesian computation (ABC). All scenarios modelled three populations (corresponding to *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*) derived from a single ancestral population. The order of population splits was (castaneus,(musculus,domesticus)) — reflecting the phylogeny in **Figure 6.4** — and was kept fixed across all scenarios. Eight scenarios were tested: (I) constant population size, no migration; (II) recent bottleneck shared by *M. m. domesticus* and *M. m. musculus*; (III) recent bottleneck, followed by exponential growth; (IV) distant bottleneck, followed by exponential growth; (V) constant population size, with migration; (VI) exponential growth at independent rates, no migration; (VII) recent bottleneck, with migration; (VIII) exponential growth, with migration.

Briefly, $100,000$ simulations were performed for each model using parameter values drawn from uninformative or weakly-informative prior distributions. Fifteen summary statistics were calculated from the joint SFS generated by each simulation: number of segregating sites in each population (3); Tajima's D, Fu and Li's D and F in each population (9); and F_{st} between all population pairs (3). The same set of statistics was computed for the observed joint SFS. The 0.1% of simulations with smallest Euclidean distance to the observed summary statistics were retained.

[1] I am the sole author of the `sfsr` package, which may eventually be described in more detail elsewhere.

Posterior distributions were computed via kernel smoothing over the parameter values of the retained simulations using the `R` package `abc`[415].

Models were compared via their Bayes factors, calculated using the `postpr()` function in the `abc` packge. To confirm the fidelity of the best-fitting model, summary statistics for pseudo-observed datasets (*i.e.* simulations from the posterior disributions) were checked against the observed summary statistics.

6.5.6 Analyses of gene expression

Hybrid zone mice. Turner *et al.*[151] measured gene expression in whole testes of 175 hybrid mice using the Agilent 4x44k Whole Mouse Genome Microarray. Raw hybridization data was obtained from NCBI Gene Expression Omnibus (GSE61417). Arrays were pre-processed using the `R` package `Agi4x44PreProcess` and companion annotation package `mgug4122a.db`. Because the array was designed prior to incorporation of most of the Y chromosome sequence into the reference genome, we re-mapped all probe sequences to the current reference genome (mm10) using `blat` to identify Y-linked probes. A total of 23 probes were mapped to the Y. These probes were assigned to their overlapping genes in Ensembl 85[115]. Genes targeted by probes but listed as "predicted" (with symbols "GmXXXX") were gathered into one of the co-amplified gene families on the Y via the "mmusculus_paralog_ensembl_gene" field in the "mmusculus_gene_ensembl" table of the Ensembl Biomart.

Mice in this study were genotyped using the Mouse Diversity Array. Y chromosome haplogroups (*musculus* or *domesticus*) were assigned by performing PCA on 35 Y-linked probes.

To test for differential gene expression between Y haplogroups we used the empirical Bayes procedure implemented in the `R` package `limma`[416]. False discovery rates were calculated using the Benjaimini-Hochberg method[417].

Reciprocal F1 hybrids. Mack *et al.*[401] measured gene expression in whole testes of three males from each of four F_1 crosses — CZECHII/EiJ×PWK/PhJ; LEWES/EiJ×PWK/PhJ; PWK/PhJ×LEWES/EiJ; and WSB/EiJ×LEWES/EiJ — using RNA-seq. Reads were retrieved from NCBI Short Read Archive (PRJNA286765). Transcript-level expression was estimated using `kallisto`[300] using the Ensembl 85 transcript catatlog augmented with all *Slx/y*, *Sstx/y* and *Srsx/y* transcripts identified in[306]. In the presence of redundant transcripts (*i.e.* from multiple copies of a co-amplified gene family), `kallisto` uses an expectation-maximization algorithm to distribute the "weight" of each read

across transcripts without double-counting. Transcript-level expression estimates were aggregated to the gene level for differential expression testing using the `R` package `tximport`. As for the microarray data, "predicted" genes (with symbols "GmXXXX") on the Y chromosome were assigned to a co-amplified family where possible using Ensembl Biomart.

Gene-level expression estimates were transformed to log scale and gene-wise dispersion parameters estimated using the `voom()` function in the `R` package `limma`. Genes with total normalized abundance (length-scaled transcripts per million, TPM) < 10 in aggregate across all samples were excluded. Differential expression testing was performed in the same way as for the microarray data.

To compare the relative expression levels of X- and Y-linked members of co-amplified gene families, we defined the "expression ratio" $z = \frac{y}{(x+y)}$ and transformed it to a log-odds "expression score" R:

$$R = \log \frac{z}{1-z}$$

The standard error of this quantity was calculated within each cross and within each gene family by the delta method[418] as implemented in the `deltamethod()` function of `R` package `msm`.

CHAPTER 7

Concluding remarks

In this thesis I have described several lines of investigation into three basic forces governing the level and distribution of genetic variation in populations: recombination, mutation and intragenomic conflict.

When I began my graduate studies, I was originally interested in the role of paternal age in genomic instability and its potential contribution to offspring risk for psychiatric disease in humans. Large epidemiological studies had demonstrated modest but robust correlation between incidence of psychiatric diseases and advanced paternal age[419,420,421,422]. Case-control and pedigree studies had established rare and *de novo* mutations, especially CNVs, as a risk factor for schizophrenia, autism and developmental delay[423,424,425,426,427]. This raised the possibility that an age-related increase in the burden of mutations transmitted by older fathers might provide the casual link between paternal age and offspring risk for polygenic diseases in addition to Mendelian syndromes with a firmly-established age effect[66,70]. However, the accumulating evidence from pedigrees, twin studies and population-based association studies no longer supports this hypothesis[428]. The narrow-sense heritability of diseases such as schizophrenia ($h^2 \approx 0.7$) alone constrains the fraction of cases that can be attributed to *de novo* mutations in the absence of underlying inherited liability. Given empirical estimates of disease prevalence and fecundity of affected men, age-related mutations are expected to contribute only about 10% of disease risk even if the mutation rate is high and the resulting alleles have high penetrance. Excess risk in offspring of older fathers is best explained by a genetic correlation between liability to psychiatric disease and late fatherhood[428].

It is fortunate, then, that my early interest in the effects of paternal age on mutation motivated a more general inquiry into variation in the rates of mutation and recombination along the genome, and how that variation is influenced by functional differences between the male and female germline. Sex effects on the point mutation rate and some aspects of recombination have received much theoretical attention and are well known in humans and other primates[26,76]. However,

most analyses — especially those that rely on high-throughput sequencing — have been restricted to "well-behaved" unique genomic sequence. The approximately 10% of mammalian genomes occupied by segmental duplications and other repetitive sequences is more poorly characterized but is also the most polymorphic within and between species. I chose to focus on that portion of the genome and to use the unique resources available for the house mouse — a well-annotated genome assembly; numerous inbred strains; randomized mapping populations like the Collaborative Cross and Diversity Outbred; and a diverse collection of wild mice from around the globe — to overcome technical obstacles that hamper studies of duplicated sequences in humans. I have been exceedingly fortunate to collaborate with many generous students, postdocs and senior scientists, without whom little of this work would have been possible.

In this concluding chapter I summarize the key findings from these studies in the context of our understanding of genome evolution in mammals. Many of these topics are treated in greater depth in previous chatpers; the reader is directed to those pages for further details.

7.1 Recombination in the male germline

In **Chapter 3** I used the CC and DO to replicate well-known differences between the overall rate and genomic distribution of (crossover) recombination between males and females. Although the excess of recombination near the telomeric end of chromosomes in the male germline has been observed repeatedly, our study is, to my knowledge, the first to connect this pattern to the temporal ordering of events in male meiosis. We observed that recombinant chromosomes tend to have a crossover near the distal end regardless of the total number of crossovers on the chromosome. We conclude that the position of this crossover is regulated. Since chromosome pairing and synapsis proceed from the telomere towards the centromere in males, beginning with the tethering of telomeres to the nuclear envelope early in the first meiotic prophase (the "bouquet"), we predict that crossovers — whose formation is dependent on the scaffold provided by the synaptonemal complex — are formed telomeric-end first also. Crossover patterns in other species suggest that this is the ancestral program and that the more uniform distribution of crossovers in female meiosis evolved along with the lengthening of female prophase.

Two other findings are completely novel. The first is the discovery of one or more modifiers of the male recombination rate on the Y chromosome. The presence of recombination modifiers on

both the X and Y chromosomes emphasizes the importance of the sex chromosomes in the function of the male germline and is further evidence of the close link between recombination and speciation. Although it is well-known that hybrid incompatibilities accumulate faster on sex chromosomes than autosomes, most attention to date has focused on the X chromosome. Y chromosomes have an important role in hybrid incompatibility in *Drosophila*[429] but their contribution in mouse is less clear[260]. Theory predicts that male-advantageous alleles should be preferentially Y-linked. How changes in the overall rate of recombination might be related to male fitness remains to be seen.

The second novel finding is that crossovers are strongly suppressed near large structural variants. We observed the effect first in mouse but showed that it applies in dogs as well, and likely is a general feature of mammalian meiosis. Although simple copy-number variation is useful as a proxy for coldspots, copy number alone is a surprisingly weak predictor of crossover patterns at any given CNV locus. We hypothesize that what we observe as CNVs are in fact more complex structural rearrangements. In the same way that crossovers are suppressed within cytologically-visible inversions, megabase-sized structural variants are likely an obstacle to normal pairing and synapsis[430]. But this leads to a paradox: the same regions that are coldspots for recombination are thought to sprout new alleles by illegitimate recombination. We predict that DSBs in coldspots are preferentially repaired either from paralogous sequence on the same chromatid, or from the sister chromatid, rather than from the homolog. Erroneous intrachromatid or sister-chromatid recombination has a broader mutational spectrum than recombination with the homolog but never produces crossovers.

Further study will be required to understand the population-genetic significance of coldspots. In theory, coldspots might promote the formation of selfish elements such as segregation distorters. Background selection is expected to be somewhat stronger in the vicinity of coldspots, and genetic diversity should consequently be diminished. On the other hand, the net result of these forces — a reduction in local effective population size — should hasten the loss of structural alleles at the coldspot itself. If coldspots are cold because of structural heterozygosity, they may be prone to lose their "coldness" by drift.

7.2 Structural variation and the "last frontier" of mammalian genomes

The preponderance of long duplicated sequences (segmental duplications, SDs) in mammalian genomes was one of the first and most striking observations of the genome-sequencing era[86,41]. It

was recognized almost immediately that these regions are also much more variable in populations than the rest of the genome[42,43]. But, owing to the technical challenges of studying repeated sequences, especially using high-throughput shotgun methods, SDs and their associated structural variants remain poorly characterized. There are exceptions: the human HLA locus (formally the "major histocompatibility complex", MHC) is highly polymorhpic but has been analyzed in detail because of its importance for susceptibility to dozens of common and rare diseases[431]. The idea that most SDs are neutral genomic flotsam has been challenged by careful studies of several loci in humans[432,433,434,33]. These studies have consistently revealed more structural complexity and higher levels of both sequence and structural polymorphism than what was estimated from shotgun sequencing. I have shown in **Chapter 3, Chapter 4** and **Chapter 6** that large structural variants and the genes they may carry have important effects on meiosis and gametogenesis in mouse. The work of characterizing these loci will benefit greatly from continued improvements in "third-generation" sequencing technologies that yield reads in the $10 - 100$ kb range and from the application of high-throughput physical mapping of megabase-sized fragments.

Deeper understanding of structural variation will also benefit interpretations of other classes of variation. Loci that have experienced cycles of duplication and loss — the "genomic revolving door" discussed in **Chapter 4** — have idiosyncratic patterns of polymorphism if diverged copies of an ancestral sequence have fixed along different lineages. The resulting excess of variation may be falsely interpreted as evidence for balancing selection, population differentiation or other non-neutral processes.

Even for intensively-studied organisms such mouse and human there remain portions of the genome whose sequence and structure are almost completely unknown. Cytological studies in mouse from the era before genome sequencing identified a number of gross chromatin features that are variable in populations and useful for phylogeny but have no direct relationship to a specific locus in a reference assembly[435,59,436,58]. The most prominent example is centromeres: mammalian centromeres, unlike those of yeast, are defined by their functional properties rather than their sequence. Kinetochore proteins assemble into centromeres at large tracts of so-called "satellite sequences," tandem arrays of hundreds to thousands of copies of $50 - 60$ bp repeating unit[437]. Centromeric sequences are heterochromatic[438] but clearly not functionally inert, as they mediate faithful segregation of chromosomes during both mitosis and meiosis. Although the degree of

population-level sequence variation in centromeric satellite arrays is little understood in mouse, human or any other vertebrate, variation in centromere "strength" influences segregation ratios in females[61]. Characterizing the landscape of centromeric sequence variation in a tractable model system such as mouse would be a major step forward in understanding karyotype evolution.

The combination of whole-genome sequencing with a segregating population like the CC or DO is a powerful approach for mapping repetitive sequences. Provided a locus is variable in some way that can be detected in WGS reads, QTL mapping can be applied to determine both its physical location(s) and the founder strains in which it is present. As proof of principle, I show in **Figure 7.1** that a satellite repeat motif identified by *de novo* assembly of CAST/EiJ sequence [1] is found exclusively in DO mice that inherit proximal chromosome 2 from CAST/EiJ. I have mapped more than $2,000$ satellite repeat motifs (not necessarily independent) in this way and have found that about half are variable in copy number among the CC founders, and 15% are private to a single strain. Those that can be assigned to a chromosome are differentiated by population in wild mice (data not shown.) Further investigation is warranted to identify chromosome- or lineage-specific centromeric satellites and to characterize their higher-order organization.

7.3 Genetic conflict, structural variation and the sex chromosomes

Two types of genetic conflict have been discussed in this thesis: hybrid incompatibilities between loci that lead to speciation (such as the recombination-modifing *Prdm9* and *Hstx1* loci in mouse, **Chapter 3**) and intragenomic conflict between selfish alleles competing for transmission at meiosis (**Chapter 5** and **Chapter 6**). Several authors have proposed that the former type of conflict and its ultimate consequence, reproductive isolation, are a side effect of the latter[439]. The sex chromosomes are uniquely vulnerable to intragenomic conflict because they are heteromorphic and non-recombining. But sex-ratio drive should prompt the rapid increase of suppressors[440]. If these suppressors have pleitropic effects on reproduction, the result should be hybrid sterility in the heterogametic sex (Haldane's rule[255]) and an excess of sex-linked sterility loci. Both predictions are fulfilled in the mouse and the mechanistic basis of male hybrid infertility has been studied in detail[142]. No active sex-ratio driver has been identified in mouse, but traces of one may exist on the

[1] Thomas Keane, personal communication

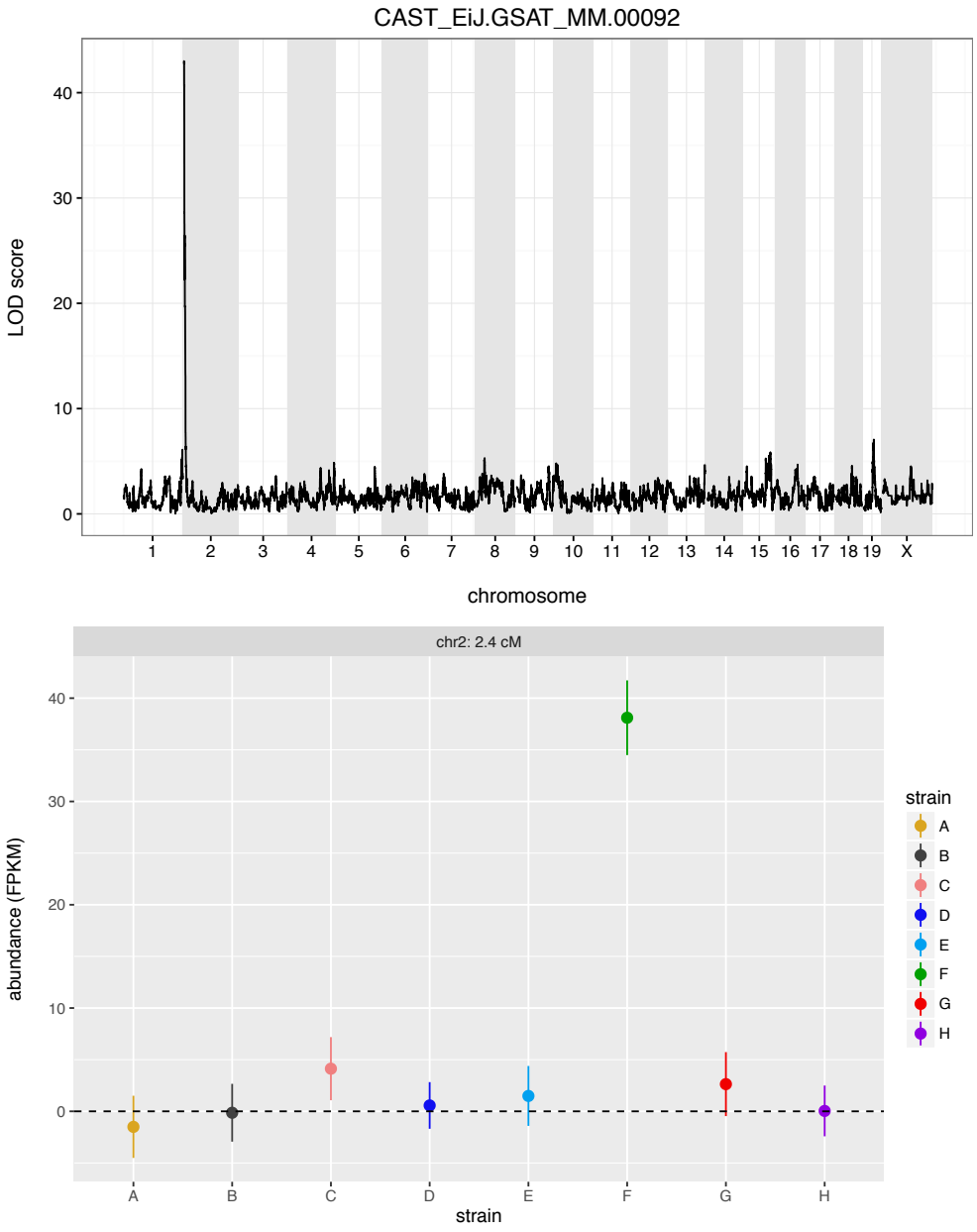

Figure 7.1: A strain-specific centromeric repeat in CAST/EiJ. Abundance of a short sequence from the mouse γ-satellite family was quantified in WGS reads from 228 DO mice. Variation in abundance maps exclusively to the centromeric end of chromosome 2 (top panel); the sequence is only present in mice inheriting that region from CAST/EiJ.

X and Y chromosomes — discussed at length in **Chapter 6**.

Both autosomal and X-linked transmission distorters tend to be two-component systems in which a *trans*-acting *distorter* influences transmission at a second *responder* locus that may be sensitive or resistant. The system is most successful, in the Darwinian sense, if the *distorter* and (resistant) *responder* are tightly linked. Recombination is an important defense against such systems, and so it is perhaps not surprising that modifiers of the rate and distribution of recombination have a central role in hybrid sterility.

Meiotic drive would appear to be an exception to the above rules since it is limited in mammals to the homogametic sex (*i.e.* females) — but somewhat counterintuitively, theory predicts that the male germline should evolve to promote fairness in female meiosis too[441]. The work presented in **Chapter 5** is one of few examples of transmission distorters in mammals, and (to my knowledge) is the only example of a meiotic drive allele that can sweep to fixation despite strong effect on fitness in the heterozygote. The rise of $R2d2^{HC}$ alleles is indistinguishable from a true selective sweep on a beneficial mutation, and without the tightly-controlled breeding scheme of the DO we would have been unlikely to unmask it as a selfish allele. This urges caution in the interpretation of scans for selection in population samples. There is no reason to believe *a priori* that intragenomic conflict should be any more rare than true positive selection.

Heterochromatin is a recurring theme in selfish systems across the tree of life[442]. Failure to properly maintain or segregate heterocrhomatic regions underlies the homogeneously-staining region (HSR) meiotic drive locus in *M. m. domesticus*[443], knobs in maize[444,445] and multiple sex-ratio drivers in *Drosophila*[446]. The rodent-specific genes on the long arm of the mouse Y chromosome (**Chapter 6**) and their X-gametologs all function in maintaining the heterochromatin-like state of meiotic sex chromosome inactivation (MSCI)[411,17,447]. MSCI and its more general form, meiotic silencing of unsynapsed chromatin, have analogs in other taxa[448], suggesting that they may be an important defense against transcription from selfish elements of many kinds — from transmission distorters to transposons — in the germline.

♢ ♢ ♢

Taken together, the work presented in this thesis underscores the ways in which the content and

organization of mammalian genomes is influenced by the biology of the male and female germlines, and the pervasive influence of intragenomic conflict in shaping genetic variation in populations.

APPENDIX A

On the number of observable meioses in the Diversity Outbred

A.1 On the number of observable meioses in the Diversity Outbred

Consider a randomly-mating diploid population of size N propagated in k non-overlapping generations. At generation k, there exist $2Nk$ meioses in the history of the population: one representing each of the two gametes giving rise to each individual. Define a meiosis in generation $1 \leq j \leq k$ to be *observable* at generation t if one of its products is segregating in the population at generation $j \leq t \leq k$. Define a meiosis to be *observed* at generation t if it is observable at generation t and is present in a sample of n_t individuals drawn randomly and with replacement from the population at generation t. Finally, let a meiosis be *uniquely observed* at generation t if it is observed at generation t but not at generation $i \leq t$, given sample sizes $n_j, \ldots, n_{t-1}$.

The set of meioses uniquely observed at each $t \in j, \ldots, k$ are clearly independent. Let U_{jt} be the probability that a meiosis at generation j is uniquely observed at generation $t \leq k$. Then the probability P_{jk} that it is observed by generation k is

$$P_{jk} = U_{jk} \prod_{t=j}^{k-1} 1 - U_{jt}$$

That is, P_{jk} is the probability that a meiosis at generation j is not observed *until* generation k.

The quantities U above arise via two independent stochastic processes: the inheritance process and the process of randomly sampling chromosomes at each generation. We model inheritance as a first-order Markov chain using the standard Wright-Fisher process. The chain has $2N + 1$ states representing allele counts $0, \ldots, 2N$; the first and last of these are absorbing states representing loss and fixation, respectively. Let $\mathbf{f}_t$ be a vector of size $2N + 1$ representing the state occupancy distribution at time t. The chain's transition matrix $\mathbf{A}$ has entries a_{qr} of the form

$$a_{qr} = \binom{2N}{r} (f_{tq})^r \, (1 - f_{tq})^{2N-r}$$

where $q \in 0, \ldots 2N$ represents the state at time t and r the state at time $t + 1$.

The chain is initialized at $\mathbf{f}_j = (0, 1, \ldots, 0)$ since a (product of) meiosis enters the population in

a single copy. The state distribution at generation k is then

$$\mathbf{f}_k = \mathbf{f}_j \mathbf{A}^{k-j}$$

The sampling process is independent of the inheritance process. Let W_{jt} be the probability of *not* observing a (product of) meiosis arising at generation j at generation $j \leq t \leq k$, conditional on its frequency at generation t and the sample size n_t:

$$W_{jt} = \sum_{i=0}^{2N} f_{ti}(1 - f_{ti})^{n_t}$$

Now, because samples drawn from different generations are independent, we can obtain the value of U_{jt}:

$$U_{jt} = W_{jt} \prod_{i=j}^{t-1} 1 - W_{ji}$$

and, hence, finally, the value of P_{jk}.

Because meiosis is independent across generations, the P_{jk} are disjoint for all $1 \leq j \leq k$. We can therefore estimate the number of observed meioses M in a multi-generational sample with sizes $\mathbf{n} = (n_j, \ldots, n_k)$ as

$$M = 2N \sum_{i=j}^{k} P_{ji} + 2 \sum_{i=j}^{k} n_i$$

where the first term gives the expected number of observed meioses inherited from the main pedigree, and the second term gives the two fully-observed meioses in each sampled individual.

In practice, the second term in the sum dominates. **Figure A.1** plots values of P_{jk} against $(j - k)$, given the population size of the DO ($N = 350$) and the empirical sample sizes at each generation. The rate of decay is approximately 95% per generation. Note that this does not imply that the majority of observed crossovers are unique: products of meioses in later generations transmit more inherited crossovers relative to new crossovers, as we see in the next section.

A.2 On the accumulation of recombination events in the Diversity Outbred

Every individual is the product of exactly two meioses (for the autosomes). Each meiotic product transits r new crossovers, plus one-half of any inherited crossovers. We can write the following

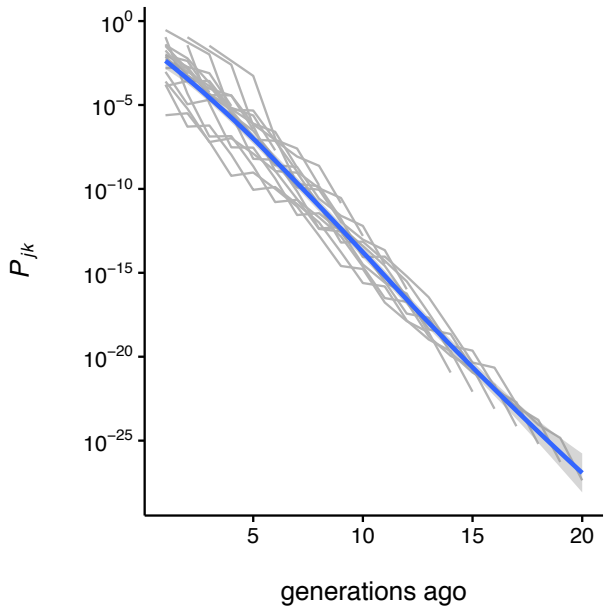

Figure A.1: Values of P_{jk} are plotted against $(j-k)$, the number of generations ago a meiosis occurred relative to to the generation of sampling. Each grey line represents one (j,k) pair; the line of best fit is superposed in blue.

recursion for the expected number of crossovers C_t per genome as a function of generation t:

$$C_t = 2r + \frac{1}{2}p_1 + \frac{1}{2}p_2$$

where p_1 and p_2 are the number of inherited crossovers in the genomes of an individuals two parents. For simplicity assume that $p_1 = p_2 = C_{t-1}$ and that the recombination rate r is constant in time and uniform across the population. Then

$$C_t = 2r + C_{t-1}$$

and we can take $C_0 = 0$, since the founders of the DO were inbred strains and we measure crossovers with respect to these founder haplotypes. Hence

$$C_t = 2r + (2r + (2r + \dots)) = 2rt$$

and the expected number of crossovers per genome is just a linear function of generation.

REFERENCES

1. Handel, M. A. & Schimenti, J. C. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat Rev Genet* **11**, 124–136 (2010). URL http://www.nature.com/nrg/journal/v11/n2/full/nrg2723.html.

2. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).

3. Chakravarti, A. *et al.* Nonuniform recombination within the human beta-globin gene cluster. *Am. J. Hum. Genet.* **36**, 1239–1258 (1984).

4. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* **4**, 587–597 (2003). URL http://www.nature.com/nrg/journal/v4/n8/full/nrg1123.html#B11.

5. Hughes, J. F. & Page, D. C. The Biology and Evolution of Mammalian Y Chromosomes. *Annual Review of Genetics* **49**, 507–527 (2015). URL http://dx.doi.org/10.1146/annurev-genet-112414-055311.

6. Berta, P. *et al.* Genetic evidence equating SRY and the testis-determining factor. *Nature* **348**, 448–450 (1990). URL http://www.nature.com/nature/journal/v348/n6300/abs/348448a0.html.

7. Eicher, E. M. & Washburn, a. L. L. Genetic Control of Primary Sex Determination in Mice. *Annual Review of Genetics* **20**, 327–360 (1986). URL http://dx.doi.org/10.1146/annurev.ge.20.120186.001551.

8. McLaren, A. Primordial germ cells in the mouse. *Developmental Biology* **262**, 1–15 (2003). URL http://www.sciencedirect.com/science/article/pii/S0012160603002148.

9. Hilscher, B. *et al.* Kinetics of gametogenesis. I. Comparative histological and autoradiographic studies of oocytes and transitional prospermatogonia during oogenesis and prespermatogenesis. *Cell Tissue Res.* **154**, 443–470 (1974).

10. Drost, J. B. & Lee, W. R. Biological basis of germline mutation: Comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environ. Mol. Mutagen.* **25**, 48–64 (1995). URL http://onlinelibrary.wiley.com/doi/10.1002/em.2850250609/abstract.

11. Hassold, T., Hall, H. & Hunt, P. The origin of human aneuploidy: where we have been, where we are going. *Hum. Mol. Genet.* **16**, R203–R208 (2007). URL http://hmg.oxfordjournals.org/content/16/R2/R203.

12. Lamb, N. E., Sherman, S. L. & Hassold, T. J. Effect of meiotic recombination on the production of aneuploid gametes in humans. *Cytogenet. Genome Res.* **111**, 250–255 (2005).

13. Villena, F. P.-M. d. & Sapienza, C. Recombination is proportional to the number of chromosome arms in mammals. *Incorporating Mouse Genome* **12**, 318–322 (2001). URL http://link.springer.com/article/10.1007/s003350020005.

14. Pardo-Manuel de Villena, F. & Sapienza, C. Nonrandom segregation during meiosis: the

unfairness of females. *Mamm. Genome* **12**, 331–339 (2001).

15. Ottolini, C. S. *et al.* Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. *Nat Genet* **47**, 727–735 (2015). URL http://www.nature.com/ng/journal/v47/n7/full/ng.3306.html.

16. Clermont, Y. Kinetics of spermatogenesis in mammals: seminiferous epithelium cycle and spermatogonial renewal. *Physiol. Rev.* **52**, 198–236 (1972).

17. Royo, H. *et al.* Evidence that Meiotic Sex Chromosome Inactivation Is Essential for Male Fertility. *Current Biology* **20**, 2117–2123 (2010). URL http://www.cell.com/current-biology/abstract/S0960-9822(10)01435-1.

18. Turner, J. M. A. *et al.* Silencing of unsynapsed meiotic chromosomes in the mouse. *Nat. Genet.* **37**, 41–47 (2005).

19. Turner, J. M. A. Meiotic sex chromosome inactivation. *Development* **134**, 1823–1831 (2007). URL http://dev.biologists.org/content/134/10/1823.

20. McKee, B. D. & Handel, M. A. Sex chromosomes, recombination, and chromatin conformation. *Chromosoma* **102**, 71–80 (1993).

21. Lyon, M. F. Transmission ratio distortion in mice. *Annu. Rev. Genet.* **37**, 393–408 (2003).

22. Lyon, M. F. Transmission ratio distortion in mouse t-haplotypes is due to multiple distorter genes acting on a responder locus. *Cell* **37**, 621–628 (1984).

23. Hammer, M. F., Schimenti, J. & Silver, L. M. Evolution of mouse chromosome 17 and the origin of inversions associated with t haplotypes. *Proc Natl Acad Sci U S A* **86**, 3261–3265 (1989). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC287110/.

24. Jaenike, J. Sex Chromosome Meiotic Drive. *Annual Review of Ecology and Systematics* **32**, 25–49 (2001). URL http://dx.doi.org/10.1146/annurev.ecolsys.32.081501.113958.

25. Mueller, J. L. *et al.* The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nat. Genet.* **40**, 794–799 (2008).

26. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15**, 47–70 (2014).

27. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). URL http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html.

28. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**, 822–826 (2015). URL http://www.nature.com/ng/journal/v47/n7/full/ng.3292.html.

29. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011). URL http://www.nature.com/nature/journal/v477/

n7364/full/nature10413.html.

30. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**, 435–445 (2004). URL http://www.nature.com/nrg/journal/v5/n6/full/nrg1348.html.

31. Dallas, J. F. Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. *Mamm. Genome* **3**, 452–456 (1992).

32. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat Genet* **44**, 1161–1165 (2012). URL http://www.nature.com/ng/journal/v44/n10/full/ng.2398.html.

33. Huddleston, J. & Eichler, E. E. An Incomplete Understanding of Human Genetic Variation. *Genetics* **202**, 1251–1254 (2016). URL http://www.genetics.org/content/202/4/1251.

34. Itsara, A. *et al.* De novo rates and selection of large copy number variation. *Genome Res* **20**, 1469–1481 (2010). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2963811/.

35. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet* **29**, 575–584 (2013). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3785239/.

36. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).

37. Perry, G. H. *et al.* Hotspots for copy number variation in chimpanzees and humans. *PNAS* **103**, 8006–8011 (2006). URL http://www.pnas.org/content/103/21/8006.

38. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).

39. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat Genet* **47**, 296–303 (2015). URL http://www.nature.com/ng/journal/v47/n3/abs/ng.3200.html.

40. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).

41. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14**, 789–801 (2004).

42. Sharp, A. J. *et al.* Segmental Duplications and Copy-Number Variation in the Human Genome. *Am J Hum Genet* **77**, 78–88 (2005). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1226196/.

43. She, X., Cheng, Z., Zöllner, S., Church, D. M. & Eichler, E. E. Mouse Segmental Duplication and Copy-Number Variation. *Nat Genet* **40**, 909–914 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2574762/.

44. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in

genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).

45. Sturtevant, A. H. The Effects of Unequal Crossing over at the Bar Locus in Drosophila. *Genetics* **10**, 117–147 (1925). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1200852/.

46. Smith, G. P. Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528–535 (1976).

47. Turner, D. J. *et al.* The rates of de novo meiotic deletions and duplications causing several genomic disorders in the male germline. *Nat Genet* **40**, 90–95 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2669897/.

48. Waldman, A. S. & Liskay, R. M. Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Mol. Cell. Biol.* **8**, 5350–5357 (1988).

49. Payen, C., Koszul, R., Dujon, B. & Fischer, G. Segmental Duplications Arise from Pol32-Dependent Repair of Broken Forks through Two Alternative Replication-Based Mechanisms. *PLoS Genet* **4** (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2518615/.

50. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).

51. Sturtevant, A. H. A Case of Rearrangement of Genes in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **7**, 235–237 (1921).

52. Bush, G. L., Case, S. M., Wilson, A. C. & Patton, J. L. Rapid speciation and chromosomal evolution in mammals. *Proc Natl Acad Sci U S A* **74**, 3942–3946 (1977). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431793/.

53. Charlesworth, B. The evolution of sex chromosomes. *Science* **251**, 1030–1033 (1991). URL http://science.sciencemag.org/content/251/4997/1030.

54. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).

55. Chinwalla, A. T. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002). URL http://www.nature.com/nature/journal/v420/n6915/abs/nature01262.html.

56. Nellåker, C. *et al.* The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol* **13**, R45 (2012). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3446317/.

57. Forejt, J. Centromeric heterochromatin polymorphism in the house mouse. *Chromosoma* **43**, 187–201 (1973). URL http://link.springer.com/article/10.1007/BF00483378.

58. Britton-Davidian, J., Cazaux, B. & Catalan, J. Chromosomal dynamics of nucleolar organizer regions (NORs) in the house mouse: micro-evolutionary insights. *Heredity* **108**, 68–74 (2012). URL http://www.nature.com/hdy/journal/v108/n1/full/hdy2011105a.html.

59. Agulnik, S., Plass, C., Traut, W. & Winking, H. Evolution of a long-range repeat family in Chromosome 1 of the genus Mus. *Mammalian Genome* **4**, 704–710 (1993). URL http://link.springer.com/article/10.1007/BF00357793.

60. Batchelor, A. L., Phillips, R. J. & Searle, A. G. A comparison of the mutagenic effectiveness of chronic neutron- and gamma-irradiation of mouse spermatogonia. *Mutat. Res.* **3**, 218–229 (1966).

61. Chmátal, L. *et al.* Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Curr. Biol.* **24**, 2295–2300 (2014).

62. Jones, K. L., Smith, D. W., Harvey, M. A., Hall, B. D. & Quan, L. Older paternal age and fresh gene mutation: data on additional disorders. *J. Pediatr.* **86**, 84–88 (1975).

63. Risch, N., Reich, E. W., Wishnick, M. M. & McCarthy, J. G. Spontaneous mutation and parental age in humans. *Am J Hum Genet* **41**, 218–248 (1987). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1684215/.

64. Weinberg, W. Zur vererbung des zwergwuchses. *Arch Rass Ges Biol* **9**, 710–718 (1912). Bibtex: weinberg1912vererbung.

65. Haldane, J. B. S. The Mutation Rate of the Gene for Haemophilia, and Its Segregation Ratios in Males and Females. *Annals of Eugenics* **13**, 262–271 (1946). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1946.tb02367.x/abstract.

66. Crow, J. F. The high spontaneous mutation rate: Is it a health risk? *PNAS* **94**, 8380–8386 (1997). URL http://www.pnas.org/content/94/16/8380.

67. Goriely, A., McVean, G. A. T., Röjmyr, M., Ingemarsson, B. & Wilkie, A. O. M. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science* **301**, 643–646 (2003).

68. Goriely, A. & Wilkie, A. O. Paternal Age Effect Mutations and Selfish Spermatogonial Selection: Causes and Consequences for Human Disease. *Am J Hum Genet* **90**, 175–200 (2012). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3276674/.

69. Project, t. . G. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712–714 (2011). URL http://www.nature.com/ng/journal/v43/n7/full/ng.862.html#ref2.

70. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).

71. Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F. & Makova, K. D. Strong and Weak Male Mutation Bias at Different Sites in the Primate Genomes: Insights from the Human-Chimpanzee Comparison. *Mol Biol Evol* **23**, 565–573 (2006). URL http://mbe.oxfordjournals.org/content/23/3/565.

72. Hehir-Kwa, J. Y. *et al.* De novo copy number variants associated with intellectual disability have a paternal origin and age bias. *J. Med. Genet.* **48**, 776–778 (2011).

73. Wang, B. *et al.* CNV analysis in Chinese children of mental retardation highlights a sex differentiation in parental contribution to de novo and inherited mutational burdens. *Sci Rep* **6**, 25954 (2016).

74. Duyzend, M. H. *et al.* Maternal Modifiers and Parent-of-Origin Bias of the Autism-Associated 16p11.2 CNV. *Am. J. Hum. Genet.* **98**, 45–57 (2016).

75. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015). URL http://genome.cshlp.org/content/25/6/792.

76. Scally, A. Mutation rates and the evolution of germline structure. *Phil. Trans. R. Soc. B* **371**, 20150137 (2016). URL http://rstb.royalsocietypublishing.org/content/371/1699/20150137.

77. Lynch, M. *The Origins of Genome Architecture* (Sinauer Associates Inc, Sunderland, Mass, 2007), 1 edition edn.

78. Gillespie, J. H. *Population Genetics: A Concise Guide* (Johns Hopkins University Press, Baltimore, Md, 2004), 2nd edition edn.

79. Kimura, M. On the Probability of Fixation of Mutant Genes in a Population. *Genetics* **47**, 713–719 (1962). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1210364/.

80. Kimura, M. & Ohta, T. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics* **61**, 763–771 (1969). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1212239/.

81. Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97–159 (1931). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1201091/.

82. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (The Blackburn Press, New Jersey, 1970).

83. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**, 256–276 (1975).

84. Maynard Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genetics Research* **23**, 23–35 (1974). URL https://www.cambridge.org/core/journals/genetics-research/article/hitch-hiking-effect-of-a-favourable-gene/918291A3B62BD50E1AE5C1F22165EF1B#.

85. Green, J. P. *et al.* The Genetic Basis of Kin Recognition in a Cooperatively Breeding Mammal. *Current Biology* **25**, 2631–2641 (2015). URL http://www.sciencedirect.com/science/article/pii/S0960982215010167.

86. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). URL http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html.

87. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).

88. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2788925/.

89. Wang, D. G. *et al.* Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280**, 1077–1082 (1998). URL http://science.sciencemag.org/content/280/5366/1077.

90. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet* **24**, 381–386 (2000). URL http://www.nature.com/ng/journal/v24/n4/full/ng0400_381.html.

91. Shifman, S. *et al.* A High-Resolution Single Nucleotide Polymorphism Genetic Map of the Mouse Genome. *PLOS Biol* **4**, e395 (2006). URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040395.

92. Yang, H. *et al.* A customized and versatile high-density genotyping array for the mouse. *Nat Meth* **6**, 663–666 (2009). URL http://www.nature.com/nmeth/journal/v6/n9/full/nmeth.1359.html.

93. Galbraith, D. W. & Edwards, J. Applications of Microarrays for Crop Improvement: Here, There, and Everywhere. *BioScience* **60**, 337–348 (2010). URL http://bioscience.oxfordjournals.org/content/60/5/337.

94. Peiffer, D. A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* **16**, 1136–1148 (2006). URL http://genome.cshlp.org/content/16/9/1136.

95. Morgan, A. P. *et al.* The Mouse Universal Genotyping Array: From Substrains to Subspecies. *G3* **6**, 263–279 (2016). URL http://www.g3journal.org/content/6/2/263.

96. Morgan, A. P. argyle : An R Package for Analysis of Illumina Genotyping Arrays. *G3* **6**, 281–286 (2016). URL http://www.g3journal.org/content/6/2/281.

97. Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005). URL http://genome.cshlp.org/content/15/11/1496.

98. Didion, J. P. *et al.* Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* **13**, 34 (2012). URL http://dx.doi.org/10.1186/1471-2164-13-34.

99. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333–351 (2016). URL http://www.nature.com/nrg/journal/v17/n6/full/nrg.2016.49.html.

100. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).

101. Little, C. & Tyzzer, E. A Mendelian explanation of rejection and susceptibility. *J. Med. Res.* **33**, 393–425 (1916). Bibtex: little1916mendelian.

102. Snell, G. D. Methods for the study of histocompatibility genes. *J. Genet.* **49**, 87–108 (1948).

103. Castle, W. E. & Allen, G. M. The Heredity of Albinism. *Proceedings of the American Academy of Arts and Sciences* **38**, 603–622 (1903). URL http://www.jstor.org/stable/20021812.

104. Castle, W. E. & Little, C. C. On a Modified Mendelian Ratio Among Yellow Mice. *Science* **32**, 868–870 (1910). URL http://science.sciencemag.org/content/32/833/868.

105. Silvers, W. K. *The Coat Colors of Mice: A Model for Mammalian Gene Action and Interaction* (Springer, New York, 1979).

106. Lyon, M. F. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190**, 372–373 (1961).

107. Gardner, M. B., Rasheed, S., Pal, B. K., Estes, J. D. & O'Brien, S. J. Akvr-1, a dominant murine leukemia virus restriction gene, is polymorphic in leukemia-prone wild mice. *PNAS* **77**, 531–535 (1980). URL http://www.pnas.org/content/77/1/531.

108. Ingalls, A. M., Dickie, M. M. & Snell, G. D. Obese, a new mutation in the house mouse. *J. Hered.* **41**, 317–318 (1950).

109. Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–432 (1994).

110. Russell, W. L., Russell, L. B. & Kelly, E. M. Radiation dose rate and mutation frequency. *Science* **128**, 1546–1550 (1958).

111. Paigen, K. One Hundred Years of Mouse Genetics: An Intellectual History. I. The Classical Period (1902-1980). *Genetics* **163**, 1–7 (2003). URL http://www.genetics.org/content/163/1/1.

112. Henig, R. M. *The Monk in the Garden: The Lost and Found Genius of Gregor Mendel, the Father of Genetics* (Mariner Books, Boston, 2001).

113. Bono, H., Kasukawa, T., Furuno, M., Hayashizaki, Y. & Okazaki, Y. FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones. *Nucleic Acids Res* **30**, 116–118 (2002). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC99137/.

114. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014). URL http://www.nature.com/nature/journal/v515/n7527/full/nature13992.html.

115. Yates, A. *et al.* Ensembl 2016. *Nucl. Acids Res.* **44**, D710–D716 (2016). URL http://nar.oxfordjournals.org/content/44/D1/D710.

116. Berry, R. J. & Scriven, P. N. The house mouse: a model and motor for evolutionary understanding. *Biological Journal of the Linnean Society* **84**, 335–347 (2005). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8312.2005.00438.x/abstract.

117. Miller, R. A. *et al.* Exotic mice as models for aging research: polemic and prospectus. *Neurobiol. Aging* **20**, 217–231 (1999).

118. Didion, J. P. & de Villena, F. P.-M. Deconstructing Mus gemischus: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. *Mamm Genome* **24**, 1–20 (2013). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4034049/.

119. Phifer-Rixey, M. & Nachman, M. W. Insights into mammalian biology from the wild house mouse Mus musculus. *eLife* **4**, e05959 (2015). URL https://elifesciences.org/content/4/e05959v1.

120. Hardouin, E. A. *et al.* House mouse colonization patterns on the sub-Antarctic Kerguelen Archipelago suggest singular primary invasions and resilience against re-invasion. *BMC Evolutionary Biology* **10**, 325 (2010). URL http://dx.doi.org/10.1186/1471-2148-10-325.

121. Gray, M. M. *et al.* Demographic history of a recent invasion of house mice on the isolated Island of Gough. *Mol Ecol* **23**, 1923–1939 (2014). URL http://onlinelibrary.wiley.com/doi/10.1111/mec.12715/abstract.

122. Babiker, H. & Tautz, D. Molecular and phenotypic distinction of the very recently evolved insular subspecies Mus musculus helgolandicus ZIMMERMANN, 1953. *BMC Evolutionary Biology* **15**, 160 (2015). URL http://dx.doi.org/10.1186/s12862-015-0439-5.

123. Boursot, P., Auffray, J.-C., Britton-Davidian, J. & Bonhomme, F. The Evolution of House Mice. *Annual Review of Ecology and Systematics* **24**, 119–152 (1993). URL http://www.jstor.org/stable/2097175.

124. Wilson, D. E. & Reeder, D. M. (eds.). *Mammal Species of the World : A Taxonomic and Geographic Reference* (Johns Hopkins University Press, Baltimore, 2005), 3rd edition edn.

125. Suzuki, H., Shimada, T., Terashima, M., Tsuchiya, K. & Aplin, K. Temporal, spatial, and ecological modes of evolution of Eurasian Mus based on mitochondrial and nuclear gene sequences. *Mol. Phylogenet. Evol.* **33**, 626–646 (2004).

126. Bonhomme, F. *et al.* Biochemical diversity and evolution in the genus Mus. *Biochem. Genet.* **22**, 275–303 (1984).

127. Prager, E. M., Orrego, C. & Sage, R. D. Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. *Genetics* **150**, 835–861 (1998). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1460354/.

128. Cucchi, T., Vigne, J.-D. & Auffray, J.-C. First occurrence of the house mouse (Mus musculus domesticus Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. *Biological Journal of the Linnean Society* **84**, 429–445 (2005). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8312.2005.00445.x/abstract.

129. Rajabi-Maham, H. *et al.* The south-eastern house mouse Mus musculus castaneus (Rodentia: Muridae) is a polytypic subspecies. *Biol J Linn Soc Lond* **107**, 295–306 (2012). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8312.2012.01957.x/abstract.

130. Geraldes, A. *et al.* Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol. Ecol.* **17**, 5349–5363 (2008).

131. Yonekawa, H. *et al.* Hybrid origin of Japanese mice "Mus musculus molossinus": evidence from restriction analysis of mitochondrial DNA. *Mol Biol Evol* **5**, 63–78 (1988). URL http://mbe.oxfordjournals.org/content/5/1/63.

132. Macholán, M., Baird, S. J. E., Munclinger, P. & Piálek, J. (eds.). *Evolution of the House Mouse* (Cambridge University Press, New York, 2012), 1 edition edn.

133. Gabriel, S. I., Jóhannesdóttir, F., Jones, E. P. & Searle, J. B. Colonization, mouse-style. *BMC Biology* **8**, 131 (2010). URL http://dx.doi.org/10.1186/1741-7007-8-131.

134. Sage, R. D., Atchley, W. R. & Capanna, E. House Mice as Models in Systematic Biology. *Syst Biol* **42**, 523–561 (1993). URL http://sysbio.oxfordjournals.org/content/42/4/523.

135. Din, W. *et al.* Origin and radiation of the house mouse: clues from nuclear genes. *Journal of Evolutionary Biology* **9**, 519–539 (1996). URL http://onlinelibrary.wiley.com/doi/10.1046/j.1420-9101.1996.9050519.x/abstract.

136. Bonhomme, F. *et al.* Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among house mouse subspecies. *Genome Biology* **8**, R80 (2007). URL http://dx.doi.org/10.1186/gb-2007-8-5-r80.

137. Bishop, C. E., Boursot, P., Baron, B., Bonhomme, F. & Hatat, D. Most classical Mus musculus domesticus laboratory mouse strains carry a Mus musculus musculus Y chromosome. *Nature* **315**, 70–72 (1985).

138. Teeter, K. C. *et al.* Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res* **18**, 67–76 (2008). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2134771/.

139. Liu, K. J. *et al.* Interspecific introgressive origin of genomic diversity in the house mouse. *PNAS* **112**, 196–201 (2015). URL http://www.pnas.org/content/112/1/196.

140. Britton-Davidian, J., Fel-Clair, F., Lopez, J., Alibert, P. & Boursot, P. Postzygotic isolation between the two European subspecies of the house mouse: estimates from fertility patterns in wild and laboratory-bred hybrids. *Biological Journal of the Linnean Society* **84**, 379–393 (2005). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8312.2005.00441.x/abstract.

141. Forejt, J. & Iványi, P. Genetic studies on male sterility of hybrids between laboratory and wild mice (Mus musculus L.). *Genet. Res.* **24**, 189–206 (1974).

142. Forejt, J. Hybrid sterility in the mouse. *Trends in Genetics* **12**, 412–417 (1996). URL http://www.sciencedirect.com/science/article/pii/0168952596100408.

143. Good, J. M., Handel, M. A. & Nachman, M. W. Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution* **62**, 50–65 (2008).

144. Bhattacharyya, T. *et al.* Mechanistic basis of infertility of mouse intersubspecific hybrids. *PNAS* **110**, E468–E477 (2013). URL http://www.pnas.org/content/110/6/E468.

145. Forejt, J., Gregorová, S. & Goetz, P. XY pair associates with the synaptonemal complex

of autosomal male-sterile translocations in pachytene spermatocytes of the mouse (Mus musculus). *Chromosoma* **82**, 41–53 (1981). URL http://link.springer.com/article/10.1007/BF00285748.

146. Campbell, P., Good, J. M. & Nachman, M. W. Meiotic sex chromosome inactivation is disrupted in sterile hybrid male house mice. *Genetics* **193**, 819–828 (2013).

147. Forejt, J., Vincek, V., Klein, J., Lehrach, H. & Loudová-Micková, M. Genetic mapping of thet-complex region on mouse chromosome 17 including the Hybrid sterility-1 gene. *Mammalian Genome* **1**, 84 (1991). URL http://link.springer.com/article/10.1007/BF02443783.

148. Good, J. M., Dean, M. D. & Nachman, M. W. A Complex Genetic Basis to X-Linked Hybrid Male Sterility Between Two Species of House Mice. *Genetics* **179**, 2213–2228 (2008). URL http://www.genetics.org/content/179/4/2213.

149. Turner, L. M., White, M. A., Tautz, D. & Payseur, B. A. Genomic Networks of Hybrid Sterility. *PLOS Genet* **10**, e1004162 (2014). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004162.

150. Albrechtová, J. *et al.* Sperm-related phenotypes implicated in both maintenance and breakdown of a natural species barrier in the house mouse. *Proc. Biol. Sci.* **279**, 4803–4810 (2012).

151. Turner, L. M. & Harr, B. Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. *eLife* **3**, e02504 (2014). URL https://elifesciences.org/content/3/e02504v1.

152. Storchová, R. *et al.* Genetic analysis of X-linked hybrid sterility in the house mouse. *Mamm Genome* **15**, 515–524 (2004). URL http://link.springer.com/article/10.1007/s00335-004-2386-0.

153. Parvanov, E. D., Petkov, P. M. & Paigen, K. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science* **327**, 835–835 (2010). URL http://science.sciencemag.org/content/327/5967/835.

154. Festing, M. Inbred strains of mice: a vital resource for biomedical research. *Mouse Genome* (1997). Bibtex: festing1997inbred.

155. Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nat Genet* **24**, 23–25 (2000). URL http://www.nature.com/ng/journal/v24/n1/full/ng0100_23.html.

156. Yalcin, B. *et al.* Commercially Available Outbred Mice for Genome-Wide Association Studies. *PLOS Genet* **6**, e1001085 (2010). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001085.

157. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* **43**, 648–655 (2011). URL http://www.nature.com.libproxy.lib.unc.edu/ng/journal/v43/n7/full/ng.847.html.

158. Bonhomme, F., Guenet, J.-L., Dod, B., Moriwaki, K. & Bulfield, G. The polyphyletic origin of laboratory inbred mice and their rate of evolution. *Biological Journal of the Linnean Society* **30**,

51–58 (1987). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8312.1987.tb00288.x/abstract.

159. Wade, C. M. *et al.* The mosaic structure of variation in the laboratory mouse genome. *Nature* **420**, 574–578 (2002). URL http://www.nature.com/nature/journal/v420/n6915/abs/nature01252.html.

160. Wiltshire, T. *et al.* Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3380–3385 (2003).

161. Frazer, K. A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053 (2007). URL http://www.nature.com/nature/journal/v448/n7157/abs/nature06067.html.

162. Yang, H., Bell, T. A., Churchill, G. A. & Pardo-Manuel de Villena, F. On the subspecific origin of the laboratory mouse. *Nat Genet* **39**, 1100–1107 (2007). URL http://www.nature.com/ng/journal/v39/n9/abs/ng2087.html.

163. Guénet, J.-L. & Bonhomme, F. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends in Genetics* **19**, 24–31 (2003). URL http://www.sciencedirect.com/science/article/pii/S0168952502000070.

164. Ideraabdullah, F. Y. *et al.* Genetic and Haplotype Diversity Among Wild-Derived Mouse Inbred Strains. *Genome Res.* **14**, 1880–1887 (2004). URL http://genome.cshlp.org/content/14/10a/1880.

165. Harr, B. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**, 730–737 (2006).

166. Payseur, B. A. & Place, M. Searching the Genomes of Inbred Mouse Strains for Incompatibilities That Reproductively Isolate Their Wild Relatives. *J Hered* **98**, 115–122 (2007). URL http://jhered.oxfordjournals.org/content/98/2/115.

167. Mirzaghaderi, G. & Hörandl, E. The evolution of meiotic sex and its alternatives. *Proceedings of the Royal Society B: Biological Sciences* **283**, 20161221 (2016). URL http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2016.1221.

168. Otto, S. P. & Lenormand, T. Resolving the paradox of sex and recombination. *Nat Rev Genet* **3**, 252–261 (2002). URL http://www.nature.com.libproxy.lib.unc.edu/nrg/journal/v3/n4/full/nrg761.html.

169. Morgan, T. H. Random Segregation Versus Coupling in Mendelian Inheritance. *Science* **34**, 384–384 (1911). URL http://science.sciencemag.org/content/34/873/384.

170. Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J. Exp. Zool.* **14**, 43–59 (1913). URL http://onlinelibrary.wiley.com/doi/10.1002/jez.1400140104/abstract.

171. Fledel-Alon, A. *et al.* Broad-Scale Recombination Patterns Underlying Proper Disjunction in Humans. *PLOS Genet* **5**, e1000658 (2009). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000658.

172. Bernstein, H., Bernstein, C. & E., R. Meiosis as an Evolutionary Adaptation for DNA Repair. In *DNA Repair* (ed. Kruman, I.) (InTech, 2011). URL http://www.intechopen.com/books/dna-repair/meiosis-as-an-evolutionary-adaptation-for-dna-repair.

173. Baudat, F., Imai, Y. & de Massy, B. Meiotic recombination in mammals: localization and regulation. *Nat Rev Genet* **14**, 794–806 (2013). URL http://www.nature.com/nrg/journal/v14/n11/full/nrg3573.html.

174. Dumont, B. L., Broman, K. W. & Payseur, B. A. Variation in Genomic Recombination Rates Among Heterogeneous Stock Mice. *Genetics* **182**, 1345–1349 (2009). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2728871/.

175. Dumont, B. L. & Payseur, B. A. Genetic Analysis of Genome-Scale Recombination Rate Evolution in House Mice. *PLOS Genet* **7**, e1002116 (2011). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002116.

176. Dumont, B. L., White, M. A., Steffy, B., Wiltshire, T. & Payseur, B. A. Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Res.* **21**, 114–125 (2011). URL http://genome.cshlp.org/content/21/1/114.

177. Sun, H., Treco, D., Schultes, N. P. & Szostak, J. W. Double-strand breaks at an initiation site for meiotic gene conversion. *Nature* **338**, 87–90 (1989). URL http://www.nature.com/nature/journal/v338/n6210/abs/338087a0.html.

178. Mahadevaiah, S. K. *et al.* Recombinational DNA double-strand breaks in mice precede synapsis. *Nat Genet* **27**, 271–276 (2001). URL http://www.nature.com.libproxy.lib.unc.edu/ng/journal/v27/n3/full/ng0301_271.html.

179. Scherthan, H. *et al.* Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing. *The Journal of Cell Biology* **134**, 1109–1125 (1996). URL http://jcb.rupress.org/content/134/5/1109.

180. Ding, X. *et al.* SUN1 Is Required for Telomere Attachment to Nuclear Envelope and Gametogenesis in Mice. *Developmental Cell* **12**, 863–872 (2007). URL http://www.sciencedirect.com/science/article/pii/S1534580707001189.

181. Wettstein, D., Rasmussen, S. W. & Holm, a. P. B. The Synaptonemal Complex in Genetic Segregation. *Annual Review of Genetics* **18**, 331–411 (1984). URL http://dx.doi.org/10.1146/annurev.ge.18.120184.001555.

182. Martin, R. H., Ko, E. & Rademaker, A. Distribution of aneuploidy in human gametes: Comparison between human sperm and oocytes. *Am. J. Med. Genet.* **39**, 321–331 (1991). URL http://onlinelibrary.wiley.com/doi/10.1002/ajmg.1320390315/abstract.

183. Gray, S. & Cohen, P. E. Control of Meiotic Crossovers: From Double-Strand Break Formation to Designation. *Annual Review of Genetics* **50**, null (2016). URL http://dx.doi.org/10.1146/annurev-genet-120215-035111.

184. Mather, K. The Determination of Position in Crossing-over. II: The chromosome length-chiasma frequency relation. *Cytologia* 514–526 (1937). Bibtex: mather1937determination.

185. Muller, H. J. The Mechanism of Crossing-Over. *The American Naturalist* **50**, 193–221 (1916). URL http://www.journals.uchicago.edu/doi/10.1086/279534.

186. Santos, T. d. l. *et al.* The Mus81-Mms4 Endonuclease Acts Independently of Double-Holliday Junction Resolution to Promote a Distinct Subset of Crossovers During Meiosis in Budding Yeast. *Genetics* **164**, 81–94 (2003). URL http://www.genetics.org/content/164/1/81.

187. Petkov, P. M., Broman, K. W., Szatkiewicz, J. P. & Paigen, K. Crossover interference underlies sex differences in recombination rates. *Trends in Genetics* **23**, 539–542 (2007). URL http://www.sciencedirect.com/science/article/pii/S0168952507003009.

188. Dunn, L. C. & Bennett, D. Sex differences in recombination of linked genes in animals. *Genetics Research* **9**, 211–220 (1967). URL https://www.cambridge.org/core/journals/genetics-research/article/sex-differences-in-recombination-of-linked-genes-in-animals/D07111CA9F7FE8A3D147BCD1657DD771.

189. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination. *The American Journal of Human Genetics* **63**, 861–869 (1998). URL http://www.sciencedirect.com/science/article/pii/S0002929707613895.

190. Cox, A. *et al.* A New Standard Genetic Map for the Laboratory Mouse. *Genetics* **182**, 1335–1344 (2009). URL http://www.genetics.org.libproxy.lib.unc.edu/content/182/4/1335.

191. Johnston, S. E., Bérénos, C., Slate, J. & Pemberton, J. M. Conserved Genetic Architecture Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep (Ovis aries). *Genetics* **203**, 583–598 (2016). URL http://www.genetics.org/content/203/1/583.

192. Samollow, P. B. *et al.* First-Generation Linkage Map of the Gray, Short-Tailed Opossum, Monodelphis domestica, Reveals Genome-Wide Reduction in Female Recombination Rates. *Genetics* **166**, 307–329 (2004). URL http://www.genetics.org/content/166/1/307.

193. Wong, A. K. *et al.* A Comprehensive Linkage Map of the Dog Genome. *Genetics* **184**, 595–605 (2010). URL http://www.genetics.org/content/184/2/595.

194. Ma, L. *et al.* Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLOS Genet* **11**, e1005387 (2015). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005387.

195. Burgoyne, P. S. Genetic homology and crossing over in the X and Y chromosomes of mammals. *Hum Genet* **61**, 85–90 (1982). URL http://link.springer.com/article/10.1007/BF00274192.

196. Kauppi, L. *et al.* Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science* **331**, 916–920 (2011).

197. Perry, J., Palmer, S., Gabriel, A. & Ashworth, A. A Short Pseudoautosomal Region in Laboratory Mice. *Genome Res* **11**, 1826–1832 (2001). URL http://www.ncbi.nlm.nih.gov/pmc/

articles/PMC311143/.

198. Balcova, M. *et al.* Hybrid Sterility Locus on Chromosome X Controls Meiotic Recombination Rate in Mouse. *PLOS Genet* **12**, e1005906 (2016). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005906.

199. Jeffreys, A. J., Murray, J. & Neumann, R. High-Resolution Mapping of Crossovers in Human Sperm Defines a Minisatellite-Associated Recombination Hotspot. *Molecular Cell* **2**, 267–273 (1998). URL http://www.sciencedirect.com/science/article/pii/S1097276500801380.

200. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science* **310**, 321–324 (2005). URL http://science.sciencemag.org/content/310/5746/321.

201. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* **40**, 1124–1129 (2008). URL http://www.nature.com.libproxy.lib.unc.edu/ng/journal/v40/n9/full/ng.213.html.

202. Hayashi, K., Yoshida, K. & Matsui, Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature* **438**, 374–378 (2005). URL http://www.nature.com.libproxy.lib.unc.edu/nature/journal/v438/n7066/full/nature04112.html.

203. Baudat, F. *et al.* PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327**, 836–840 (2010). URL http://science.sciencemag.org/content/327/5967/836.

204. Borde, V. *et al.* Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO J* **28**, 99–111 (2009). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2634730/.

205. Oliver, P. L. *et al.* Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. *PLOS Genet* **5**, e1000753 (2009). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000753.

206. Grey, C. *et al.* Mouse PRDM9 DNA-Binding Specificity Determines Sites of Histone H3 Lysine 4 Trimethylation for Initiation of Meiotic Recombination. *PLOS Biol* **9**, e1001176 (2011). URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001176.

207. Walker, M. *et al.* Affinity-seq detects genome-wide PRDM9 binding sites and reveals the impact of prior chromatin modifications on mammalian recombination hotspot usage. *Epigenetics & Chromatin* **8**, 31 (2015). URL http://dx.doi.org/10.1186/s13072-015-0024-6.

208. Myers, S. *et al.* Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science* **327**, 876–879 (2010). URL http://science.sciencemag.org.libproxy.lib.unc.edu/content/327/5967/876.

209. Baker, C. L. *et al.* PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus

through Haplotype-Specific Initiation of Meiotic Recombination. *PLOS Genet* **11**, e1004916 (2015). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004916.

210. Cole, F., Keeney, S. & Jasin, M. Comprehensive, Fine-Scale Dissection of Homologous Recombination Outcomes at a Hot Spot in Mouse Meiosis. *Molecular Cell* **39**, 700–710 (2010). URL http://www.sciencedirect.com/science/article/pii/S1097276510006271.

211. de Boer, E., Jasin, M. & Keeney, S. Local and sex-specific biases in crossover vs. noncrossover outcomes at meiotic recombination hot spots in mice. *Genes Dev.* **29**, 1721–1733 (2015).

212. Singhal, S. *et al.* Stable recombination hotspots in birds. *Science* **350**, 928–932 (2015). URL http://science.sciencemag.org/content/350/6263/928.

213. Axelsson, E. *et al.* Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Res.* **22**, 51–63 (2012). URL http://genome.cshlp.org/content/22/1/51.

214. Lam, I. & Keeney, S. Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* **350**, 932–937 (2015). URL http://science.sciencemag.org/content/350/6263/932.

215. Auton, A. *et al.* Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLOS Genet* **9**, e1003984 (2013). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003984.

216. Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C. & Forejt, J. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. *Science* **323**, 373–375 (2009). URL http://science.sciencemag.org.libproxy.lib.unc.edu/content/323/5912/373.

217. Davies, B. *et al.* Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* **530**, 171–176 (2016). URL http://www.nature.com.libproxy.lib.unc.edu/nature/journal/v530/n7589/abs/nature16931.html.

218. Smagulova, F. *et al.* Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* **472**, 375–378 (2011). URL http://www.nature.com/nature/journal/v472/n7343/full/nature09869.html.

219. Coop, G. & Przeworski, M. An evolutionary view of human recombination. *Nat Rev Genet* **8**, 23–34 (2007). URL http://www.nature.com/nrg/journal/v8/n1/full/nrg1947.html.

220. Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet* **36**, 1133–1137 (2004). URL http://www.nature.com/ng/journal/v36/n11/full/ng1104-1133.html.

221. Consortium, C. C. The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population. *Genetics* **190**, 389–401 (2012). URL http://www.genetics.org.libproxy.lib.unc.edu/content/190/2/389.

222. Svenson, K. L. *et al.* High-Resolution Genetic Mapping Using the Mouse Diversity Outbred

Population. *Genetics* **190**, 437–447 (2012). URL http://www.genetics.org.libproxy.lib.unc.edu/content/190/2/437.

223. Flint, J. & Eskin, E. Genome-wide association studies in mice. *Nat Rev Genet* **13**, 807–817 (2012). URL http://www.nature.com.libproxy.lib.unc.edu/nrg/journal/v13/n11/full/nrg3335.html.

224. Broman, K. W. & Weber, J. L. Characterization of Human Crossover Interference. *The American Journal of Human Genetics* **66**, 1911–1926 (2000). URL http://www.sciencedirect.com/science/article/pii/S0002929707635435.

225. Broman, K. W., Rowe, L. B., Churchill, G. A. & Paigen, K. Crossover Interference in the Mouse. *Genetics* **160**, 1123–1131 (2002). URL http://www.genetics.org/content/160/3/1123.

226. Dumont, B. L. & Payseur, B. A. Evolution of the Genomic Recombination Rate in Murid Rodents. *Genetics* **187**, 643–657 (2011). URL http://www.genetics.org/content/187/3/643.

227. Henderson, S. A. & Edwards, R. G. Chiasma Frequency and Maternal Age in Mammals. *Nature* **218**, 22–28 (1968). URL http://www.nature.com/nature/journal/v218/n5136/abs/218022a0.html.

228. Hassold, T., Merrill, M., Adkins, K., Freeman, S. & Sherman, S. Recombination and maternal age-dependent nondisjunction: molecular studies of trisomy 16. *Am J Hum Genet* **57**, 867–874 (1995). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1801507/.

229. Lamb, N. E. *et al.* Susceptible chiasmate configurations of chromosome 21 predispose to non–disjunction in both maternal meiosis I and meiosis II. *Nat Genet* **14**, 400–405 (1996). URL http://www.nature.com/ng/journal/v14/n4/abs/ng1296-400.html.

230. Robinson, W. P. *et al.* Maternal Meiosis I Non-Disjunction of Chromosome 15: Dependence of the Maternal Age Effect on Level of Recombination. *Hum. Mol. Genet.* **7**, 1011–1019 (1998). URL http://hmg.oxfordjournals.org/content/7/6/1011.

231. Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nat Genet* **36**, 1203–1206 (2004). URL http://www.nature.com/ng/journal/v36/n11/full/ng1445.html.

232. Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D. & Auton, A. Escape from crossover interference increases with maternal age. *Nature Communications* **6**, 6260 (2015). URL http://www.nature.com/doifinder/10.1038/ncomms7260.

233. Hussin, J., Roy-Gagnon, M.-H., Gendron, R., Andelfinger, G. & Awadalla, P. Age-Dependent Recombination Rates in Human Pedigrees. *PLOS Genet* **7**, e1002251 (2011). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002251.

234. Martin, H. C. *et al.* Multicohort analysis of the maternal age effect on recombination. *Nature Communications* **6**, 7846 (2015). URL http://www.nature.com/doifinder/10.1038/ncomms8846.

235. Vrooman, L. A., Nagaoka, S. I., Hassold, T. J. & Hunt, P. A. Evidence for Paternal Age-Related Alterations in Meiotic Chromosome Dynamics in the Mouse. *Genetics* **196**, 385–396 (2014). URL http://www.genetics.org/content/196/2/385.

236. Shi, Q. *et al.* Absence of Age Effect on Meiotic Recombination between Human X and Y Chromosomes. *The American Journal of Human Genetics* **71**, 254–261 (2002). URL /ajhg/abstract/S0002-9297(07)60471-6.

237. Pardo-Manuel de Villena, F. & Sapienza, C. Genetic mapping of DXYMov15-associated sequences in the pseudoautosomal region of the C57BL/6J strain. *Mamm. Genome* **7**, 237–239 (1996).

238. White, M. A., Ikeda, A. & Payseur, B. A. A pronounced evolutionary shift of the pseudoautosomal region boundary in house mice. *Mamm Genome* **23**, 454–466 (2012). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3519421/.

239. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**, 552–564 (2006). URL http://www.nature.com.libproxy.lib.unc.edu/nrg/journal/v7/n7/full/nrg1895.html.

240. Larson, E. L. *et al.* Contrasting Levels of Molecular Evolution on the Mouse X Chromosome. *Genetics* **203**, 1841–1857 (2016). URL http://www.genetics.org/content/203/4/1841.

241. Campbell, C. L., Bhérer, C., Morrow, B. E., Boyko, A. R. & Auton, A. A Pedigree-Based Map of Recombination in the Domestic Dog Genome. *G3* (2016). URL http://www.g3journal.org/content/early/2016/09/01/g3.116.034678.

242. Beadle, G. W. A Possible Influence of the Spindle Fibre on Crossing-Over in Drosophila. *Proc Natl Acad Sci U S A* **18**, 160–165 (1932). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1076180/.

243. Jensen-Seaman, M. I. *et al.* Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Res.* **14**, 528–538 (2004). URL http://genome.cshlp.org/content/14/4/528.

244. Meredith, R. W. *et al.* Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science* **334**, 521–524 (2011). URL http://science.sciencemag.org/content/334/6055/521.

245. Huxley, J. S. Sexual difference of linkage in Gammarus chevreuxi. *Journ. of Gen.* **20**, 145–156 (1922). URL http://link.springer.com.libproxy.lib.unc.edu/article/10.1007/BF02983136.

246. Brandvain, Y. & Coop, G. Scrambling eggs: meiotic drive and the evolution of female recombination rates. *Genetics* **190**, 709–723 (2012).

247. Liu, E. Y. *et al.* High-Resolution Sex-Specific Linkage Maps of the Mouse Reveal Polarized Distribution of Crossovers in Male Germline. *Genetics* **197**, 91–106 (2014). URL http://www.genetics.org/content/197/1/91.

248. Sakamoto, T. *et al.* A Microsatellite Linkage Map of Rainbow Trout (Oncorhynchus mykiss)

Characterized by Large Sex-Specific Differences in Recombination Rates. *Genetics* **155**, 1331–1345 (2000). URL http://www.genetics.org/content/155/3/1331.

249. Reid, D. P. *et al.* A Genetic Linkage Map of Atlantic Halibut (Hippoglossus hippoglossus L.). *Genetics* **177**, 1193–1205 (2007). URL http://www.genetics.org/content/177/2/1193.

250. Brelsford, A., Dufresnes, C. & Perrin, N. High-density sex-specific linkage maps of a European tree frog (Hyla arborea) identify the sex chromosome without information on offspring sex. *Heredity* **116**, 177–181 (2016). URL http://www.nature.com.libproxy.lib.unc.edu/hdy/journal/v116/n2/full/hdy201583a.html.

251. Thomson, G. J. & Feldman, M. W. Population genetics of modifiers of meiotic drive. II linkage modification in the segregation distortion system. *Theoretical Population Biology* **5**, 155–162 (1974). URL http://www.sciencedirect.com/science/article/pii/0040580974900380.

252. Haig, D. & Grafen, A. Genetic scrambling as a defence against meiotic drive. *Journal of Theoretical Biology* **153**, 531–558 (1991). URL http://www.sciencedirect.com/science/article/pii/S0022519305801559.

253. Jenkins, T. G., Aston, K. I., Pflueger, C., Cairns, B. R. & Carrell, D. T. Age-Associated Sperm DNA Methylation Alterations: Possible Implications in Offspring Disease Susceptibility. *PLOS Genet* **10**, e1004458 (2014). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004458.

254. Rowsey, R., Gruhn, J., Broman, K. W., Hunt, P. A. & Hassold, T. Examining Variation in Recombination Levels in the Human Female: A Test of the Production-Line Hypothesis. *The American Journal of Human Genetics* **95**, 108–112 (2014). URL http://www.cell.com/ajhg/abstract/S0002-9297(14)00270-5.

255. Haldane, J. B. S. Sex ratio and unisexual sterility in hybrid animals. *Journ. of Gen.* **12**, 101–109 (1922). URL http://link.springer.com/article/10.1007/BF02983075.

256. Charlesworth, B., Coyne, J. A. & Barton, N. H. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *The American Naturalist* **130**, 113–146 (1987). URL http://www.journals.uchicago.edu/doi/10.1086/284701.

257. Payseur, B. A., Krenz, J. G. & Nachman, M. W. Differential Patterns of Introgression across the X Chromosome in a Hybrid Zone between Two Species of House Mice. *Evolution* **58**, 2064–2078 (2004). URL http://www.jstor.org/stable/3449455.

258. Qvarnström, A. & Bailey, R. I. Speciation through evolution of sex-linked genes. *Heredity* **102**, 4–15 (2008). URL http://www.nature.com/hdy/journal/v102/n1/full/hdy200893a.html.

259. Kousathanas, A., Halligan, D. L. & Keightley, P. D. Faster-X Adaptive Protein Evolution in House Mice. *Genetics* **196**, 1131–1143 (2014). URL http://www.genetics.org/content/196/4/1131.

260. Campbell, P., Good, J. M., Dean, M. D., Tucker, P. K. & Nachman, M. W. The contribution of

the Y chromosome to hybrid male sterility in house mice. *Genetics* **191**, 1271–1281 (2012).

261. Campbell, P. & Nachman, M. W. X-y interactions underlie sperm head abnormality in hybrid male house mice. *Genetics* **196**, 1231–1240 (2014).

262. Baker, C. L. *et al.* Multimer Formation Explains Allelic Suppression of PRDM9 Recombination Hotspots. *PLoS Genet.* **11**, e1005512 (2015).

263. Gatti, D. M. *et al.* Quantitative Trait Locus Mapping Methods for Diversity Outbred Mice. *G3* **4**, 1623–1633 (2014). URL http://www.g3journal.org/content/4/9/1623.

264. Kirkpatrick, M. How and Why Chromosome Inversions Evolve. *PLOS Biol* **8**, e1000501 (2010). URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000501.

265. Shin, H. S., Flaherty, L., Artzt, K., Bennett, D. & Ravetch, J. Inversion in the H-2 complex of t-haplotypes in mice. *Nature* **306**, 380–383 (1983).

266. Hallast, P., Balaresque, P., Bowden, G. R., Ballereau, S. & Jobling, M. A. Recombination Dynamics of a Human Y-Chromosomal Palindrome: Rapid GC-Biased Gene Conversion, Multi-kilobase Conversion Tracts, and Rare Inversions. *PLOS Genet* **9**, e1003666 (2013). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003666.

267. Yalcin, B. *et al.* Sequence-based characterization of structural variation in the mouse genome. *Nature* **477**, 326–329 (2011). URL http://www.nature.com/nature/journal/v477/n7364/full/nature10432.html.

268. Pezer, Ž., Harr, B., Teschke, M., Babiker, H. & Tautz, D. Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. *Genome Res.* **25**, 1114–1124 (2015). URL http://genome.cshlp.org.libproxy.lib.unc.edu/content/25/8/1114.

269. Chesler, E. J. *et al.* The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm Genome* **19**, 382–389 (2008). URL http://link.springer.com/article/10.1007/s00335-008-9135-8.

270. Chesler, E. J. *et al.* Diversity Outbred Mice at 21: Maintaining Allelic Variation in the Face of Selection. *G3* **6**, 3893–3902 (2016). URL http://www.g3journal.org/content/early/2016/09/29/g3.116.035527.

271. Liu, E. Y., Zhang, Q., McMillan, L., Villena, F. P.-M. d. & Wang, W. Efficient genome ancestry inference in complex pedigrees with inbreeding. *Bioinformatics* **26**, i199–i207 (2010). URL http://bioinformatics.oxfordjournals.org/content/26/12/i199.

272. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010). URL http://bioinformatics.oxfordjournals.org/content/26/22/2867.

273. Karlin, S. & Altschul, S. F. Methods for assessing the statistical significance of molecular

sequence features by using general scoring schemes. *PNAS* **87**, 2264–2268 (1990). URL http://www.pnas.org/content/87/6/2264.

274. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013). URL http://arxiv.org/abs/1303.3997, arXiv: 1303.3997.

275. Faust, G. G. & Hall, I. M. SAMBLASTER fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014). URL http://bioinformatics.oxfordjournals.org.libproxy.lib.unc.edu/content/30/17/2503.

276. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269–276 (2011). URL http://www.nature.com/ng/journal/v43/n3/full/ng.768.html.

277. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17**, 556–565 (2007). URL http://genome.cshlp.org/content/17/5/556.

278. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).

279. Dopman, E. B. & Hartl, D. L. A portrait of copy-number polymorphism in Drosophila melanogaster. *PNAS* **104**, 19920–19925 (2007). URL http://www.pnas.org.libproxy.lib.unc.edu/content/104/50/19920.

280. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).

281. Armengol, L., Pujana, M. A., Cheung, J., Scherer, S. W. & Estivill, X. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* **12**, 2201–2208 (2003). URL http://hmg.oxfordjournals.org/content/12/17/2201.

282. Hurles, M. Are 100,000 "SNPs" Useless? *Science* **298**, 1509–1509 (2002). URL http://science.sciencemag.org.libproxy.lib.unc.edu/content/298/5598/1509.

283. Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E. & Matsuda, G. Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Syst Biol* **28**, 132–163 (1979). URL http://sysbio.oxfordjournals.org.libproxy.lib.unc.edu/content/28/2/132.

284. Dover, G. Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117 (1982).

285. Nagylaki, T. & Petes, T. D. Intrachromosomal Gene Conversion and the Maintenance of Sequence Homogeneity Among Repeated Genes. *Genetics* **100**, 315–337 (1982). URL http://www.genetics.org.libproxy.lib.unc.edu/content/100/2/315.

286. Didion, J. P. *et al.* A Multi-Megabase Copy Number Gain Causes Maternal Transmission Ratio Distortion on Mouse Chromosome 2. *PLOS Genet* **11**, e1004850 (2015). URL http://journals.plos.org/plosgenetics/article?id=10.1371/

journal.pgen.1004850.

287. Chevret, P., Veyrunes, F. & Britton-Davidian, J. Molecular phylogeny of the genus Mus (Rodentia: Murinae) based on mitochondrial and nuclear data. *Biological Journal of the Linnean Society* **84**, 417–427 (2005). URL http://onlinelibrary.wiley.com.libproxy.lib.unc.edu/doi/10.1111/j.1095-8312.2005.00444.x/abstract.

288. Muffato, M., Louis, A., Poisnel, C.-E. & Crollius, H. R. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**, 1119–1121 (2010). URL http://bioinformatics.oxfordjournals.org.libproxy.lib.unc.edu/content/26/8/1119.

289. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969–1973 (2012). URL http://mbe.oxfordjournals.org.libproxy.lib.unc.edu/content/29/8/1969.

290. Swallow, J. G., Carter, P. A. & Garland, T. Artificial selection for increased wheel-running behavior in house mice. *Behav. Genet.* **28**, 227–237 (1998).

291. Salcedo, T., Geraldes, A. & Nachman, M. W. Nucleotide variation in wild and inbred mice. *Genetics* **177**, 2277–2291 (2007).

292. Halligan, D. L. *et al.* Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. *PLOS Genet* **9**, e1003995 (2013). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003995.

293. Moran, P. A. Wandering distributions and the electrophoretic profile. *Theor Popul Biol* **8**, 318–330 (1975).

294. Egan, C. M., Sridhar, S., Wigler, M. & Hall, I. M. Recurrent DNA copy number variation in the laboratory mouse. *Nat Genet* **39**, 1384–1389 (2007). URL http://www.nature.com.libproxy.lib.unc.edu/ng/journal/v39/n11/full/ng.2007.19.html.

295. Yeh, T.-C. *et al.* Splicing Factor Cwc22 Is Required for the Function of Prp2 and for the Spliceosome To Escape from a Futile Pathway. *Mol. Cell. Biol.* **31**, 43–53 (2011). URL http://mcb.asm.org.libproxy.lib.unc.edu/content/31/1/43.

296. Li, H. *et al.* TreeFam a curated database of phylogenetic trees of animal gene families. *Nucl. Acids Res.* **34**, D572–D580 (2006). URL http://nar.oxfordjournals.org.libproxy.lib.unc.edu/content/34/suppl_1/D572.

297. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13**, R51 (2012).

298. Crowley, J. J. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat Genet* **47**, 353–360 (2015). URL http://www.nature.com.libproxy.lib.unc.edu/ng/journal/v47/n4/full/ng.3222.html.

299. Phifer-Rixey, M., Bomhoff, M. & Nachman, M. W. Genome-Wide Patterns of Differentiation Among House Mouse Subspecies. *Genetics* **198**, 283–297 (2014). URL http://www.genetics.org/content/198/1/283.

300. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* **34**, 525–527 (2016). URL http://www.nature.com/nbt/journal/v34/n5/abs/nbt.3519.html.

301. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* **8**, 762–775 (2007). URL http://www.nature.com.libproxy.lib.unc.edu/nrg/journal/v8/n10/full/nrg2193.html.

302. Cole, F. *et al.* Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat Genet* **46**, 1072–1080 (2014). URL http://www.nature.com.libproxy.lib.unc.edu/ng/journal/v46/n10/full/ng.3068.html.

303. Demuth, J. P., Bie, T. D., Stajich, J. E., Cristianini, N. & Hahn, M. W. The Evolution of Mammalian Gene Families. *PLOS ONE* **1**, e85 (2006). URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000085.

304. Dumont, B. L. & Eichler, E. E. Signals of Historical Interlocus Gene Conversion in Human Segmental Duplications. *PLOS ONE* **8**, e75949 (2013). URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0075949.

305. Melamed, C. & Kupiec, M. Effect of donor copy number on the rate of gene conversion in the yeast Saccharomyces cerevisiae. *Molec. Gen. Genet.* **235**, 97–103 (1992). URL http://link.springer.com.libproxy.lib.unc.edu/article/10.1007/BF00286186.

306. Soh, Y. Q. S. *et al.* Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**, 800–813 (2014).

307. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003). URL http://www.nature.com.libproxy.lib.unc.edu/nature/journal/v423/n6942/full/nature01723.html.

308. Liao, D., Pavelitz, T., Kidd, J. R., Kidd, K. K. & Weiner, A. M. Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. *The EMBO Journal* **16**, 588–598 (1997). URL http://emboj.embopress.org/content/16/3/588.

309. Eickbush, T. H. & Eickbush, D. G. Finely Orchestrated Movements: Evolution of the Ribosomal RNA Genes. *Genetics* **175**, 477–485 (2007). URL http://www.genetics.org.libproxy.lib.unc.edu/content/175/2/477.

310. Schindelhauer, D. & Schwarz, T. Evidence for a Fast, Intrachromosomal Conversion Mechanism From Mapping of Nucleotide Variants Within a Homogeneous α-Satellite DNA Array. *Genome Res.* **12**, 1815–1826 (2002). URL http://genome.cshlp.org.libproxy.lib.unc.edu/content/12/12/1815.

311. Shi, J. *et al.* Widespread Gene Conversion in Centromere Cores. *PLOS Biol* **8**, e1000327 (2010). URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000327.

312. Scavetta, R. J. & Tautz, D. Copy Number Changes of CNV Regions in Intersubspecific

Crosses of the House Mouse. *Mol Biol Evol* **27**, 1845–1856 (2010). URL http://mbe.oxfordjournals.org.libproxy.lib.unc.edu/content/27/8/1845.

313. Laan, R. v. d. *et al.* Ubiquitin ligase Rad18Sc localizes to the XY body and to other chromosomal regions that are unpaired and transcriptionally silenced during male meiotic prophase. *Journal of Cell Science* **117**, 5023–5033 (2004). URL http://jcs.biologists.org.libproxy.lib.unc.edu/content/117/21/5023.

314. Baarends, W. M. *et al.* Silencing of Unpaired Chromatin and Histone H2A Ubiquitination in Mammalian Meiosis. *Mol. Cell. Biol.* **25**, 1041–1053 (2005). URL http://mcb.asm.org.libproxy.lib.unc.edu/content/25/3/1041.

315. Turner, J. M. A. *et al.* BRCA1, Histone H2AX Phosphorylation, and Male Meiotic Sex Chromosome Inactivation. *Current Biology* **14**, 2135–2142 (2004). URL http://www.sciencedirect.com/science/article/pii/S0960982204009194.

316. Lindholm, A. K. *et al.* The Ecology and Evolutionary Dynamics of Meiotic Drive. *Trends in Ecology & Evolution* **31**, 315–326 (2016). URL http://www.sciencedirect.com/science/article/pii/S0169534716000434.

317. Hauffe, H. C. & Searle, J. B. Extreme Karyotypic Variation in a Mus musculus domesticus Hybrid Zone: The Tobacco Mouse Story Revisited. *Evolution* **47**, 1374–1395 (1993). URL http://www.jstor.org/stable/2410154.

318. Nachman, M. W., Boyer, S. N., Searle, J. B. & Aquadro, C. F. Mitochondrial DNA Variation and the Evolution of Robertsonian Chromosomal Races of House Mice, Mus Domesticus. *Genetics* **136**, 1105–1120 (1994). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205867/.

319. Steckelberg, A.-L., Boehm, V., Gromadzka, A. M. & Gehring, N. H. CWC22 Connects Pre-mRNA Splicing and Exon Junction Complex Assembly. *Cell Reports* **2**, 454–461 (2012). URL http://www.sciencedirect.com/science/article/pii/S2211124712002574.

320. Chick, J. M. *et al.* Defining the consequences of genetic variation on a proteome-wide scale. *Nature* **534**, 500–505 (2016). URL http://www.nature.com/nature/journal/v534/n7608/full/nature18270.html.

321. Green, M. R. & Sambrook, J. *Molecular Cloning: A Laboratory Manual (Fourth Edition): Three-volume set* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y, 2012), 4th edition edn.

322. Morgan, A. P. *et al.* Genome Report: Whole Genome Sequence of Two Wild-Derived Mus musculus domesticus Inbred Strains, LEWES/EiJ and ZALENDE/EiJ, with Different Diploid Numbers. *G3* g3.116.034751 (2016). URL http://www.g3journal.org/content/early/2016/10/07/g3.116.034751.

323. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]* (2012). URL http://arxiv.org/abs/1207.3907, arXiv: 1207.3907.

324. Holt, J. & McMillan, L. Merging of multi-string BWTs with applications. *Bioinformatics* **30**,

3524–3531 (2014). URL http://bioinformatics.oxfordjournals.org.libproxy.lib.unc.edu/content/30/24/3524.

325. Dobin, A. *et al.* STAR ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). URL http://bioinformatics.oxfordjournals.org.libproxy.lib.unc.edu/content/29/1/15.

326. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* **29**, 644–652 (2011). URL http://www.nature.com/nbt/journal/v29/n7/abs/nbt.1883.html.

327. Edgar, R. C. MUSCLE multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792–1797 (2004). URL http://nar.oxfordjournals.org.libproxy.lib.unc.edu/content/32/5/1792.

328. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). URL http://bioinformatics.oxfordjournals.org.libproxy.lib.unc.edu/content/30/9/1312.

329. Staubach, F. *et al.* Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus). *PLOS Genet* **8**, e1002891 (2012). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002891.

330. Williamson, S. H. *et al.* Localizing Recent Adaptive Evolution in the Human Genome. *PLOS Genet* **3**, e90 (2007). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030090.

331. Grossman, S. R. *et al.* Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell* **152**, 703–713 (2013). URL http://www.sciencedirect.com/science/article/pii/S0092867413000871.

332. Colonna, V. *et al.* Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology* **15**, R88 (2014). URL http://dx.doi.org/10.1186/gb-2014-15-6-r88.

333. Fu, W. & Akey, J. M. Selection and Adaptation in the Human Genome. *Annu. Rev. Genom. Hum. Genet.* **14**, 467–489 (2013). URL http://www.annualreviews.org/doi/10.1146/annurev-genom-091212-153509.

334. Bryk, J. & Tautz, D. Copy number variants and selective sweeps in natural populations of the house mouse (Mus musculus domesticus). *Mus musculus domesticus* **5**, 153 (2014). URL http://journal.frontiersin.org/article/10.3389/fgene.2014.00153/full.

335. Kaplan, N. L., Hudson, R. R. & Langley, C. H. The "hitchhiking effect" revisited. *Genetics* **123**, 887–899 (1989). URL http://www.genetics.org/content/123/4/887.

336. Pelz, H.-J. *et al.* The Genetic Basis of Resistance to Anticoagulants in Rodents. *Genetics* **170**, 1839–1847 (2005). URL http://www.genetics.org/content/170/4/1839.

337. Bersaglieri, T. *et al.* Genetic Signatures of Strong Recent Positive Selection at the Lactase

Gene. *The American Journal of Human Genetics* **74**, 1111–1120 (2004). URL http://www.sciencedirect.com/science/article/pii/S0002929707628389.

338. Sandler, L. & Novitski, E. Meiotic Drive as an Evolutionary Force. *The American Naturalist* **91**, 105–110 (1957). URL http://www.journals.uchicago.edu/doi/10.1086/281969.

339. White, M. J. D. *Modes of Speciation* (W.H.Freeman & Co Ltd, San Francisco, 1978), 1st edition edition edn.

340. Henikoff, S. & Malik, H. S. Centromeres as selfish drivers. *Nature* **417**, 227–227 (2002). URL http://www.nature.com/nature/journal/v417/n6886/full/417227a.html.

341. Derome, N., Métayer, K., Montchamp-Moreau, C. & Veuille, M. Signature of Selective Sweep Associated With the Evolution of sex-ratio Drive in Drosophila simulans. *Genetics* **166**, 1357–1366 (2004). URL http://www.genetics.org/content/166/3/1357.

342. Villena, F. P.-M. d. & Sapienza, C. Female Meiosis Drives Karyotypic Evolution in Mammals. *Genetics* **159**, 1179–1189 (2001). URL http://www.genetics.org/content/159/3/1179.

343. Presgraves, D. C., Gérard, P. R., Cherukuri, A. & Lyttle, T. W. Large-Scale Selective Sweep among Segregation Distorter Chromosomes in African Populations of Drosophila melanogaster. *PLOS Genet* **5**, e1000463 (2009). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000463.

344. Babcock, C. S. & Anderson, W. W. Molecular evolution of the Sex-Ratio inversion complex in Drosophila pseudoobscura: analysis of the Esterase-5 gene region. *Mol Biol Evol* **13**, 297–308 (1996). URL http://mbe.oxfordjournals.org/content/13/2/297.

345. Dyer, K. A., Charlesworth, B. & Jaenike, J. Chromosome-wide linkage disequilibrium as a consequence of meiotic drive. *PNAS* **104**, 1587–1592 (2007). URL http://www.pnas.org/content/104/5/1587.

346. Derome, N., Baudry, E., Ogereau, D., Veuille, M. & Montchamp-Moreau, C. Selective Sweeps in a 2-Locus Model for Sex-Ratio Meiotic Drive in Drosophila simulans. *Mol Biol Evol* **25**, 409–416 (2008). URL http://mbe.oxfordjournals.org/content/25/2/409.

347. Kingan, S. B., Garrigan, D. & Hartl, D. L. Recurrent Selection on the Winters sex-ratio Genes in Drosophila simulans. *Genetics* **184**, 253–265 (2010). URL http://www.genetics.org/content/184/1/253.

348. Nolte, V., Pandey, R. V., Kofler, R. & Schlötterer, C. Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in Drosophila mauritiana. *Genome Res.* **23**, 99–110 (2013). URL http://genome.cshlp.org/content/23/1/99.

349. Fishman, L. & Saunders, A. Centromere-Associated Female Meiotic Drive Entails Male Fitness Costs in Monkeyflowers. *Science* **322**, 1559–1562 (2008). URL http://science.sciencemag.org/content/322/5907/1559.

350. Siracusa, L. D., Alvord, W. G., Bickmore, W. A., Jenkins, N. A. & Copeland, N. G. Interspecific backcross mice show sex-specific differences in allelic inheritance. *Genetics* **128**, 813–821 (1991).

URL http://www.genetics.org/content/128/4/813.

351. Montagutelli, X., Turner, R. & Nadeau, J. H. Epistatic Control of Non-Mendelian Inheritance in Mouse Interspecific Crosses. *Genetics* **143**, 1739–1752 (1996). URL http://www.genetics.org/content/143/4/1739.

352. Eversley, C. D. *et al.* Genetic mapping and developmental timing of transmission ratio distortion in a mouse interspecific backcross. *BMC Genetics* **11**, 98 (2010). URL http://dx.doi.org/10.1186/1471-2156-11-98.

353. Kelly, S. A. *et al.* Parent-of-origin effects on voluntary exercise levels and body composition in mice. *Physiological Genomics* **40**, 111–120 (2010). URL http://physiolgenomics.physiology.org/content/40/2/111.

354. Kelly, S. A. *et al.* Genetic architecture of voluntary exercise in an advanced intercross line of mice. *Physiological Genomics* **42**, 190–200 (2010). URL http://physiolgenomics.physiology.org/content/42/2/190.

355. Albrechtsen, A., Moltke, I. & Nielsen, R. Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. *Genetics* **186**, 295–308 (2010). URL http://www.genetics.org/content/186/1/295.

356. Song, Y. *et al.* Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. *Current Biology* **21**, 1296–1301 (2011). URL http://www.sciencedirect.com/science/article/pii/S0960982211007160.

357. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002). URL http://www.nature.com/nature/journal/v419/n6909/full/nature01140.html.

358. Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M. & Servin, B. Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. *Genetics* **193**, 929–941 (2013). URL http://www.genetics.org/content/193/3/929.

359. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLOS Biol* **4**, e72 (2006). URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040072.

360. Lyon, M. F. Male sterility of the mouse t-complex is due to homozygosity of the distorter genes. *Cell* **44**, 357–363 (1986).

361. Hartl, D. L. Complementation Analysis of Male Fertility Among the Segregation Distorter Chromosomes of Drosophila Melanogaster. *Genetics* **73**, 613–629 (1973). URL http://www.genetics.org/content/73/4/613.

362. Hedrick, P. W. The Establishment of Chromosomal Variants. *Evolution* **35**, 322–332 (1981). URL http://www.jstor.org/stable/2407841.

363. Hernandez, R. D. *et al.* Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science* **331**, 920–924 (2011). URL http://science.sciencemag.org/content/331/6019/920.

364. Leamy, L. J., Kelly, S. A., Hua, K. & Pomp, D. Exercise and diet affect quantitative trait loci for body weight and composition traits in an advanced intercross population of mice. *Physiological Genomics* **44**, 1141–1153 (2012). URL http://physiolgenomics.physiology.org/content/44/23/1141.

365. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). URL http://bioinformatics.oxfordjournals.org/content/25/16/2078.

366. Stevens, E. L. *et al.* Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. *PLOS Genet* **7**, e1002287 (2011). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002287.

367. Pocock, M. J. O., Hauffe, H. C. & Searle, J. B. Dispersal in house mice. *Biological Journal of the Linnean Society* **84**, 565–583 (2005). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8312.2005.00455.x/abstract.

368. Purcell, S. *et al.* PLINK A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007). URL http://www.sciencedirect.com/science/article/pii/S0002929707613524.

369. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genet* **8**, e1002967 (2012). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002967.

370. Scheet, P. & Stephens, M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics* **78**, 629–644 (2006). URL http://www.sciencedirect.com/science/article/pii/S000292970763701X.

371. Szpiech, Z. A. & Hernandez, R. D. selscan An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Mol Biol Evol* **31**, 2824–2827 (2014). URL http://mbe.oxfordjournals.org/content/31/10/2824.

372. Browning, B. L. & Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **194**, 459–471 (2013). URL http://www.genetics.org/content/194/2/459.

373. Laurie, C. C. *et al.* Linkage Disequilibrium in Wild Mice. *PLOS Genet* **3**, e144 (2007). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030144.

374. Stephens, J. C. *et al.* Dating the Origin of the CCR5-Δ32 AIDS-Resistance Allele by the Coalescence of Haplotypes. *The American Journal of Human Genetics* **62**, 1507–1515 (1998). URL http://www.sciencedirect.com/science/article/pii/S0002929707627943.

375. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).

376. Cortez, D. *et al.* Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493 (2014). URL http://www.nature.com.libproxy.lib.unc.edu/

nature/journal/v508/n7497/full/nature13151.html.

377. Lahn, B. T. & Page, D. C. Functional Coherence of the Human Y Chromosome. *Science* **278**, 675–680 (1997). URL http://science.sciencemag.org.libproxy.lib.unc.edu/content/278/5338/675.

378. Ellis, P. J. I., Bacon, J. & Affara, N. A. Association of Sly with sex-linked gene amplification during mouse evolution: a side effect of genomic conflict in spermatids? *Hum. Mol. Genet.* **20**, 3010–3021 (2011).

379. Cocquet, J. *et al.* A Genetic Basis for a Postmeiotic X Versus Y Chromosome Intragenomic Conflict in the Mouse. *PLOS Genet* **8**, e1002900 (2012). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002900.

380. Hurst, L. D. Heredity - Abstract of article: Embryonic growth and the evolution of the mammalian Y chromosome. I. The Y as an attractor for selfish growth factors. *Heredity* **73**, 223–232 (1994). URL http://www.nature.com.libproxy.lib.unc.edu/hdy/journal/v73/n3/abs/hdy1994127a.html.

381. Charlesworth, B. The evolution of chromosomal sex determination and dosage compensation. *Current Biology* **6**, 149–162 (1996). URL http://www.sciencedirect.com/science/article/pii/S0960982202004487.

382. Graves, J. A. M. Sex Chromosome Specialization and Degeneration in Mammals. *Cell* **124**, 901–914 (2006). URL http://www.cell.com/cell/abstract/S0092-8674(06)00241-8.

383. Nishioka, Y. & Lamothe, E. Isolation and characterization of a mouse Y chromosomal repetitive sequence. *Genetics* **113**, 417–432 (1986).

384. Eicher, E. M., Hutchison, K. W., Phillips, S. J., Tucker, P. K. & Lee, B. K. A repeated segment on the mouse Y chromosome is composed of retroviral-related, Y-enriched and Y-specific sequences. *Genetics* **122**, 181–192 (1989).

385. McLaren, A. *et al.* Location of the genes controlling H-Y antigen expression and testis determination on the mouse Y chromosome. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 6442–6445 (1988).

386. Burgoyne, P. S., Mahadevaiah, S. K., Sutcliffe, M. J. & Palmer, S. J. Fertility in mice requires X-Y pairing and a Y-chromosomal "Spermiogenesis" gene mapping to the long arm. *Cell* **71**, 391–398 (1992). URL http://www.cell.com/cell/abstract/0092-8674(92)90509-B.

387. Touré, A. *et al.* A New Deletion of the Mouse Y Chromosome Long Arm Associated With the Loss of Ssty Expression, Abnormal Sperm Development and Sterility. *Genetics* **166**, 901–912 (2004). URL http://www.genetics.org/content/166/2/901.

388. Hendriksen, P. J. *et al.* Postmeiotic transcription of X and Y chromosomal genes during spermatogenesis in the mouse. *Dev. Biol.* **170**, 730–733 (1995).

389. Uchimura, A. *et al.* Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res.* **25**, 1125–1134 (2015). URL http://genome.cshlp.org/content/25/8/1125.

390. Tucker, P. K., Lee, B. K., Lundrigan, B. L. & Eicher, E. M. Geographic origin of the Y chromosomes in "old" inbred strains of mice. *Mamm. Genome* **3**, 254–261 (1992).

391. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014). URL http://dx.doi.org/10.1186/s12859-014-0356-4.

392. Hudson, R. R., Kreitman, M. & Aguadé, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 (1987).

393. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

394. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).

395. Achaz, G. Testing for neutrality in samples with sequencing errors. *Genetics* **179**, 1409–1424 (2008).

396. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring Coalescence Times From DNA Sequence Data. *Genetics* **145**, 505–518 (1997). URL http://www.genetics.org/content/145/2/505.

397. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**, 1791–1798 (1999). URL http://mbe.oxfordjournals.org/content/16/12/1791.

398. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).

399. Auffray, J.-C., Vanlerberghe, F. & Britton-Davidian, J. The house mouse progression in Eurasia: a palaeontological and archaeozoological approach. *Biological Journal of the Linnean Society* **41**, 13–25 (1990). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8312.1990.tb00818.x/abstract.

400. Hofreiter, M. *et al.* Lack of phylogeography in European mammals before the last glaciation. *Proc Natl Acad Sci U S A* **101**, 12963–12968 (2004). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC516467/.

401. Mack, K. L., Campbell, P. & Nachman, M. W. Gene regulation and speciation in house mice. *Genome Res.* **26**, 451–461 (2016). URL http://genome.cshlp.org.libproxy.lib.unc.edu/content/26/4/451.

402. Maly, M. S., Knuth, B. A. & Barrett, G. W. Effects of Resource Partitioning on Dispersal Behavior of Feral House Mice. *Journal of Mammalogy* **66**, 148–153 (1985). URL http://www.jstor.org/stable/1380971.

403. Rowe, F. P., Quy, R. J. & Swinney, T. Recolonization of the buildings on a farm by house mice. *Acta Theriologica* **32**, 3–19 (1987). URL http://rcin.org.pl/ibs/dlibra/docmetadata?id=11218&from=publication.

404. Lidicker, W. Z. Social Behaviour and Density Regulation in House Mice Living in Large

Enclosures. *Journal of Animal Ecology* **45**, 677–697 (1976). URL http://www.jstor.org/stable/3575.

405. Macholán, M. *et al.* Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone? *BMC Evolutionary Biology* **8**, 271 (2008). URL http://dx.doi.org/10.1186/1471-2148-8-271.

406. Webster, T. H. & Wilson Sayres, M. A. Genomic signatures of sex-biased demography: progress and prospects. *Current Opinion in Genetics & Development* **41**, 62–71 (2016). URL http://www.sciencedirect.com/science/article/pii/S0959437X1630106X.

407. Pool, J. E. & Nielsen, R. Population Size Changes Reshape Genomic Patterns of Diversity. *Evolution* **61**, 3001–3006 (2007). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.2007.00238.x/abstract.

408. Hudson, R. R. & Kaplan, N. L. The Coalescent Process and Background Selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **349**, 19–23 (1995). URL http://rstb.royalsocietypublishing.org/content/349/1327/19.

409. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat Genet* **38**, 463–467 (2006). URL http://www.nature.com.libproxy.lib.unc.edu/ng/journal/v38/n4/full/ng1754.html.

410. Li, G. *et al.* Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res.* **23**, 1486–1495 (2013). URL http://genome.cshlp.org/content/23/9/1486.

411. Cocquet, J. *et al.* The Multicopy Gene Sly Represses the Sex Chromosomes in the Male Mouse Germline after Meiosis. *PLOS Biol* **7**, e1000244 (2009). URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000244.

412. Bulatova, N. & Kotenkova, E. Variants of the Y-chromosome in sympatric taxa of Mus in southern USSR. *Bolletino di zoologia* **57**, 357–360 (1990). URL http://dx.doi.org/10.1080/11250009009355719.

413. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

414. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

415. Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* **3**, 475–479 (2012). URL http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00179.x/abstract.

416. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

417. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful

Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol* **57**, 289–300 (1995). URL http://www.jstor.org/stable/2346101.

418. Oehlert, G. W. A Note on the Delta Method. *Am Stat* **46**, 27–29 (1992). URL http://www.jstor.org/stable/2684406.

419. Zammit, S. *et al.* Paternal age and risk for schizophrenia. *The British Journal of Psychiatry* **183**, 405–408 (2003). URL http://bjp.rcpsych.org/content/183/5/405.

420. Petersen, L., Mortensen, P. B. & Pedersen, C. B. Paternal Age at Birth of First Child and Risk of Schizophrenia. *AJP* **168**, 82–88 (2011). URL http://ajp.psychiatryonline.org/doi/abs/10.1176/appi.ajp.2010.10020252.

421. D'Onofrio, B. M. *et al.* Paternal Age at Childbearing and Offspring Psychiatric and Academic Morbidity. *JAMA Psychiatry* **71**, 432–438 (2014). URL http://jamanetwork.com/journals/jamapsychiatry/fullarticle/1833092.

422. McGrath, J. J. *et al.* A Comprehensive Assessment of Parental Age and Psychiatric Disorders. *JAMA Psychiatry* **71**, 301–309 (2014). URL http://jamanetwork.com/journals/jamapsychiatry/fullarticle/1814892.

423. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* **13**, 537–551 (2012). URL http://www.nature.com/nrg/journal/v13/n8/full/nrg3240.html.

424. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012). URL http://www.nature.com/nature/journal/v485/n7397/full/nature11011.html.

425. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012). URL http://www.nature.com/nature/journal/v485/n7397/full/nature10945.html.

426. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012). URL http://www.nature.com/nature/journal/v485/n7397/full/nature10989.html.

427. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014). URL http://www.nature.com/nature/journal/v515/n7526/full/nature13908.html.

428. Gratten, J. *et al.* Risk of psychiatric illness from advanced paternal age is not predominantly from de novo mutations. *Nat Genet* **48**, 718–724 (2016). URL http://www.nature.com/ng/journal/v48/n7/abs/ng.3577.html.

429. Johnson, N. A., Hollocher, H., Noonburg, E. & Wu, C. I. The effects of interspecific Y chromosome replacements on hybrid sterility within the Drosophila simulans clade. *Genetics* **135**, 443–453 (1993).

430. Winking, H., Reuter, C. & Traut, W. Meiotic synapsis of homogeneously staining regions (HSRs) in chromosome 1 of Mus musculus. *Chromosome Res.* **1**, 37–44 (1993).

431. Raymond, C. K. *et al.* Ancient haplotypes of the HLA Class II region. *Genome Res.* **15**, 1250–1257 (2005). URL http://genome.cshlp.org/content/15/9/1250.

432. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**, 881–885 (2012). URL http://www.nature.com/ng/journal/v44/n8/full/ng.2334.html.

433. Antonacci, F. *et al.* Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat Genet* **46**, 1293–1302 (2014). URL http://www.nature.com/ng/journal/v46/n12/full/ng.3120.html.

434. Usher, C. L. *et al.* Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat Genet* **47**, 921–925 (2015). URL http://www.nature.com/ng/journal/v47/n8/abs/ng.3340.html.

435. Suzuki, H., Kurihara, Y., Kanehisa, T. & Moriwaki, K. Variation in the distribution of silver-staining nucleolar organizer regions on the chromosomes of the wild mouse, Mus musculus. *Mol Biol Evol* **7**, 271–282 (1990). URL http://mbe.oxfordjournals.org/content/7/3/271.

436. Zurita, F., Sanchez, A., Burgos, M., Jimenez, R. & de la Guardia, R. D. Interchromosomal, intercellular and interindividual variability of NORs studied with silver staining and in situ hybridization. *Heredity* **78**, 229–234 (1997). URL http://www.nature.com/hdy/journal/v78/n3/abs/hdy199736a.html.

437. Mitchell, A. R. The mammalian centromere: its molecular architecture. *Mutat. Res.* **372**, 153–162 (1996).

438. Guenatri, M., Bailly, D., Maison, C. & Almouzni, G. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J Cell Biol* **166**, 493–505 (2004). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2172221/.

439. Frank, S. A. Divergence of Meiotic Drive-Suppression Systems as an Explanation for Sex-Biased Hybrid Sterility and Inviability. *Evolution* **45**, 262–267 (1991). URL http://www.jstor.org/stable/2409661.

440. Hamilton, W. D. Extraordinary Sex Ratios. *Science* **156**, 477–488 (1967). URL http://science.sciencemag.org/content/156/3774/477.

441. Brandvain, Y. & Coop, G. Sperm should evolve to make female meiosis fair. *Evolution* **69**, 1004–1014 (2015). URL http://onlinelibrary.wiley.com/doi/10.1111/evo.12621/abstract.

442. Meiklejohn, C. D. Heterochromatin and genetic conflict. *PNAS* **113**, 3915–3917 (2016). URL http://www.pnas.org/content/113/15/3915.

443. Weichenhan, D., Traut, W., Kunze, B. & Winking, H. Distortion of Mendelian recovery ratio for a mouse HSR is caused by maternal and zygotic effects. *Genet. Res.* **68**, 125–129 (1996).

444. Mcclintock, B. Chromosome Morphology in Zea Mays. *Science* **69**, 629–629 (1929). URL http://science.sciencemag.org/content/69/1798/629.

445. Rhoades, M. M. Preferential Segregation in Maize. *Genetics* **27**, 395–407 (1942). URL http://www.genetics.org/content/27/4/395.

446. Bayes, J. J. & Malik, H. S. Altered Heterochromatin Binding by a Hybrid Sterility Protein in Drosophila Sibling Species. *Science* **326**, 1538–1541 (2009). URL http://science.sciencemag.org/content/326/5959/1538.

447. Comptour, A. *et al.* SSTY proteins co-localize with the post-meiotic sex chromatin and interact with regulators of its expression. *FEBS J.* **281**, 1571–1584 (2014).

448. Shiu, P. K. T., Raju, N. B., Zickler, D. & Metzenberg, R. L. Meiotic Silencing by Unpaired DNA. *Cell* **107**, 905–916 (2001). URL http://www.sciencedirect.com/science/article/pii/S0092867401006092.