GENETIC ANALYSIS OF MEIOTIC DRIVE SYSTEMS IN THE MOUSE USING GENOTYPING ARRAYS

John P. Didion

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology.

Chapel Hill
2014

Approved by:

Fernando Pardo-Manuel de Villena

Shawn M. Gomez

Gary A. Churchill

Terrence S. Furey

Jeremy B. Searle

ABSTRACT

John P. Didion: Genetic Analysis of Meiotic Drive Systems in the Mouse Using Genotyping Arrays.
(Under the direction of Fernando Pardo-Manuel de Villena.)

Mendel's laws are key to our understanding of genetics and evolution. The Law of Segregation states that alleles at each genetic locus segregate randomly to the gametes such that each parental allele has an equal chance of passing to offspring. Though the processes governing chromosomal segregation are among the best conserved in eukaryotic biology, there are multiple examples of alleles that depart significantly from Mendelian inheritance ratios (transmission ratio distortion, TRD) and cannot be explained by natural selection on organismal fitness. Such deviations are thought to result from intragenomic conflict, in which "selfish" genetic elements have evolved mechanisms to propagate regardless of their effect on fitness. Meiotic drive is a type of intragenomic conflict in which a selfish allele is able to exploit asymmetric meiosis in order to have a significantly greater than random chance of being transmitted to the gamete. In mammals, only female meiosis is asymmetric due to the requirement that most of the cellular volume is transferred to a single haploid oocyte that is able to develop into an embryo upon fertilization. In some meiotic drive systems, non-random segregation is observed at or near the centromere. The Centromeric Drive theory predicts that meiotic drive acting on competing centromeric alleles leads to fixation of chromosomal rearrangements and thereby evolution of karyotypes with possible implications for speciation. In other meiotic drive systems, the locus exhibiting non-random segregation is not in direct linkage to the centromere. In order to characterize genetic factors that influence non-random chromosome segregation, I studied two different populations of the western house mouse (*Mus musculus domesticus*), each of which exhibited a different type of meiotic drive. First, there are over 100 chromosomal races of the house mouse, each of which has fixed one or more Robertsonian (Rb)

translocations (fusions between acrocentric chromosomes). It has been hypothesized that the karyotypic diversity of the house mouse is due to a segregating meiotic drive system in which the ancestral allele favors transmission of acrocentric chromosomes to the oocyte while the derived allele favors metacentric chromosomes. I conducted a genome-wide association study of karyotypically divergent wild mice and identified a locus on Chr 13 that was significantly associated with accumulation of Rb translocations. Second, I characterized a novel meiotic drive system in the Collaborative Cross and Diversity Outbred mouse populations. I found a large region of Chr 2 in a wild-derived strain, WSB/EiJ, is preferentially transmitted during female meiosis when in heterozygosity with alleles from several other classical inbred strains. We identified a promising candidate causal allele, a 127 kb copy number variant with 33 additional copies in WSB/EiJ. We mapped the candidate allele to a 900 kb region that is distal to the single copy that exists in the mouse reference genome. There was striking similarity in both the number of copies and the sequence similarity between WSB/EiJ and SPRET/EiJ, a strain derived from a different species (*Mus spretus*), which also exhibits Chr 2 TRD in crosses with C57BL/6J. I also found that both the presence and level of meiotic drive were variable and dependent on genetic background. Some backgrounds exhibited drive approaching 100%, a level unprecedented in mammalian meiotic drive systems. We identified multiple QTL that approached significant associated with the presence and level of TRD in a relatively small sample of CC and DO hybrid females. This work contributes substantially to the understanding of meiotic drive, provides several important methods, data sets and mouse resources, and may have future implications for evolutionary theory, human health and biotechnological applications.

ACKNOWLEDGMENTS

This research was only possible because of the support given by my mentors, collaborators, lab members and funding sources. There have been many over the years who have contributed in ways large and small. If I have forgotten anyone, it is due to lack of memory rather than appreciation.

First and foremost, I am grateful to my advisor, Fernando Pardo-Manuel de Villena. He has been instrumental in my development as a scientist and in providing the vision for my research.

I have been very fortunate to work with many excellent post-docs, students and staff in the lab: David Aylor, John Calaway, Andrew Morgan, Leeanna Hyacinth, Tim Bell, Darla Miller, Ryan Buus, Justin Gooch, Mark Calaway, Ginger Shaw, Nicole Miller, Sara Cates, Teresa Mascenik, Stephanie Hansen and Jennifer Shockley.

I am grateful to have had funding support throughout my graduate career from the following organizations: Bioinformatics and Computational Biology Training Grant, the Center for Genome Dynamics, and the Center for Integrated Systems Genomics.

I would like to thank the members of my thesis committee for encouragement and fruitful discussions: Shawn Gomez (Chair), Terry Furey, Jason Leib, Gary Churchill and Jeremy Searle.

Finally, I am thankful for the love and support of my family, my partner Kathryn and many great friends.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2N	Diploid Number
BAC	Bacterial Artificial Chromosome
BF	Bayes Factor
CC	Collaborative Cross
CNV	Copy Number Variant
CR	Chromosomal Race
DO	Diversity Outbred
GD	Gametic Disequilibrium
GRP	Genetic Reference Populations
GWAS	Genome-Wide Association Study
HMM	Hidden Markov Model
HWE	Hardy-Weinberg Equilibrium
IBD	Identity By Descent
IBS	Identity By State
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MDA	Mouse Diversity Array
MI	Meiosis I
MII	Meiosis II
MSA	Multiple Sequence Alignment

MUGA	Mouse Universal Genotyping Array
MYA	Million Years Ago
OTV	Off-Target Variant
PAR	Pseudo-Autosomal Region
PCA	Principal Component Analysis
SDP	Strain Distribution Pattern
Rb	Robertsonian
SNP	Single Nucleotide Polymorphism
ST	Standard Population
SV	Structural Variant
TRD	Transmission Ratio Distortion
VINO	Variable Intensity Oligonucleotide
WART	Whole-Arm Reciprocal Translocation

Chapter 1

BACKGROUND AND INTRODUCTION

This work examines the phenomenon of meiotic drive in the context of two different systems in the house mouse. The primary aim of these studies is to identify genetic factors that contribute to the presence and variability of non-random chromosomal segregation during female meiosis (meiotic drive). In this chapter, I give an introduction to meiotic drive, followed by an overview of these two studies.

1.1 Meiosis and meiotic drive

Since their rediscovery in the early 20th century, Mendel's Laws have formed the theoretical basis for our understanding of the genetics of inheritance. The Law of Segregation ("First Law") states that alleles at homologous loci segregate from each other during meiotic cell division and are transmitted randomly to gametes. The Law of Independent Assortment ("Second Law") states that alleles at unlinked loci segregate independently from each other. Together, Mendel's Laws create one of the strongest predictions in biology: that, for sexually reproducing species, each individual's genome is a random collection of alleles equally derived from its mother and father.

As scientists have unraveled the underlying mechanisms of inheritance, they have also discovered exceptions requiring the expansion and revision of Mendel's Laws. For example, the discovery of a direct relationship between the physical proximity of two genes on a chromosome and the likelihood of their co-segregation (genetic linkage) required an amendment

to the Law of Independent Assortment that we now take for granted. Early geneticists also discovered that some genes violated the Law of Segregation [1, 2]. When a violation of the First Law is found to be significant and reproducible, regardless of the cause, it is referred to as transmission ratio distortion (TRD) [3]. Most observations of TRD are due to selection in favor of alleles that increase the fitness of individuals with respect to their environment (ecological selection) or their sexual competitors (sexual selection). Selection may also act upon the products of meiosis (gamete selection) or fertilization (differential embryonic survival). However, an increasing number of observations of TRD can be ascribed to competition between "selfish" genetic elements, which promote their own preferential transmission irrespective of their effects on individual fitness. This so-called intragenomic conflict has been observed in a wide variety of eukaryotic species [4, 5, 6]. Intragenomic conflict can take many forms, but generally follows one (or more) of three strategies: 1) *interference*, in which an allele prevents the transmission of other alleles or prevents the carriers of alternate alleles from passing them on; 2) *overreplication*, in which an allele increases its chances of being transmitted by increasing its prevalence in the genome, by duplication or transcriptional upregulation; or 3) *gonotaxis*, in which an allele moves preferentially into the genetic material that is passed on to subsequent generations [7].

Meiosis is the process by which the germ cells of multicellular, sexually reproducing eukaryotes give rise to the gamete cells, which in turn fuse during fertilization and transmit the genetic information from parents to their offspring. Although the processes governing meiotic chromosomal segregation are among the most well-conserved features of eukaryotic biology [8], there are examples in many species of genes that selfishly subvert the redundancies and safe-guards. Meiotic drive is a type of intragenomic conflict (specifically gonotaxis) that results in the differential inclusion of parental alleles in the products of meiosis [4].

Meiosis is well characterized at the cellular level (reviewed in [9]). Figure 1.1 presents a simplified schematic of a typical mammalian meiosis. The stages of meiosis that are important to the understanding of meiotic drive are 1) synthesis (S phase), in which a duplicate copy of

the entire genome is made and each chromosome becomes a complex of two identical sister chromatids; 2) meiosis I (MI), in which each pair of homologous chromosomes is segregated into two haploid daughter cells; and 3) meiosis II (MII), in which each daughter cell divides and the sister chromatids of each chromosome are segregated. An important distinction between MI and MII is that during prophase of MI (prophase I), homologous chromosomes pair (synapsis) and attach to one another at points called chiasmata. When the meiotic spindle pulls the homologous chromosomes apart, some chiasmata are resolved as crossovers that result in the exchange of genetic material between the homologous chromosomes (recombination).

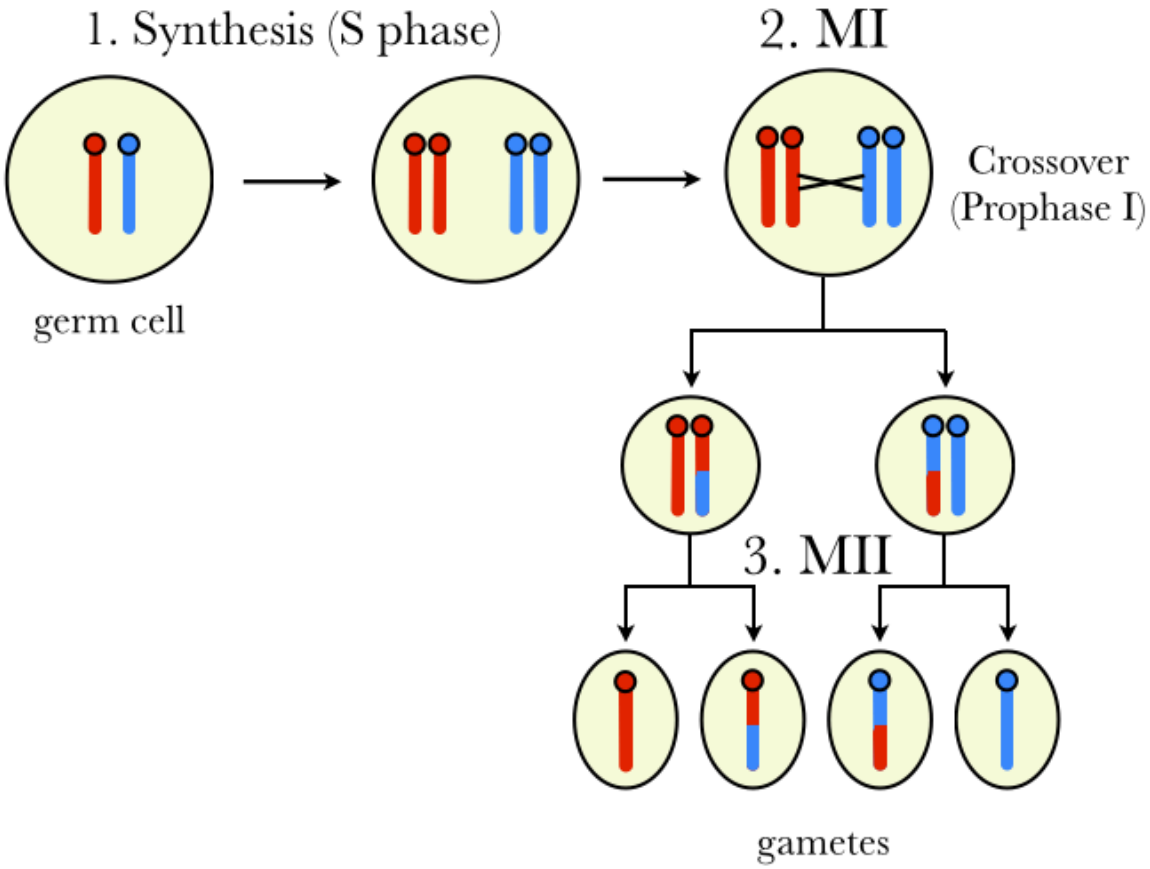


Figure 1.1: Simplified schematic of a mammalian meiosis. Two parental chromosomes (red and blue) are replicated during S phase (1). Homologues crossover at least once per chromosome arm during the first meiotic cell division (2). Finally, a second meiotic division (3) results in haploid gamete cells.

In all but a few known exceptions [10], male and female meiosis differ with respect to their resulting products. In male meiosis, each cell division results in two viable daughter

cells of roughly equal size. In contrast, each cell division in female meiosis results in one viable product containing the vast majority of the volume of the progenitor cell, and one non-viable "polar body." Therefore, each female meiosis results in a single gamete (ovum, or egg), while each male meiosis results in four gametes (spermatids). This difference is primarily due to the requirement that the ovum, in addition to carrying the maternal genetic complement, must carry all of the material required for development into a new organism as well as a protective enclosure within which the new organism may develop. Intuitively, it is a better strategy for a female germ cell to put all of its energy into creating a single, robust gamete rather than four smaller viable gametes. However, the asymmetry of female meiosis presents an opportunity for selection: if one allele of a locus that is directly involved in meiotic chromosomal segregation has greater "fitness" than its homologue, it can get into the ovum more than 50% of the time, thereby increasing its frequency. Meiotic drive that results from an allele exploiting asymmetric female meiosis to gain a segregation advantage is called female meiotic drive.

1.2 Requirements for meiotic drive

A survey of meiotic drive systems revealed that the necessary and sufficient requirements for drive are only three: 1) meiotic divisions that are asymmetrical with respect to cell fate; 2) functional asymmetry of the meiotic spindle poles; and 3) functional heterozygosity at a locus that mediates attachment of a chromosome to the meiotic spindle [11] (Figure 1.2).

Asymmetric cell division solves a common need for some cells to produce daughter cells with differing contents, and is pervasive in the tree of life. In higher eukaryotes, cellular differentiation and neurological development are dependent on asymmetric mitotic division. However, meiosis is unique in its need to segregate homologous chromosomes during MI; therefore, meiotic cell division is quite different from all other asymmetrical cell divisions (reviewed in [9]). Most importantly, the meiotic spindle must first attach to the cellular cortex so that the metaphase plate is parallel to the cortex. A small area of the cortex buds

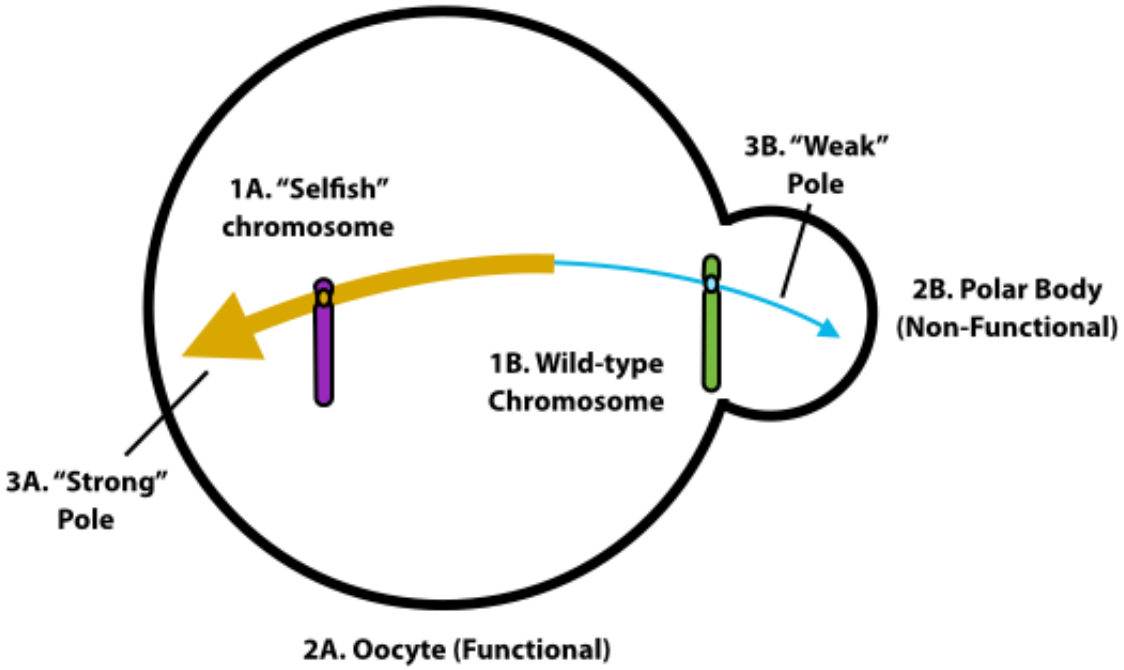


Figure 1.2: Meiotic drive is the non-Mendelian segregation of functionally different chromosomes (1) that depends on asymmetric female meiosis (2) and the inequality of meiotic spindle poles (3). Adapted from [11].

out during cytokinesis to encapsulate the chromosomes on the externally facing side of the metaphase plate, forming a polar body that degenerates and may be reabsorbed. In contrast to mammalian females, females of some plant and insect taxa undergo two successive meioses without cytokinesis. The meiotic products are arranged linearly, and the innermost product is retained by the oocyte while the remaining three products degenerate.

While the asymmetry of female meiotic cell division is a well established fact, asymmetry of the meiotic spindle is less well established. The best evidence of spindle asymmetry comes from the study of B chromosomes. B's are parasitic, supernumerary chromosomes that are prevalent in many plant, insect and rodent species. They likely derive from degenerate normal (A) chromosomes, and the majority have deleterious phenotypic effects. Studies in grasshoppers have shown that the two poles of the meiotic spindle differ in size. B chromosomes do not pair during meiosis, and they attach to the meiotic spindle in a volume-dependent manner [12]. Interestingly, it has also been observed that unpaired X chromosomes in XO female

mice preferentially segregate toward the functional meiotic product. It is speculated that preferential segregation of unpaired chromosomes may be due to a meiotic failsafe that, in the presence of an unequal number of centromeres, prefers to retain more genetic material in the oocyte [13].

The third requirement for meiotic drive may be satisfied by a selfish genetic element that, when in heterozygosity, succeeds in being transmitted to the functional product of asymmetric meiosis more than 50% of the time. B chromosomes are only one of several types of loci that exhibit drive.

1.3 Types and examples of meiotic drive systems

Meiotic drive has been reported in several species (Figure 1.3), and has been the subject of much study. The loci at which meiotic drive is observed (*responder* loci) may be classified by their chromosomal position with respect to the centromere, which determines the phase of meiosis during which drive may occur and also the level of TRD that may be observed. In theory, *distorter* loci (the loci that induce non-random segregation) may be located anywhere in the genome, although in nearly all meiotic drive systems in which a *distorter* has been identified it is tightly linked to the *responder*.

1.3.1 Drive at MI: centromeric drive

The simplest meiotic drive system is a single *responder* with two alleles that have different fitness with respect to their ability to transmit to the ovum during meiosis. In sexually reproducing species, the centromere is the ideal single-locus system. Centromeres typically consist of at least 500kb of tandem repeats surrounded by pericentric heterochromatin that may be interspersed with functional sequence. Centromeres are epigenetically defined, and are marked by nucleosomes that incorporate a centromere-specific histone. Centromeres mediate the attachment of chromosomes to the meiotic spindle via a protein complex called the kinetochore. Therefore, any centromere that can preferentially attach to the spindle pole directed toward

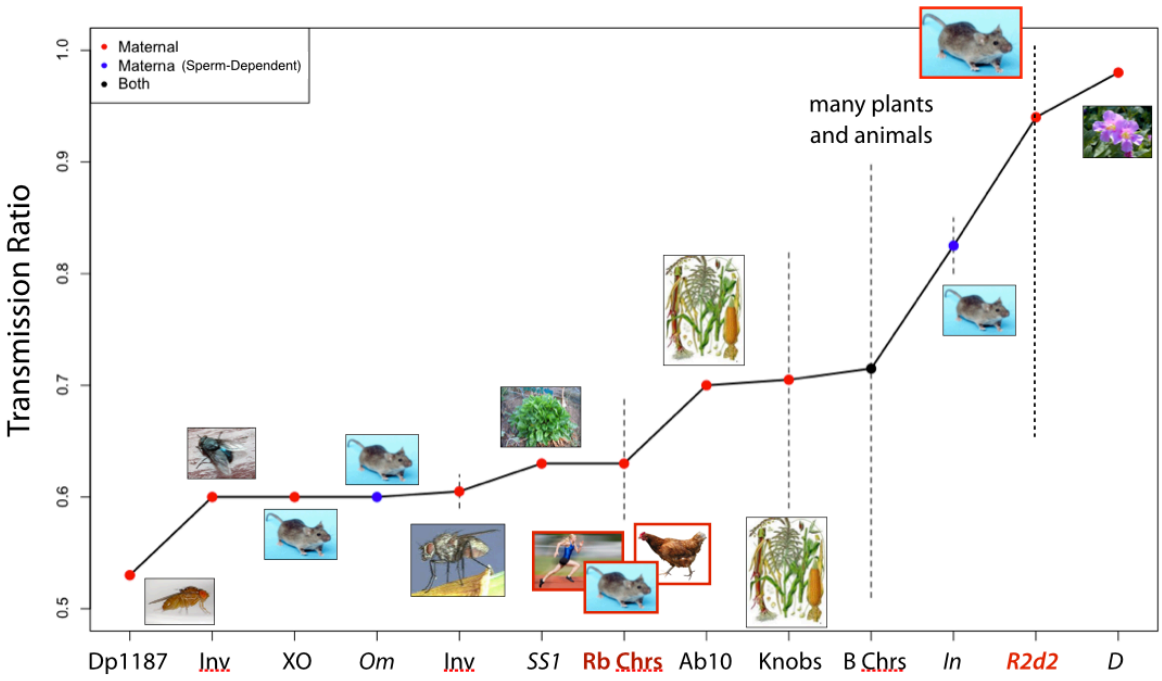


Figure 1.3: Summary of meiotic drive systems in plant and animal species. Colored points indicate the sex specificity of drive: female-only (red), female drive induced by a genetic factor in the fertilizing sperm (blue), and sex-independent. Dotted lines indicate a range of reported transmission ratios. The systems involved in the research presented in Chapters 3 and 4 are highlighted in red.

the functional cell product will increase its frequency in the population (centromeric drive). Centromeric drive must occur at MI since it is only in the primary oocyte that there may be homologous chromosomes with centromeres of different fitness (Figure 1.1). It is thought that centromeric drive may result from an "arms race" between competing centromeric repeat sequences [14].

Evidence of driving centromeres in A chromosomes is that certain chromosomal fusions segregate non-randomly during meiosis. A Robertsonian (Rb) translocation occurs when two chromosomes with terminal centromeres (acrocentrics) fuse at or near their centromeres to form a single metacentric chromosome. During meiosis in females that are heterozygous for a Rb translocation, the metacentric fusion chromosome typically pairs with the two homologous acrocentric chromosomes. The Mendelian expectation is of equal transmission of the Rb translocation and of the two acrocentric homologs; however, TRD of Rb translocations

has been observed in humans [15] and mice (reviewed in [16]) at levels $\sim 60\%$. Drive is observed independent of which chromosomes are fused. Strikingly, the direction of segregation distortion is not consistent: in mice, the acrocentric chromosomes are preferred over the metacentric, while the opposite is true in humans. In both cases, the direction of drive is consistent with the predominant chromosomal form of the species. Conceptually equivalent results are observed in chickens: heterozygous female carriers of chromosome fissions preferentially transmit the metacentric $\sim 70\%$ of the time [17]. The common feature of meiotic drive involving all of the mentioned chromosomal rearrangements and abnormalities is the unequal number of centromeres on either side of the metaphase plate during meiotic division (the unequal centromere number rule [16]). Therefore, it is most likely that drive is acting on the centromeres themselves rather than any particular DNA sequence. It is theorized that drive in favor of either greater or lesser numbers of chromosomes is a common feature across the tree of life. Furthermore, the direction of drive appears to undergo frequent reversal over evolutionary time [16], sometimes even within otherwise genetically homogeneous species. This can be observed most dramatically in the extreme karyotype diversity within *M. m. domesticus*, discussed in Chapter 3, although intraspecific karyotype variation due to Rb translocations is also known in several other small mammals, e.g., shrews [18]. Meiotic drive of Rb translocations is predicted to contribute to karyotype evolution, especially in species in which Rb translocations are the predominant type of chromosome structural change [19].

Near-complete preferential segregation of centromeres during MI is a theoretical possibility, but has not been observed until quite recently. In reciprocal backcrosses between two species of *Mimulus* (monkeyflower), transmission ratios of the *M. guttatus* allele of up to 99% were observed only when the female was heterozygous for a novel *responder* locus (*D*) [20]. The experimental design enabled the authors to rule out alternate explanations such as cytoplasmic effects and post-meiotic effects on seed development. The offspring were fertile, so it was possible to show that drive persisted in subsequent generations. The authors also examined seed development across multiple generations and found no evidence of reduced

female fecundity, meaning that the observed TRD was not due to lethality associated with not having a *M. guttatus* allele at the *D* locus. Since the theoretical limit on the maximum distortion that may occur at a non-centromeric locus ($\sim 83\%$, [21]) is less than the magnitude of drive observed at *D*, either *D* is the centromere (or tightly linked to the centromere), or *D* is driving at both MI and MII. Although an assembly of the *Mimulus* genome was not available, and thus the physical proximity of the *D* locus to the centromere was unknown, a likely *Mimulus* centromeric repeat sequence maps to the *D* locus [22]. Furthermore, the *D* locus was greatly expanded in size in species exhibiting TRD. There was also low polymorphism near *D*, with LD extending up to 2 cM, which strongly suggests that *D* is either the centromere or is bounded by structural variants that suppress recombination.

1.3.2 Drive at MII

Centromeres are not the only loci that may drive. In fact, the majority of female drive systems that have been described involve *responder* loci that are quite distant from a centromere. However, this disparity may be due to the greater difficulty of observing TRD at centromeric vs. non-centromeric loci than to any relative difference in frequency. Systems that involve distal *responders* tend to share several features in common. First, the *responders* tend to be located in large heterochromatic regions [23, 24]. In the few cases where the sequence of these loci has been examined, they appear to consist primarily of tandem repeats that have some homology to centromeric repeat sequences. These large, centromere-like regions appear to actually function as centromeres during meiosis (neocentromeres). In many cases it has been shown or predicted that structural variants are involved in suppressing recombination within the loci. Activation of neocentromeres during meiosis may be a specific instance of centromere repositioning, which appears to be frequent in mammals and may play an important role in karyotype evolution [25, 26].

By far the best studied case of female drive is the Ab10/knob system in *Zea mays* (maize) [23, 27, 21]. An abnormal variant of maize Chr 10 (Ab10) contains highly repetitive hete-

rochromatic sequences (knobs) distinct from the chromosome's normal centromere. In maize, female meiosis results in the four haploid products extending from the ovary in a row; it is only the basal product (closest to the ovary) that develops into a gamete, while the other three degenerate. The knobs of Ab10 are able to function like centromeres during meiosis and interact with the meiotic spindle to greatly increase the chance of the Ab10 homologues being outermost in the ordering of meiotic products, and thus of one of the two being basal. The knobs are active during both male and female meiosis, but it is only in female meiosis that the ordering matters since all of the male meiotic products are viable. The large knob is comprised of a 180bp repeat [28]. There are also three additional supernumerary regions on Chr 10 that function as additional neocentromeres during meiosis and consist of tandem repeats of a different, 350bp motif (TR-1). There is evidence of intragenomic conflict between the two types of repeats [29]. Cytogenetic studies have shown that neocentromeres only replicate the ability of true centromeres to move along the meiotic spindle; true centromeres and neocentromeres are otherwise functionally different [30].

Another common feature of non-centromeric *responders* is that they tend to be located at roughly the same genetic distance from the centromere. The positioning of those loci strikes a compromise between two factors that are both necessary for drive at MII but which are oppositely affected by the distance of the locus from the centromere. First, an odd number of crossovers (typically just one) are required between the centromere and the drive locus in order for the dyad to be heteromorphic at MII. The greater the distance between the centromere and the *responder*, the greater the chance of crossover. Second, the *responder* presumably needs to coordinate with the centromere in order to attach to the same spindle pole [21, 30]. Attachment to different spindle poles would result in chromosome bridging and breakage, leading either to loss of the chromosome in the meiotic product or checkpoint-mediated termination of meiosis altogether. The closer the drive locus is to the centromere, the greater the chance that it will be in (three-dimensional) physical proximity to the centromere during spindle attachment. It has been speculated that the success of Ab10 is due at least in part to the fact that Chromosome

10 is the shortest in maize and thus provides the greatest probability of having exactly one recombination between the centromere and large knob [7]. Another consideration is that recombination is undesirable near or within the region containing the genetic elements that are essential for drive. Ab10 contains several structural rearrangements, a large insertion and two nested inversions, that are proposed to suppress recombination in the distal part of the chromosome. All of these facts indicate that chromosomes are subject to natural selection on the position of meiotic drive loci, and that a position at ~ 50 cM leads to chromosomes of the greatest "fitness" in terms of meiosis, even if not in terms of overall organismal fitness.

1.3.3 Modifiers of meiotic drive

In most cases, characterizations of meiotic drive systems have only identified the *responder*. However, there is evidence to suggest that meiotic drive is a complex trait that is subject to modification by interacting elements. In Ab10, there are at least four key *distorters* that are required for drive, at least some of which must be *trans*-acting: a neocentromere-activating gene, a locus that enhances recombination between the large knob and the centromere and two loci of unknown function, one immediately proximal (*smd1/3*) and the other immediately distal to large knob (*DTF*), in which mutations have been shown to cause a loss of drive. Recombination modifiers could be common features of meiotic drive systems: a recombination modifier tightly linked to a drive locus could increase its own transmission frequency by decreasing or increasing the rate of recombination (depending on whether drive occurs at MI or MII) [31]. The neocentromere-activating gene has been parasitized by knobs located on other chromosomes of the maize genome. Knobs have been found on all 10 chromosomes and all but three chromosome arms in the maize genome [27]. All knobs show drive in the presence of Ab10, with transmission rates from 59% to 82% for different knob sizes and positions [32]. Transmission rates appear to be positively correlated with knob size [21]. When knobs of different sizes compete the larger knob appears to win (i.e., be transmitted more frequently), probably due to greater attachment to molecular motors that move the chromosome along the

meiotic spindle [30].

The only evidence for unlinked modifiers of meiotic drive come from the study of B chromosomes in the Mealybug, *Pseudococcus afinis*. Nur and Brett (1987) [33] found evidence of unlinked loci associated with different rates of B chromosome transmission, however it is not clear whether the distortion was meiotic or post-meiotic. On the other hand, a study of B chromosomes in rye identified modifiers originating from the B chromosomes themselves; however, it was unclear if the modifiers acted in *cis* only, or if a modifier on one B chromosome was able to influence the segregation of other B chromosomes in *trans* [34].

1.4 Genetic characterization of meiotic drive in the mouse

For several reasons, the genetic factors, molecular mechanisms and evolutionary consequences of most meiotic drive systems are still poorly understood. First, meiotic drive mechanisms act within a short time window. It is technically challenging to capture the dynamics of a system that is active only during female meiosis, which in mammals occurs primarily during embryo development. Most cytogenetic evidence has come from *in vitro* studies, and even then it has only been possible to capture a static picture of a process that is highly dependent on the dynamic interaction between chromosomes and the meiotic spindle and on the orientation of the spindle to the rest of the cell. Second, meiotic drive systems invariably involve large and highly repetitive genetic elements that are intractable to sequencing, cloning and manipulation. Third, by nature non-random transmission promotes the fixation of the observed allele. This limits the instances of meiotic drive systems that may be observed in nature to those we are lucky enough to catch by chance, or to those in which the deleterious effects of drive have prevented complete fixation. Fourth, meiotic drive systems are often confounded with additional factors, such as changes in fertility, fecundity and survival. In some cases these factors are directly linked to the meiotic drive system [35], though in most cases they are instead due to unrelated genetic incompatibilities that are characteristic of the crosses and natural populations in which meiotic drive is observed.

The house mouse is a good model for expanding our knowledge of the genetic components of meiotic drive. The relatively large number of meiotic drive systems described in the mouse indicate that meiotic drive is at least as frequent in the mouse as in any other species (Figure 1.3). Mice are abundant and relatively easy to capture in the wild, and there are a wealth of reproducible, genetically divergent mouse stocks for laboratory experiments. The mouse is also one of the most popular laboratory model organisms, which has encouraged the development of extensive genetic, cytogenetic, molecular and bioinformatic tools. A significant fraction of my time in the lab has been dedicated to developing technologies and bioinformatic tools to better study natural mouse populations; I discuss these in the next chapter.

My first experimental study (Chapter 3) involved a widely studied set of natural populations. Each of the ~ 100 chromosomal races (CRs) of the house mouse has a different, non-standard karyotype due to the fixation of one or more metacentric chromosomes that arose by Rb translocation. It has been proposed that meiotic drive is the primary mode of fixation, and that the direction of drive in *M. m. domesticus* has changed such that metacentric chromosomes are selected for, rather than against, during meiosis in females of the CRs. However, there has been no direct evidence of genetic factors that are involved in the change in the direction of drive. The aim of my first investigation was to assemble a large catalogue of genetic variation in wild mice, and to mine that data to determine whether any genetic loci were associated with the accumulation of Rb translocations.

In laboratory populations, reports of TRD are common in experimental crosses [36, 37] and may be directly studied to uncover the underlying mechanism. The aim of my second experimental study (Chapter 4) was to determine the mechanism underlying multiple observations of TRD of a wild-derived allele in a laboratory population, the Collaborative Cross [38, 39]. I was able to show that the observed TRD was due to meiotic drive, and furthermore that there are several unlinked *distorters* that control the presence and level of TRD.

The ability to compare and contrast these two quite different meiotic drive systems has yielded valuable insights that I discuss in my final chapter.

Chapter 2

DESIGN AND USE OF GENOTYPING ARRAYS FOR GENETIC ANALYSIS OF WILD AND INBRED MICE[1]

2.1 The house mouse, *Mus musculus*

The house mouse, *M. musculus*, is a monophyletic species that arose in central and south Asia ~ 1 MYA [43]. Between 0.25 and 0.5 MYA [44, 45], the mouse began to diverge into three distinct subspecies: *M. m. domesticus*, whose ancestral range extends westward from Turkey, throughout the Mediterranean basin, and northward to Scandinavia; *M. m. musculus*, whose ancestral range extends from eastern Europe to China; and *M. m. castaneus*, whose ancestral range is India and southeast Asia (Figure 2.1). The subspecies interact at several known hybrid zones, the largest of which extends north to south across the whole of Europe (Figure 2.2). With the development of agriculture $\sim 10,000$ years ago, mice became human commensals, and have since become established on nearly every landmass that has been vis-

[1]The work described in this chapter was accomplished in collaboration with Hyuna Yang, Gary Churchill, Chen-Ping Fu, Catherine Welch, Katy Kao and Leonard McMillan. The aim of this work was to develop efficient, low-cost, high-throughput genotyping methods capable of characterizing the genetic diversity in wild and laboratory mice while mitigating the effects of SNP ascertainment bias. This work is presented in multiple articles that have either been published or are in preparation. In Yang *et. al.* 2011 [40], I conducted sequencing and data analysis to characterize VINOs, a novel class of marker that is critical in the analysis of array data for wild mice. In Didion *et. al.* 2012a [41], I conducted all bioinformatic analysis to explore the effects of unaccounted-for variation on array data. In Didion *et. al.* 2013 [42], I was invited to write a review of the relationship between wild and laboratory mice, which highlighted the work in the previous two papers. In Fu, Didion, Welsh, *et. al.* (in prep), I contributed to the design of the MegaMUGA array, developed QC methods, and conducted experiments to demonstrate the uses of the array. I have applied these methods in several other publications: Aylor *et. al.* 2011 [38], Crowley *et al.* (submitted), Calabrese *et. al.* (submitted) and Chandler *et. al.* (in prep). In Didion *et. al.* (in prep), I present a software package for cell line validation using SNP arrays, and applies the method to the characterization of more than 100 mouse cell lines (in collaboration with Sandy Morse).

ited by human vessels. Many hybrid populations of mice have been observed as the result of secondary contact. For example, the *M. m. molossinus* hybrid subspecies in northern Japan has a mixed genome resulting from contact between *M. m. musculus* and *M. m. castaneus* [46]. Finally, the taxonomic statuses of two more recently identified populations, *M. m. gentilulus* [47] and *M. m. homoulus* [48], are still being determined.

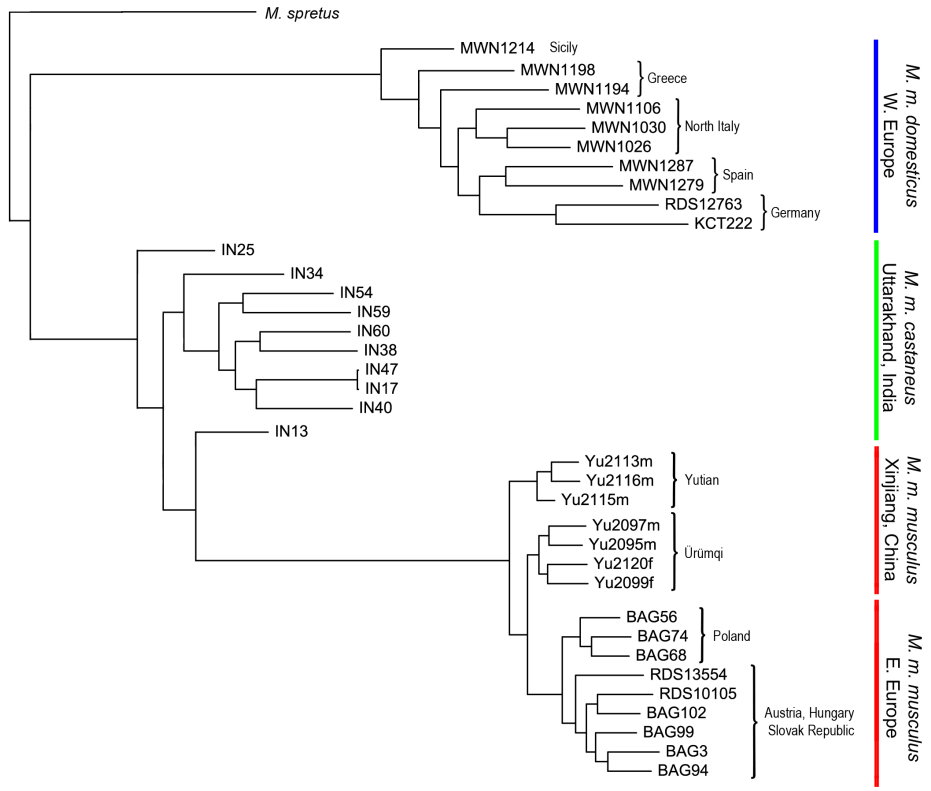


Figure 2.1: The single best maximum-likelihood tree for the phylogeny of *M. musculus*. I used RAxML [49] to analyze genotypes for 547,406 SNP markers and 118,733 VINO markers from 36 wild-caught *M. musculus* samples (10 *M. m. domesticus*, 16 *M. m. musculus* and 10 *M. m. castaneus*) [40] and a single sample of the wild-derived *M. spretus* strain SPRET/EiJ [42]. Colored lines denote subspecific clades. Blue: *M. m. domesticus*; green: *M. m. castaneus*; red: *M. m. musculus*. Geographic origin of samples is given for *M. m. domesticus* and *M. m. musculus*; all *M. m. castaneus* samples are from the state of Uttarakhand, India.

Mice were commonly kept as pets during the 19th and early 20th century. Mouse "fanciers" in Europe and Japan bred mice for certain attractive traits, such as coat color and interesting behaviors. The first laboratory inbred strains were derived from fancy mice. It is now known

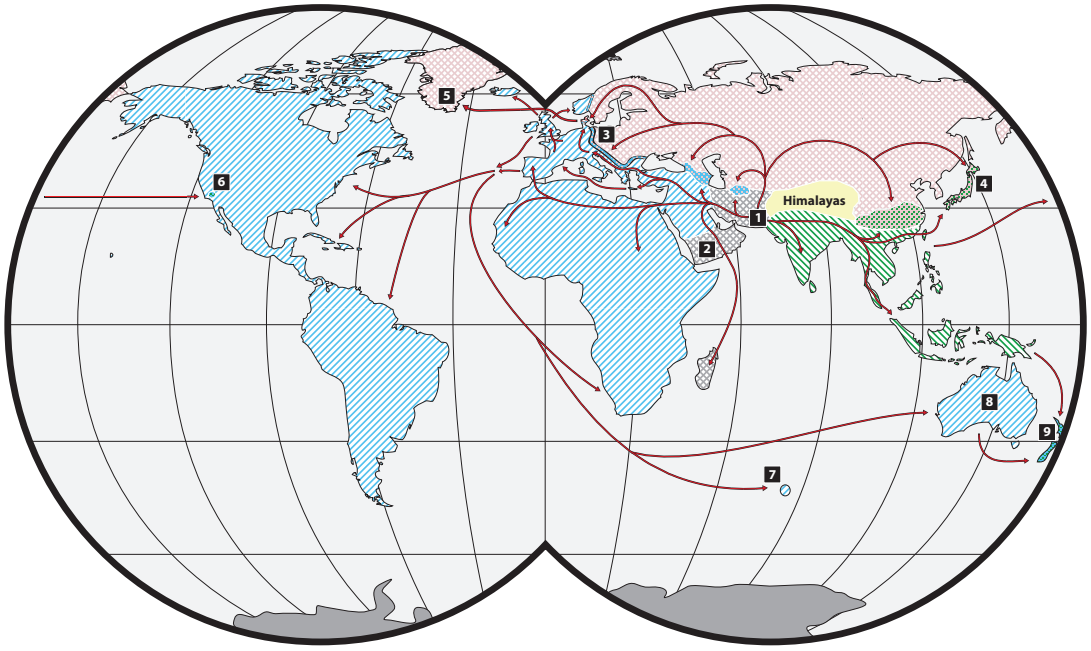


Figure 2.2: Origin and historical migrations of *M. musculus* subspecies. Hatching shows the ranges of *M. musculus* subspecies. Blue: *M. m. domesticus*; red: *M. m. musculus*; green: *M. m. castaneus*; gray: central populations (and '*gentilulus*'). Note that mice may not be found throughout the complete extent of the indicated ranges; for example, in sub-arctic areas of Russia and Canada, in the Amazon rainforest, and in the Sahara. Checkered areas represent zones of hybridization between subspecies. Red arrows represent historical migrations and commensal movements. Black boxes note interesting or important populations; see [42] Supplementary Figure 1 for a full description.

that the founders of those original lines were few in number, and their genome was a mixture of all three subspecies, probably due to mixing between European (*M. m. domesticus*) and Japanese (*M. m. molossinus*) fancy mice [50, 40]. More recently, additional laboratory strains have been derived from wild-caught mice with the goal of increasing the available genetic and phenotypic diversity [51, 42]. The Collaborative Cross (CC) is a new genetic reference panel developed from both classical and wild-derived strains [39].

2.2 SNP genotyping arrays

Single nucleotide polymorphisms (SNPs) account for a substantial fraction of the genetic variants that differentiate individuals and species. SNPs are causal for most Mendelian (i.e., single-variant, high-penetrance) traits, and they also contribute to most complex traits. SNPs are valuable as genetic markers in linkage mapping and association studies because of their quantity and stable inheritance over generations. For similar reasons, SNPs are useful in population genetic and evolutionary studies to determine the relationships between individuals, populations or taxa. The HapMap project has generated a rich, highly-annotated catalogue of human SNPs, including an estimation of their frequency in different human populations [52]. While an equivalent resource does not exist for natural populations of the mouse, several large-scale sequencing and genotyping efforts have nonetheless identified a large fraction of the SNPs present in laboratory strains [53, 54, 55, 40, 56].

SNPs are discovered by comparative analysis of homologous sequences across multiple genomes; the more divergent the genomes, the more SNPs will be identified. Modern SNP discovery efforts generally depend on the shotgun sequencing and assembly of a reference genome, although reference-free methods are sometimes used for non-model organisms [57]. A reference sequence facilitates the alignment and comparison of sequences from multiple individuals (multiple sequence alignment, MSA). MSA allows one not only to identify SNPs, but also to estimate the frequency of each allele within the surveyed population. The first catalogue of genetic variation of the mouse was assembled from comparative analysis of short

sequences in public databases [58].

Once a SNP has been identified, several types of assays may be used to determine the alleles (genotype) of additional samples. The most popular methods include restriction fragment length polymorphism (RFLP), chain-termination (Sanger) sequencing and hybridization. Hybridization assays rely on sequence-specific interactions between complementary nucleotide sequences. The greater the number of inconsistencies between the two strands, the lower their probability of binding. To create a hybridization assay, probe sequence that incorporate known SNPs are synthesized and immobilized (generally by attachment to a glass or silicon substrate). The probes and a DNA sample of interest are then exposed under appropriate conditions. Excess DNA is washed off the substrate, leaving behind only those sequences that are bound to probes. Probes with and without bound complements can be distinguished using fluorescent labels, which may be ligated to the probe sequence or may be added after hybridization by single base extension. Early hybridization arrays were created one at a time in the lab and generally only contained probes for a handful of SNPs; currently, there are several companies that manufacture high-density arrays with probes for thousands to millions of SNPs. High-resolution cameras are used to quantify the fraction of probes for each SNP that have hybridized, which is reported as the hybridization intensity value. A wide array of computational methods have been developed to convert continuous intensity values into discrete genotype calls. In addition to SNPs, hybridization arrays may also be used to assay copy-number variants (CNVs), structural variants (SVs) and differential DNA methylation.

2.3 The Mouse Diversity Array

The mouse is one of the most widely used model organisms in genetic studies, and thus the characterization of the mouse genome and genetic variation has been a high priority. The mouse was the second mammalian genome to be sequenced [53], after the human genome. The mouse reference genome is based on a single classical inbred strain, C57BL/6J. Recently, 18 additional strains were sequenced [56, 59]. To date, all SNP discovery projects have uti-

lized inbred laboratory strains. The majority of those strains were classical inbred lines, although five wild-derived strains have also been resequenced: WSB/EiJ (*M. m. domesticus*), PWK/PhJ and PWD/PhJ (*M. m. musculus* and highly related), CAST/EiJ (*M. m. castaneus*) and SPRET/EiJ (*M. spretus*, the species most closely related to *M. musculus*).

When the first large catalogue of mouse SNPs became available as a result of the NIEHS/ Perlegen resequencing project [55], the community naturally recognized the opportunity for large-scale genetic studies based on SNP arrays. The Mouse Diversity Array [60] (MDA), developed in collaboration between the Churchill and Pardo-Manuel de Villena labs, was the first widely available, high-density SNP array for the mouse. The MDA was designed to capture the known genetic diversity present in laboratory strains using a largely unbiased approach to SNP selection. In addition to the Perlegen data set, SNPs were selected from several additional sources, including bacterial artificial chromosome (BAC)-end sequences from MSM/Ms, a *M. m. molossinus*-derived inbred strain and sequences from public databases. Each category of SNP was placed with uniform spatial distribution. The final array included 623,124 SNP and 916,269 invariant (including CNV) probes.

The MDA and similar arrays based on the Affymetrix platform use genome-wide sampling to reduce genomic complexity by size-selective amplification of restriction fragments [61]. Efficient hybridization requires genomic DNA targeted by a probe set to fall within at least one restriction enzyme fragment in the selected size range (50 bp to 1 kb). The MDA was designed to use a combination of two restriction enzymes, *NspI* and *StyI*, and fragment sizes were predicted based on the mouse reference genome (NCBI mouse genome Build 36).

2.4 Ascertainment bias

SNP discovery methods are inevitably focused on a limited set of individuals, populations or clades. This creates a bias in SNPs available for array designs (ascertainment bias). Furthermore, many arrays are designed using an iterative process that selects only probes that perform well across a screening set of samples. This is done to ensure low miscall and no-call

rates, but also compounds ascertainment bias. Miscall and no-call rates can vary greatly depending on the composition of samples, and are positively correlated with genetic divergence from the reference sequence used to design the array [52, 41]. When SNP probes are excluded from analyses due to post-hoc filtering based on no-call rate, unexpected heterozygosity or departure from Hardy-Weinberg equilibrium, important information is lost (discussed below). In a recent genome-wide analysis of a large number of dog breeds, over 50% of SNPs were excluded for such reasons [62]. The cumulative effect of these SNP selection procedures can potentially skew the interpretation of experimental results and limit researchers' ability to effectively study genetically divergent samples. The MDA was designed with attention to the phylogenetic origin of SNPs, but SNP selection will still introduce some biases, especially in studies that include wild-derived strains or wild-caught mice [40].

2.5 Variable intensity oligonucleotides (VINOs)

Genotype calling programs use a variety of methods to infer discrete genotypes from continuous intensity data. Many methods, including the standard Affymetrix algorithm (`BRLMM-P 2D` [63]), employ clustering of multiple samples based on the contrast between allelic probe intensities. Samples belonging to the two clusters with a large absolute contrast are called as homozygous genotypes and samples with low contrast are called heterozygous. Samples that do not fall within any of the three clusters in the contrast dimension remain uncalled (Figure 2.3).

We previously genotyped 162 laboratory mouse strains using the MDA [40]. Contrary to our expectation of homozygosity at all SNPs in inbred mouse strains, we observed a substantial number of heterozygous genotype calls. Furthermore, the rates of both no-calls and unexpected heterozygous calls were positively correlated with divergence from the reference genome. The highest rates were observed in strains derived from species of the *Mus* genus other than *Mus musculus*, such as *M. spretus* and *M. spicilegus*, followed by strains derived from the *M. m. musculus* and *M. m. castaneus* subspecies. These findings were indicative

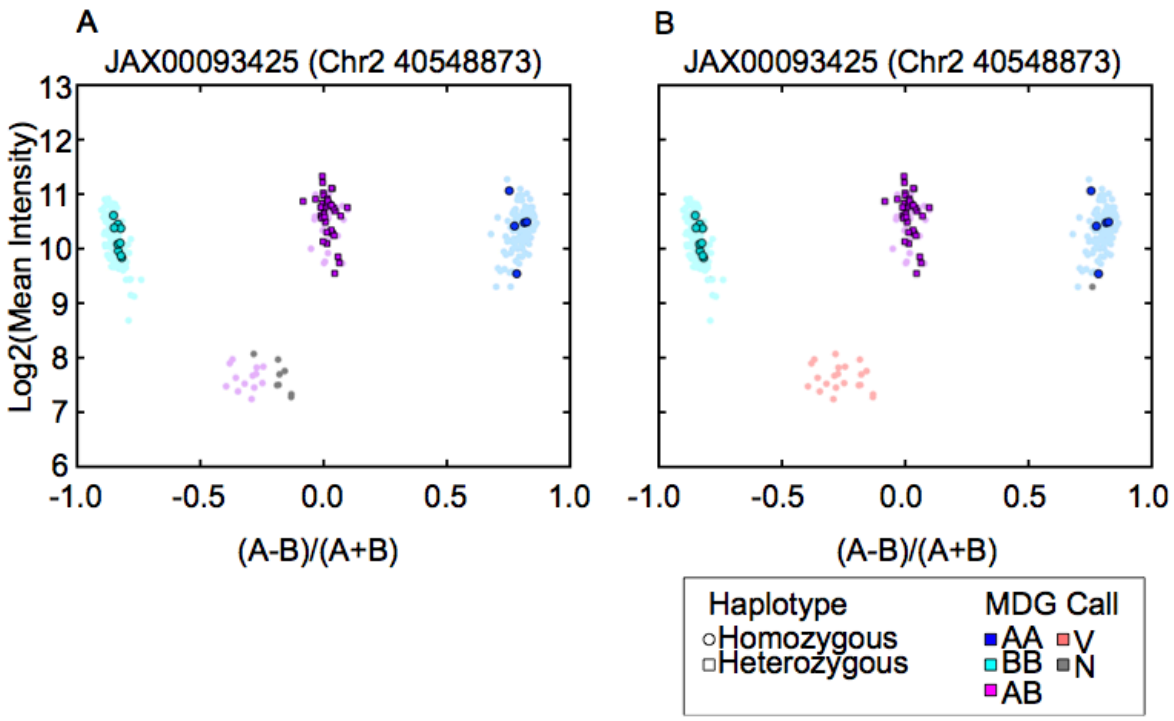


Figure 2.3: VINOs are identified as a cluster of low-intensity samples. Contrast plots of a SNP called by A) `BRLMM-P 2D` and B) `MouseDivGeno`. Probe intensities from 351 samples are shown in MA-transformed space. The sample contrast is the normalized difference between A and B allele intensities [(A-B)/(A+B)]. The y-axis shows the log2 mean of A and B allele intensities. Dark blue: AA call; light blue: BB call; purple: AB call; red: V call; gray: N call. Circles represent strains with a homozygous haplotype in the region of the SNP, while squares represent strains with a heterozygous haplotype. F1 animals with parental alleles of AA and BB are true heterozygotes and are highlighted along with their parental strains. `MouseDivGeno` software is able to identify samples in the low intensity cluster as containing an OTV and assigns a VINO (V) call, whereas `BRLMM-P 2D` assigns several different genotype calls (AB, N) to samples in this cluster.

of problems affecting all hybridization arrays, genotype calling software and studies that use those genotype data for a variety of goals. Our studies of well-characterized inbred strains provided an opportunity for investigating the underlying causes of genotyping errors.

Essentially, a no-call or incorrect genotype call is the result of abnormal hybridization intensity for a sample at a given SNP and may be due to technical or biological causes. Technical errors are generally either very obvious, such as a high no-call rate due to poor DNA quality, or slight enough that they do not affect genotype calling. On the other hand, genotype calling errors that are biological in origin can be attributed to previously uncharacterized variation

in genomic DNA, either in the sequence targeted by a probe set or in the proximal or distal restriction sites used for genome-wide amplification. These variants can reduce hybridization intensity sufficiently to eliminate or reverse the contrast between allelic probes such that an incorrect genotype call (or no-call) is made. We call such variants "off-target variants" (OTVs) to distinguish them from the expected variant targeted by the SNP probe set. We call probe sets that are affected by OTVs "variable intensity oligonucleotides" (VINOs) due to the dynamic effect of OTVs on hybridization intensity [41].

We hypothesized that OTVs were the primary cause of miscalls and no-calls. Hyuna Yang developed a novel genotype calling algorithm that also recognized clusters of samples apart from those with the standard homozygous or heterozygous genotypes (`MouseDivGeno`, [40, 41]). Probes with such clusters are considered putative VINOs, and the samples in those clusters are given a genotype call of "V" (Figure 2.3). To confirm that VINOs do indeed represent previously unidentified genetic variation, we selected 15 SNP probes with VINO calls. For each probe, I selected at least four mouse strains of each genotype (homozygous for allele A, homozygous for B or VINO) for targeted sequencing. Strains for resequencing were selected to maximally sample across subspecies and strain type (classical or wild-derived). I designed sequencing primers approximately 200 bp proximal and distal to each probe using PrimerQuest (Integrated DNA Technologies). I amplified probe regions by PCR and submitted them for automated Sanger sequencing at UNC. I aligned the resulting sequences using `Sequencher 4.9` (Gene Codes). Supplementary Table 4 of [40] lists all probes, strains and primer sequences used. I confirmed that all homozygous SNP genotype calls were concordant with the sequencing data. In addition, in 14 out of 15 probes the VINO calls were associated with the presence of one or more additional variants near the target SNP. The final case was explained by polymorphisms outside of the sequenced region that altered the cut sites for the enzymes used for genome-wide amplification.

We followed up on this work with a more thorough characterization of the effects of OTVs on hybridization intensities, and a formal description of the `MouseDivGeno` soft-

ware [41]. We first hybridized 351 mouse DNA samples on the MDA. Those data are now public (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml), and include classical inbred strains, wild-derived strains, consomic strains, recombinant inbred strains, samples from early generations of the CC, F1 hybrids and wild mice – among the largest mouse genotype datasets available. Among the 143 inbred strains in that sample (116 classical and 27 wild-derived), we observed a significant increase in both heterozygous calls and no-calls as a function of genetic distance from the reference genome (Figure 2.4). All of those strains were expected to be fully homozygous based on previous studies (for at least 99% of their genomes), therefore we assumed that most of the heterozygous calls were errors (miscalls). We called genotypes for our sample set using three different algorithms: `BRLMM-P 2D` [63], `Alchemy` [64] and `MouseDivGeno`. We found that genotype calls for the set of 351 samples were highly concordant in homozygous and heterozygous classes (97.4 - 97.8% agreement). The majority of discordant genotypes were due to homozygous calls using one of the methods that were called heterozygous using another method. Conflicts with opposite homozygous genotypes were very rare (less than 0.05% in all comparisons). The overall rate of AB genotypes was slightly lower for `MouseDivGeno` (10.26%) compared to `Alchemy` (11.45%) and `BRLMM-P 2D` (11.62%). Of the VINO calls from `MouseDivGeno`, 9.76% and 46.04% were called AB by `Alchemy` and `BRLMM-P 2D`, respectively, while 65.32% and 34.04% were called as N.

Of the 18 strains resequenced by the Sanger Institute [56, 59], 15 are *M. musculus* inbred strains that were genotyped with the MDA. I obtained and filtered SNPs and small insertions/deletions (indels) for those strains at autosomal typed loci (Appendix A). I re-annotated all MDA probes by aligning them to the latest version of the mouse genome (Build 37) using BWA [65]. Probes on the MDA were 25 bp long, and the target SNP was typically located in the center of the probe. For each probe, I identified the number, type and position of OTVs, as well as the presence of OTVs in either proximal or distal restriction sites. I used dbSNP and Ensembl to link each probe to functional classifications in public databases. I also noted whether each probe was in a region of low or missing sequence coverage for any of the Sanger

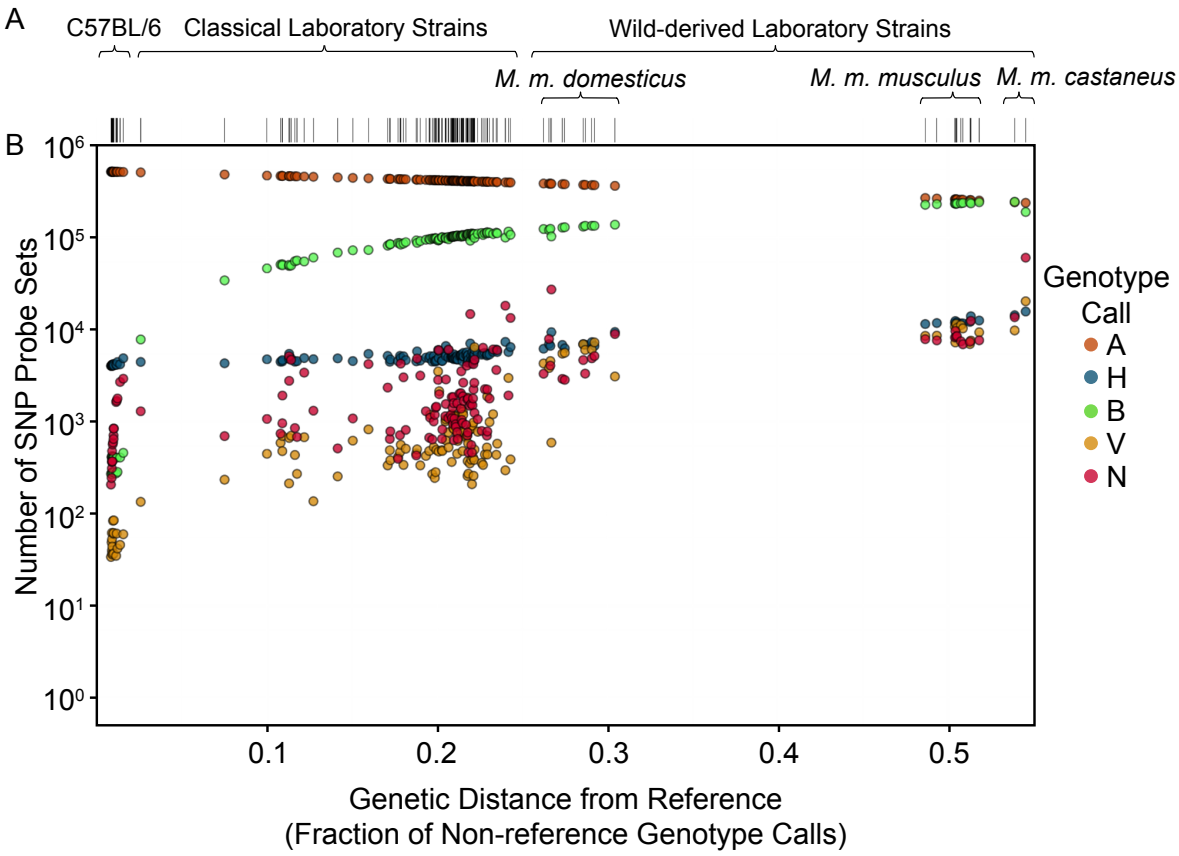


Figure 2.4: Non-homozygous VINO call rates increase with divergence from the reference genome. A) Genetic distance from the mouse reference genome for 143 laboratory inbred strains. Each strain is shown as a vertical tick mark. Strains are grouped according to their origin are arranged left-to-right in increasing order of genetic distance from the reference. Genetic distance is computed as the fraction of non-reference (non-A allele) genotype calls. B) VINO calls for each strain. For each strain, the number of SNP probe sets assigned each of the five possible calls (A, B, H, V or N) are shown as five points of different colors that sum to 526,363 SNP probe sets.

strains.

As expected due to the inbred status of the strains overlapping in the Sanger and MDA data sets, there were no heterozygous calls in the filtered Sanger genotypes. The genotypes for those samples had heterozygous call rates between 1-2%; the homozygous calls were highly concordant between the two data sets (99.8%). `MouseDivGeno` made 35,604 VINO calls (0.48% of total calls), a proportion similar to the one observed in the larger set of 351 samples. Among VINOs, 81.4% correspond to an AA or BB homozygous genotype calls in the Sanger data. Because Sanger SNPs were identified by alignment to the reference sequence, regions

that could not be aligned were inaccessible to SNP discovery and thus not comparable with array genotypes. The size of the inaccessible fraction of the genome increased with a strain's divergence from the reference. I observed an enrichment of VINO calls in inaccessible regions of the Sanger data (2,221 VINO calls compared to an expectation of 54) [56], in probes with a deleted target base (24 vs. 2 expected) and unaligned or non-uniquely aligned probes (4,361 vs. 82 expected).

I examined the correlation between hybridization intensity and OTV position relative to the target SNP for the probes that had OTVs in at least one of the strains (Figure 2.5). I found that OTVs located within the first 3 bp of either the 5' or 3' end of a target sequence (edge OTVs) had relatively minor effect on hybridization intensity. In contrast, OTVs within the central region of the probe (central OTVs) had pronounced effect on hybridization intensity, with mean intensity differing by more than one standard deviation from that of probes having no OTVs. I also found that OTVs that disrupted a restriction fragment site and increased the size of the minimum fragment length to greater than 1500 bp significantly reduced hybridization intensities. I predicted from these results that `MouseDivGeno` was undercalling VINOs by at least 1/3, since VINOs could not be recognized when the OTV was located in 6 of the 24 off-target positions. I determined the false-negative and false-positive rates for VINO calling by comparing predicted VINOs with the Sanger genotypes. Using the Sanger data as the "truth" was problematic due to miscalled or uncalled SNPs in that data set as well as known problems with the mouse genome assembly [66], but it was the best available metric. The measured false-negative rate for sequences with central OTVs was 55%. In most cases, false negatives were due to samples failing to meet the stringent requirements for VINO calling that were used to minimize the false-positive rate. The false-positive rate was 19.8%. I examined the performance of `Alchemy` and `BRLMM-P` and found a more than 30-fold increase in no-call rates for unexplained VINOs.

An additional complication in calling VINOs in wild mice, and in inbred mice with known regions of residual heterozygosity, was heterozygous OTVs. By definition, heterozygous

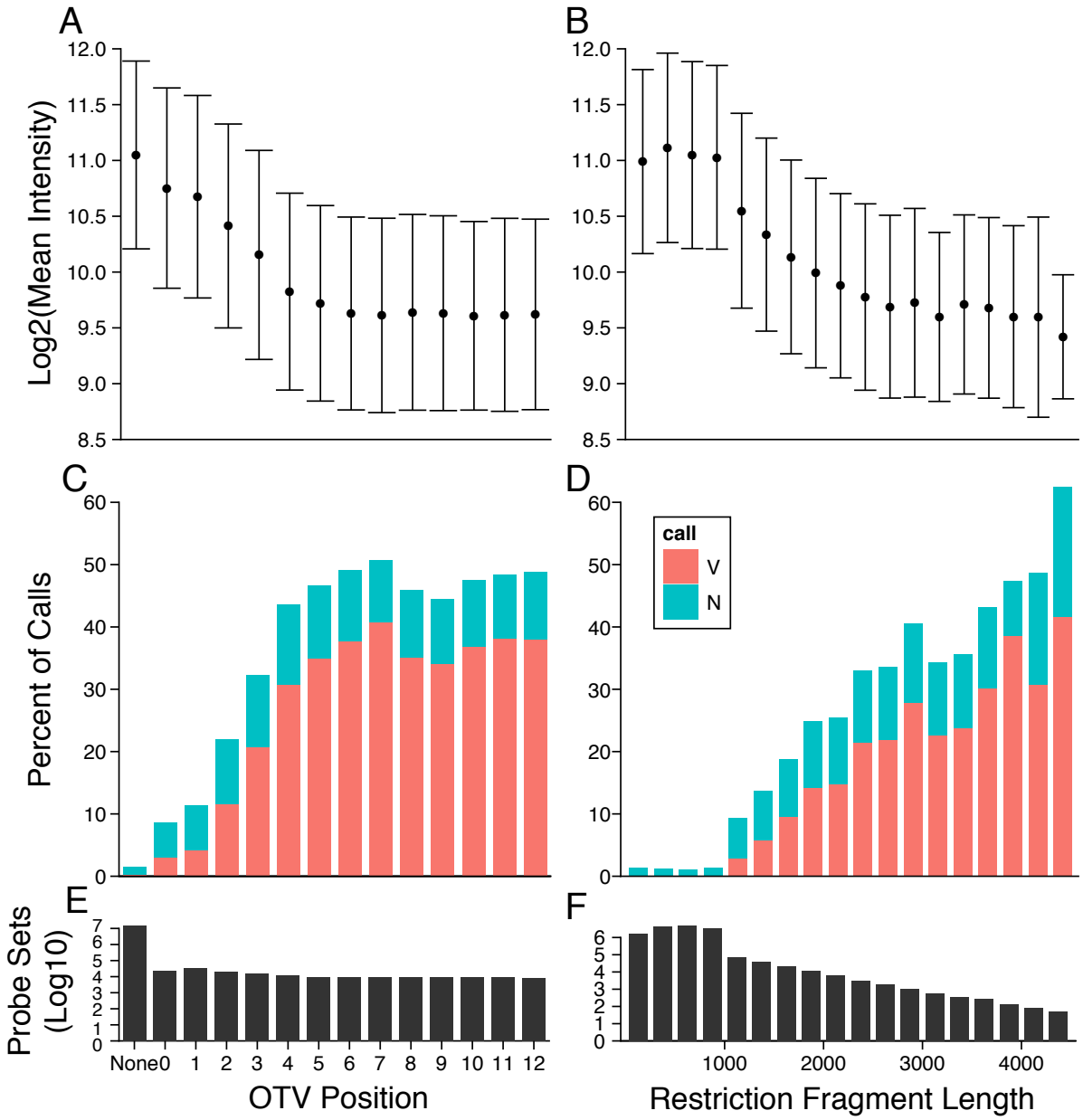


Figure 2.5: OTV position in the probe and RFLP have significant effects on hybridization intensity and VINO detection. Left panels: probe sets are grouped by the distance from the OTV to the nearest edge of the probe sequence for each possible OTV position (either none or between 0-12). Right panels: probe sets having no evidence of an OTV within the probe sequence are grouped by the size of their smallest restriction fragment (*NspI* or *StyI*) in bins of 250 bp. Top panels show the mean intensity across each subset using the four probes for the best-hybridizing allele in each probe set for A) OTV position in the probe and B) minimum restriction fragment length. Middle panels show the number of VINO and N calls (as a percentage of all genotype calls) for probe sets grouped by C) OTV position in the probe and D) minimum restriction fragment length. Bottom panels show the number of probes in each bin for E) OTV position in the probe and F) minimum restriction fragment length.

OTVs only alter one allele. Therefore, heterozygous genotypes with a nearby heterozygous OTV appeared as homozygous for the allele lacking the OTV. We called those "cryptic VINOs" (Figure 2.6). F1 hybrid mice were used to determine the extent of miscalls due to cryptic VINOs since their phase (i.e., parental origin) of haplotypes is known. We used a (C57BL/6JxCAST/EiJ)F1 with the expectation that all OTVs would be present only in the CAST/EiJ sequence. We found that 62% of SNPs with OTVs in heterozygosity were called as homozygous, leading to a low concordance rate (83.35%) between the genotypes predicted from the parental strains and the actual genotype calls for the F1 hybrid. Cryptic VINOs represent a substantial source of genotyping error, particularly since they may only be recognized if the parental genotypes are known (and heterozygous parent genotypes will also be affected by cryptic VINOs).

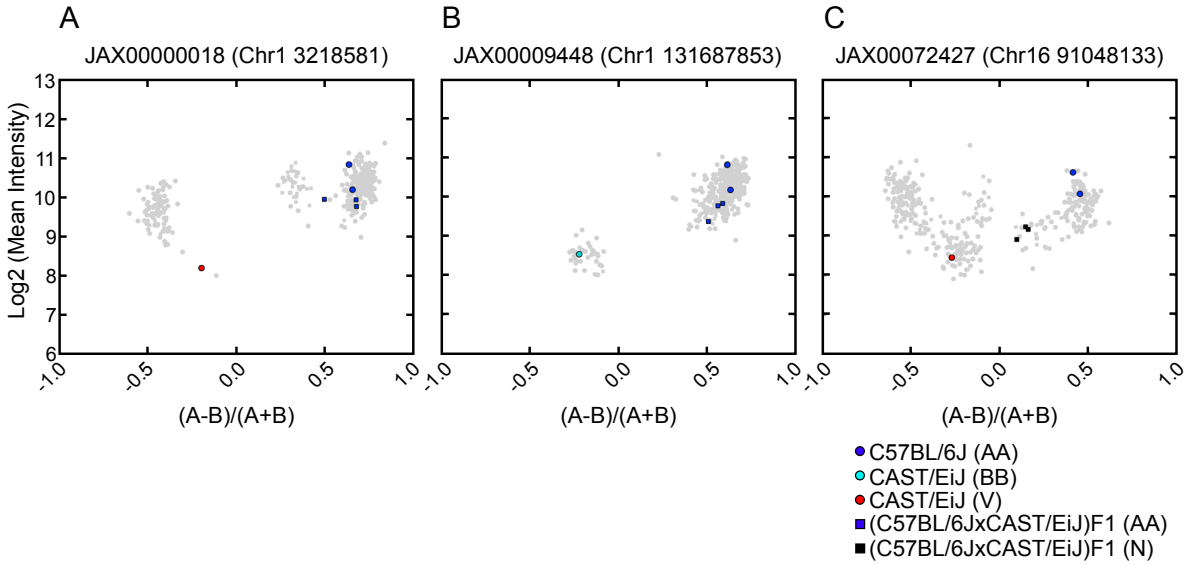


Figure 2.6: Detected and undetected VINOs in homozygosity may lead to inaccurate genotyping in heterozygosity. Circles represent parental strains: C57BL/6J (dark blue), which have the AA allele; CAST/EiJ, which has the BB genotype at its target position and also an OTV within the probe and is called either BB (light blue) or V (red) by `MouseDivGeno`; squares: (C57BL/6JxCAST/EiJ)F1 samples, which have an OTV in heterozygosity and are called AA (dark blue) or N (black) by `MouseDivGeno`. A) `MouseDivGeno` calls CAST/EiJ as V; the F1 samples are called AA due to stronger hybridization intensity for the AA allele and thus the OTV goes unrecognized. B) `MouseDivGeno` calls CAST/EiJ as BB due to the absence of a true BB cluster; the F1 samples are again called AA. C) `MouseDivGeno` calls CAST/EiJ as V but calls the F1 samples as N due to poor discrimination between genotype clusters.

Distances between consecutive SNPs are expected to follow a geometric distribution (Figure 2.7), with a significant proportion in the 0-12 bp range in species with high levels of variation and large populations size such as the house mouse. In a significant fraction of probes with OTVs, we were able to detect the reduction in hybridization intensity and discriminate the samples harboring previously undetected variation from those that do not. VINOs are biased in favor of more divergent samples in reverse proportion to the degree to which the genetic variants in a given sample were known and represented on the array at the time of design. Thus VINOs could be used to counteract SNP selection bias (discussed further below).

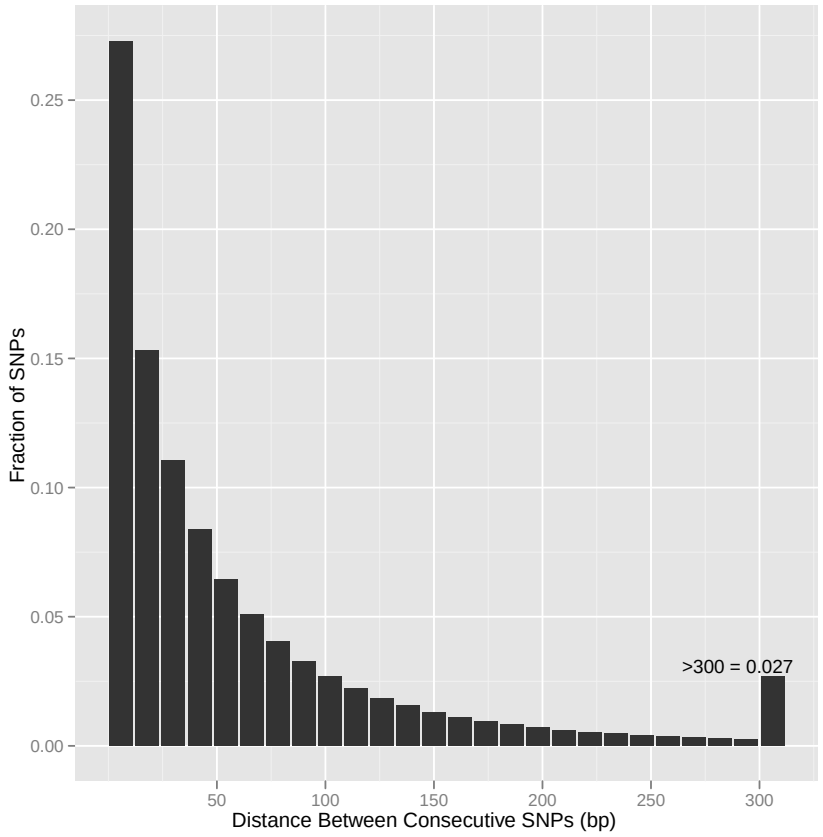


Figure 2.7: The distance between consecutive SNPs follows a geometric distribution. Histogram of distance between consecutive SNPs in 14 Sanger strains using a bin size of 12 bp. Distances greater than 300 bp are combined in the right-most bin.

The method for identifying VINOs is generalizable, and we expect that new genotyping algorithms will take the next logical step of recognizing arbitrary numbers of clusters. We

tested `MouseDivGeno` on a randomly chosen subset of human HapMap data [67]. In 70% of cases, `MouseDivGeno` either correctly called a VINO or the correct homozygous allele of the target variant. The 30% miscalls were all due to cryptic VINOs. We identified a 2:1 bias of VINOs in human YRI (Yoruban African) samples compared the other three HapMap populations. That was consistent with the greater number of genetic variants in African populations that were unknown at the time of the design of the human SNP array.

2.6 Diagnostic SNPs

An important factor in the study of natural populations is the long-distance relatedness (shared ancestry) of individuals. At each SNP, two individuals may share the same allele or have different alleles. Shared alleles may be due to shared ancestry (identity by descent, IBD), or they may have occurred by recurring mutation (homoplasy). Alleles that are exclusive to a single taxa, or that only appear at a low level in other taxa due to homoplasy, are useful for determining the ancestral origin of previously uncharacterized individuals. We call such markers diagnostic alleles, although in studies of human ancestry they are sometimes referred to as ancestry-informative markers.

We used genotypes of 36 wild-caught mice to determine the ability of each MDA SNP or VINO to discriminate between subspecies [40]. At each marker, we examined the allele frequencies within each subspecies. Alleles found in only one subspecies were considered diagnostic. These included fully informative alleles, in which subspecies are fixed for different alleles, and partially informative alleles, in which an allele was restricted to one subspecies but not fixed. We identified 251,676 SNPs and 96,188 VINOs with diagnostic alleles distributed across every chromosome. We found substantial differences between the number of SNPs and VINOs with diagnostic alleles for each the three subspecies detected. For example, 55% of all informative SNPs carried diagnostic alleles for *M. m. domesticus*, whereas only 27% and 18% carry diagnostic alleles for *M. m. musculus* and *M. m. castaneus*, respectively. This situation was reversed among VINOs, where 17%, 24% and 59% of diagnostic alleles identified *M. m.*

domesticus, *M. m. musculus* and *M. m. castaneus*, respectively. Those differences reflected the two opposing biases discussed above. On one hand, the selection criteria for inclusion of SNPs in the MDA led to the over-representation of SNPs with *M. m. domesticus* diagnostic alleles and under-representation of *M. m. castaneus* SNPs [60]. On the other hand, the deeper knowledge of the genetic variation present in the *M. m. domesticus* subspecies allowed screening of candidate SNP probes with internal polymorphisms that could create VINOs, whereas the limited knowledge of the genetic variation present in the *M. m. castaneus* subspecies in particular resulted in an excess of *M. m. castaneus* diagnostic VINOs. We constructed a phylogenetic tree of the 36 wild-caught samples and confirmed the taxonomic classification of all samples (Figure 2.1).

The method described above allowed for misclassification caused by genotyping error, homoplasy or gene flow in the wild by down-weighting (but still considering diagnostic) alleles that were detected at low frequency ($< 5\%$) in the other subspecies. We are currently using a more robust method of identifying diagnostic alleles based on a Bonferroni-corrected Chi-squared test (2 df) for markers with significantly different frequencies in one subspecies compared to the other two. Diagnostic SNPs appear to be quite robust to sampling differences. I applied the method for discovering diagnostic alleles to a larger sample and found 92% concordance with the set discovered in only 36 wild-caught mice.

2.7 VINOs and diagnostic SNPs mitigate ascertainment bias

Ascertainment bias can result in distorted allele frequencies and inaccurate phylogenies. VINOs contain important phylogenetic information that can correct for the common problem of underestimating branch lengths for highly divergent samples due to missing information (which is typically ignored by phylogeny reconstruction methods). This is dramatically illustrated by phylogenetic analysis of several different species of the *Mus* genus (Figure 2.8). I constructed maximum-likelihood phylogenetic trees using strains derived from *M. musculus*, *M. spretus*, *M. spicilegus*, *M. cypriacus* and *M. macedonicus* (Figure 2.1, see Appendix A for

methods). When only the standard genotypes were used, the discrimination between non-*M. musculus* species was poor (Figure 2.8 A). Furthermore, the length of the *M. m. domesticus* branch was grossly overestimated while non-*M. m. domesticus* branches were underestimated due to a high rate of missing information in those samples. The opposite result was observed when only VINOs were used to construct the tree by converting all genotypes to binary for the presence or absence of a VINO (Figure 2.8 B). When genotypes and VINOs were combined, discrimination between taxa increased and a representation more similar to morphology-based phylogenies emerged (Figure 2.8 C).

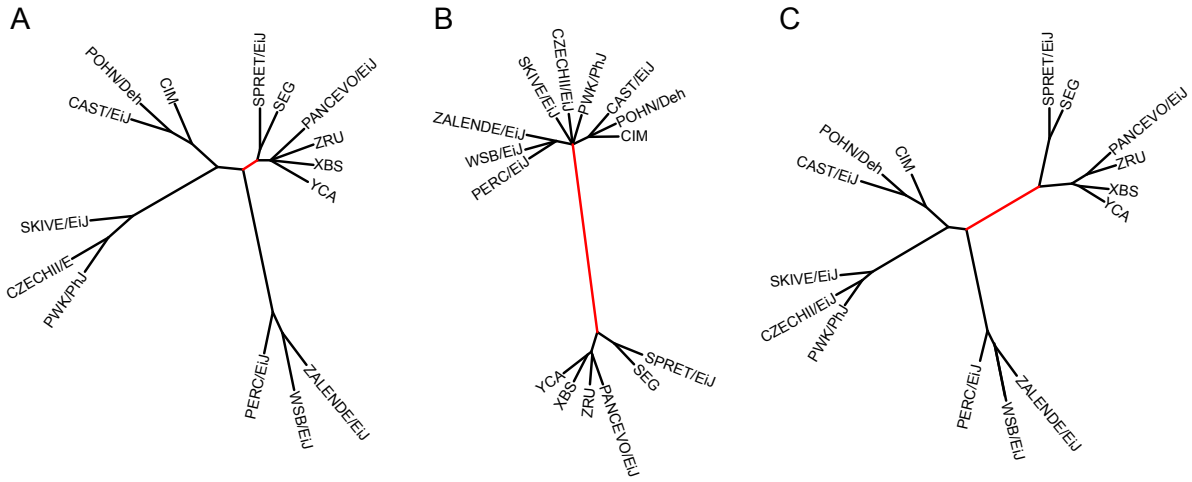


Figure 2.8: VINOs improve the topology of phylogenetic trees. Phylogenetic trees created using A) SNP genotypes only, B) VINOs only and C) both SNP genotypes and VINOs. The branch highlighted in red separates *M. musculus* and non-*M. musculus* strains and is the most significantly improved by the addition of VINOs.

2.8 Subspecific origin of laboratory mice

It has long been known that laboratory mice do not belong to a single taxa but rather represent a mosaic between multiple *M. musculus* subspecies [68, 50, 69]. Some have even suggested that the laboratory mouse be given it own taxonomic designation, *Mus gemischus* (gemisch is a Yiddish word meaning "mixture") [70, 42]. However, the quantity and distribution of the contribution from each subspecies has been fiercely debated. A popular model was that the ancestry of the laboratory mouse was a roughly equal mixture of *M. m. domes-*

ticus and *M. m. molossinus* [46]. That view had a pervasive influence in the planning and interpretation of SNP discovery efforts.

We and others have recently presented results on the subspecific origin of laboratory mice using the newly available genotyping [55, 71, 40] and sequencing [56] platforms. The sets of strains used in those studies were different but highly overlapping. In each study, the authors chose one or more samples to serve as a reference for each *M. musculus* subspecies. They then examined the local phylogenetic relationships among strains (called strain distribution patterns, SDPs) in small regions spanning the genome. Within each region, they attempted to assign a subspecific origin to each group of related strains based on the reference sample(s) that clustered with the group. Remarkably, the local concordance between SDPs was high across all studies despite the use of distinct genotype data sets that differed in density by several orders of magnitude. However, in spite of the local agreement between phylogenetic relationships, the studies drew opposite conclusions about the ancestral origin of the laboratory mouse genome. Frazer *et al.* (2007) concluded that the ratio of *M. m. domesticus* to non-*domesticus* (or unknown) ancestry in the classical strains was about 2:1, a finding that supported the traditional mosaic model. Their conclusions were based on the assumption that the four wild-derived strains were "pure" representatives of their respective subspecies. In contrast, Yang *et al.* (2007) determined that classical strains are primarily of *M. m. domesticus* origin (92%), with only a minor contribution from *M. m. musculus* and *M. m. castaneus* (6–7 and 1–2%, respectively). Their method was based on the use of diagnostic markers, and required excluding regions of the genome in which diagnostic markers were infrequent.

With the availability of the MDA, we were able to genotype wild-caught mice from the native ranges of each of the three subspecies. We used diagnostic SNPs and VINOs derived from those samples to infer the subspecific origin of every region of the genome of each of the 162 classical and wild-derived laboratory strains. An HMM was used to identify the boundaries and subspecific origin based on the cumulative scores within these regions [40]. Our results showed that the genome of classical inbred strains is predominantly derived from

M. m. domesticus (mean of 94.3% $\pm$ 2.0% per genome), with variable contribution from *M. m. musculus* (5.4% $\pm$ 1.9%) and a small contribution from *M. m. castaneus* (0.3% $\pm$ 0.1%). The contribution from subspecies other than *M. m. domesticus* was not distributed randomly across the genome or among strains, but rather lay mostly in overlapping regions of strains with some shared history. Notably, the *M. m. castaneus* and *M. m. musculus* contributions were not independent from each other, with the former frequently nested within or contiguous with the latter. This association suggested a *M. m. molossinus* origin of the *M. m. musculus* contribution to the classical inbred strains. We tested this hypothesis by comparing the *M. m. musculus* regions found in classical inbred strains to wild-caught *M. m. musculus* mice from Europe or Asia. Over 90% of the *M. m. musculus* haplotypes found in classical inbred strains clustered with Asian wild-caught mice.

Introgression is the movement of variants from one population into the gene pool of another population by the repeated backcrossing of a hybrid to one of its parent populations. Because *M. musculus* subspecies are not generally sympatric, introgression typically exists on a small scale and is difficult to observe, even with high-density genotype data. However, exceptions occur in places with a high rate of mixing between individuals of divergent genetic backgrounds [72]. Those regions are known as hybrid zones, and they may be natural or man-made. The derivation of new wild-derived strains has in large part been driven by a few fields of study, such as hybrid zone biology. This, along with the findings of [71] suggest that introgression may be widespread in wild-derived strains.

We extended the analysis of subspecific origin to test whether wild-derived strains were pure representatives of their expected subspecies. We computed the frequency of diagnostic alleles in non-overlapping 1 Mb intervals for each wild-derived strain. The majority of the genome of the 62 wild-derived laboratory strains originated from the expected subspecies or combination of subspecies (Figure 2.9). However, only 9 strains had a genome derived entirely from a single subspecies, while 18 had contributions from two subspecies and 35 had contribution from all three subspecies. The prevalence and extent of multi-subspecific origin

was a defining characteristic of wild-derived laboratory strains as a group. Our set of wild-derived strains included ten strains derived from natural intersubspecific hybrids, all of which had, unexpectedly, contributions from all three subspecies. The remarkable discordance in subspecific origin in several strains based on phylogeny (Figure 2.10) provides further evidence for intersubspecific introgression. Interestingly, we identified several shared patterns of subspecific origin between classical inbred strains and some wild-derived strains, which suggested that some of the intersubspecific introgressions in the latter group involved cross breeding with classical strains.

Diagnostic markers are also an important tool for identifying the ancestry of previously unstudied "new-world" populations (i.e., populations outside the historical ranges of the three subspecies). We obtained wild-caught mice from Southeast Farallon Island (USA) and Floreana Island (Galapagos archipelago, Ecuador). I used diagnostic markers to determine that both of these populations were primarily *M. m. domesticus*. I analyzed the diagnostic markers using `ChromoPainter` [73] and identifed population structure and shared ancestry in *M. m. domesticus* mice. There were generally two genetically divergent populations: northern Europe and the Mediterranean basin. That reflected the general consensus that the mouse colonized Europe from south to north, likely with partial isolation of the two populations due to geographic boundaries. The Farallon mice appear to be a mosaic of the northern and southern populations. I constructed separate phylogenies of the mitochondrial and Chr Ys of the samples and found that the Farallon mice clustered with mice from the northern UK in the former, and the Mediterranean basin in the later. Together, this evidence suggests multiple colonizations of the Farallon Islands by house mice of different origins.

2.9 Haplotype and sequence diversity

A useful unit for the analysis of genome organization is a haplotype block, a contiguous interval in which the number of unique sequences (haplotypes) is much smaller than the total number of sequences due to a high degree of genetic similarity (approaching identity)

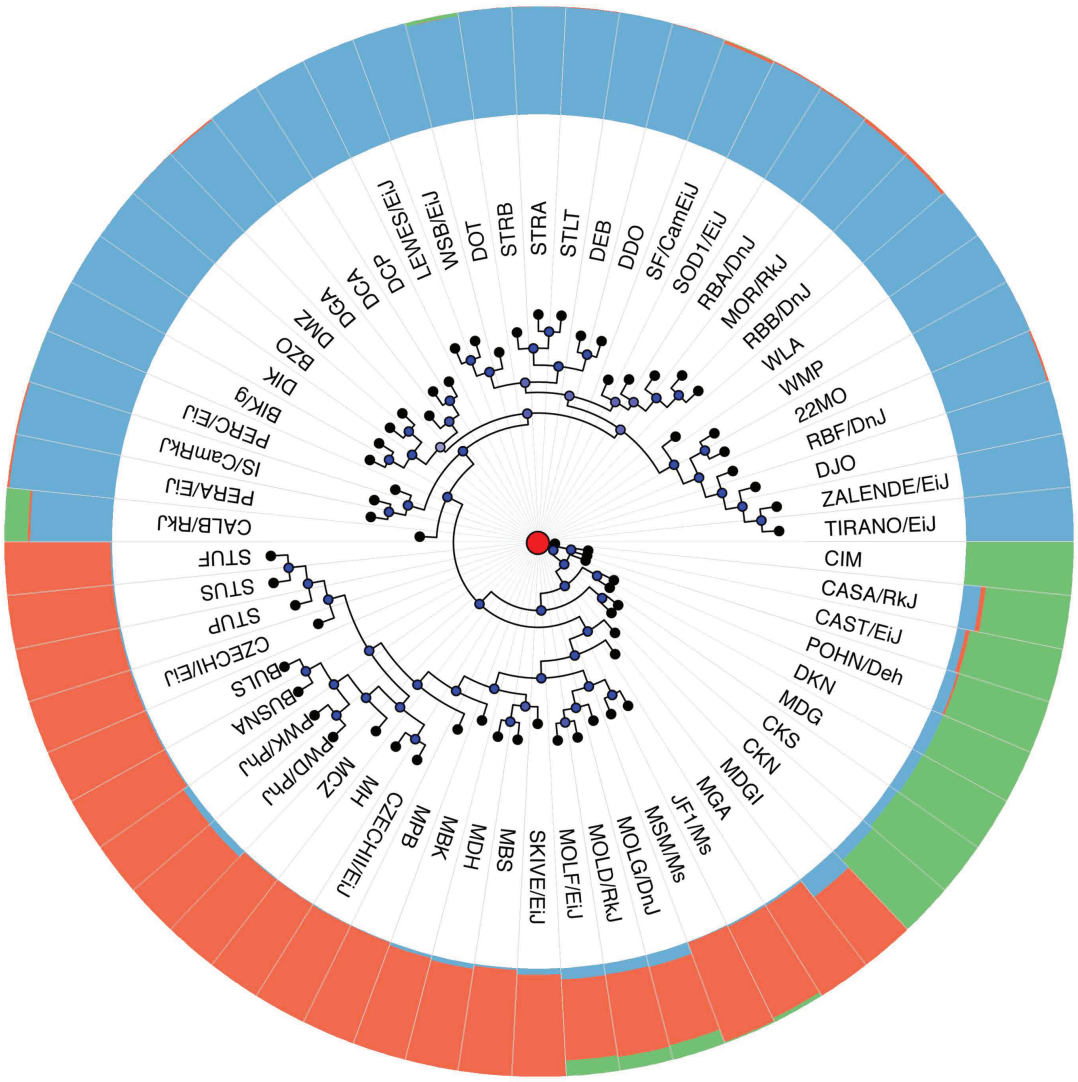


Figure 2.9: The phylogeny of wild-derived strains. Neighbor-joining phylogeny of 62 wild-derived strains based on SNP and VINO genotypes. Node colors represent bootstrap values. The outer ring of colors shows the fraction of the genome of each strain that is derived from *M. m. domesticus* (blue), *M. m. musculus* (red), and *M. m. castaneus* (green). This figure appeared on the cover of the Jaunary 2013 issue of *Mammalian Genome*.

within subsets of strains. A natural criterion to define haplotype blocks in classical strains is to identify regions of shared ancestry among multiple strains which have not recombined (compatible intervals) [74, 40] using the 4-gamete rule [75]. We used the 4-gamete rule to identify 43,285 haplotype blocks with a median size of 71 kb in 100 classical strains. The majority of blocks contained between four and six haplotypes, and there were fewer than ten haplotypes across 97% of the genome. Those findings confirmed the small size of the classical strain founder population. The larger numbers of haplotypes in the remaining 3% of the genome were due to a combination of new mutations in the past century and contributions from outside of the founder population. Blocks with large numbers of different haplotypes should be further investigated to understand their origins.

The relative lack of genetic variation in classical strains limits their utility in at least two respects. First, it constrains the phenotypic variation that exists in classical strains. Second, use of classical strains is inappropriate to study evolutionary processes since they may be invariant for many of the genes involved in speciation [76]. The extent of additional variation present in natural populations of *M. musculus* is hinted at by limited studies in wild mice [77] and by the recent whole-genome sequencing of three wild-derived strains [56], but it is not known for certain. To examine sequence variation at the genome scale, I computed the nucleotide diversity in classical, wild-derived, and wild-caught mice. I used the method of [78] to compute the average pairwise genetic distance between individuals within a population π. Overall, I found greater diversity in wild-derived and wild mice than in classical strains (π = 0.298, 0.282, and 0.203, respectively). The contrast is even more striking, however, when comparing diversity between regions with different subspecific origin (Figure 2.10). In classical strains, intervals derived from *M. m. domesticus* founders had 6 and 17 times greater diversity than intervals derived from the minority Asian fancy mouse founders (*M. m. musculus* and *M. m. castaneus*, respectively, Figure 2.10 A), whereas in wild-derived lines and wild mice (Figures 2.10 B,C), variation is similar among regions of different ancestry.

There is a significant risk of introducing bias into a study when local differences in hap-

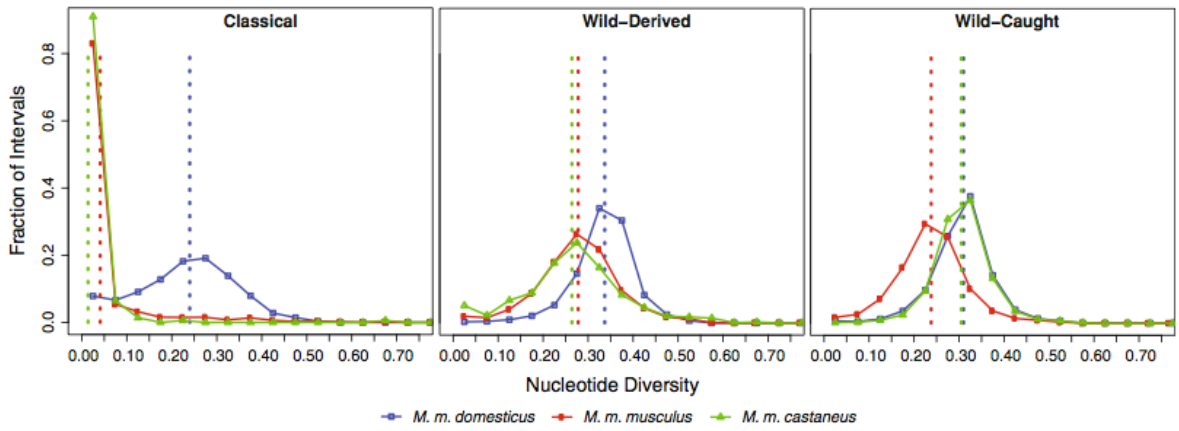


Figure 2.10: Nucleotide diversity is greater in wild mice than classical strains. We divided the genome into 16,331 intervals with no historical evidence of recombination in classical strains and measured nucleotide diversity (π) at diagnostic SNPs in each interval for classical strains, wild-derived strains, and wild-caught mice. The x axis shows π for each subspecies in bins of 0.01. The y axis shows the fraction of intervals with the given subspecific origin that is in each bin. Vertical dotted lines show mean values of π for each subspecies. Color indicates subspecific origin. Blue: *M. m. domesticus*; red: *M. m. musculus*; green: *M. m. castaneus*.

lotype diversity are not accounted for [79, 80]. Haplotype structure also has important implications for the ability to conduct genetic mapping because it can significantly affect the level and rate of decay of linkage disequilibrium (LD) [39]. Gametic disequilibrium (GD), which is also known as long-range LD, is problematic because it can introduce false genotype–phenotype associations [81, 82]. An analysis of LD decay in a panel of 88 classical strains revealed widespread GD [39], suggesting caution when interpreting the results of mapping experiments in those strains. The effect of population structure can be reduced by using a genetic reference population (such as the CC).

2.10 The MegaMUGA genotyping array

UNC is the primary breeding and distribution site for Collaborative Cross (CC) lines. The speed at which lines can be inbred is a limiting factor in the number of lines that can be made available to researchers. To increase the speed of inbreeding, the McMillan and Pardo-Manuel de Villena labs developed the Mouse Universal Genotyping Array (MUGA) [83, 39]. The array was designed on the Illumina platform, primarily due to lower cost and easier sample

preparation protocols compared to the MDA. The array contained 7,851 SNP markers that were distributed throughout the mouse genome. The markers were chosen to be maximally informative and maximally independent for the eight founder strains of the CC. Informativeness was determined by minor-allele frequency, and independence was determined by local pairwise linkage disequilibrium. The design criteria resulted in sets of three contiguous SNPs that together could uniquely identify each of the eight founders. The design was also optimal for the detection of heterozygous regions in the CC.

Soon after MUGA was developed, rapidly advancing technology made it possible to develop a new array with 10x greater density than MUGA for the same per-sample cost. This new array has the capacity to enable higher-resolution haplotype identification in the CC, and in the companion Diversity Outbred (DO) population (discussed later), and also to include markers that were informative in natural populations. Using MDA genotypes from several hundred wild mice of all three subspecies, I selected 13,238 markers for the array that I identified as segregating at high minor allele frequency in one of the subspecies while also being discriminatory for that subspecies (i.e., partially informative diagnostic markers). The majority of those markers were selected to be diagnostic in *M. m. domesticus* to support several projects studying natural populations of that subspecies. In addition I selected 1,007 SNPs that I identified as segregating in species of *Mus* closely related to *M. musculus*: *M. spretus*, *M. spicilegus*, *M. macedonicus* and *M. cypriacus*. Those diagnostic SNPs will facilitate evolutionary studies with resolution and accuracy comparable to the MDA. MegaMUGA also includes a larger number of mitochondrial markers than the MDA (42), and all of the Chr Y markers from the MDA that performed well (33). Those markers have been widely used previously for phylogenetic analyses in the mouse, and so a large database of genotypes already exist that may be compared with new samples genotyped on MegaMUGA. The remaining markers on MegaMUGA were selected to support several important mouse models, including knock-outs on C57BL/6 backgrounds and other genetically engineered mice.

2.11 Studies using the MegaMUGA array

I conducted several experiments that demonstrate the versatility of the MegaMUGA array. I first assessed the genotyping error rate of MegaMUGA and found it to be remarkably low (0.04%) when comparing biological replicates of 11 inbred lines. I also compared those samples to the Sanger genotypes and found an inconsistency rate of 2.1%. I hypothesized that most inconsistencies were systematic, either due to incorrect Sanger genotypes or poorly performing MegaMUGA probes. When I eliminated the 4,052 poorly-performing markers (5.2% of markers), the rate of inconsistency fell to 0.005%. I developed a set of QC metrics for determining the quality of array data (described in Appendix A). I implemented these methods in an R package, `megamugaQC`, that will be released along with the publication on the MegaMUGA array.

I assigned subspecific origin to the autosomes and Chr Xs of all samples. First, I identified 36,822 diagnostic SNPs using the methods described above for MDA. I developed a Hidden Markov Model (HMM) that had seven states corresponding to pure *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*, the three pairwise mixtures, and the indeterminate state in which all three subspecies are equally likely. I assigned initial values based on the predicted subspecies of each sample (0.94 for the state corresponding to the predicted subspecies and 0.01 for each other state). I used a transition probability matrix in which the diagonal values were 0.94 and all other values were 0.01. I estimated the mean and covariance matrix parameters of the multivariate normal distribution by averaging the diagnostic values for each subspecies in a 5 Mb sliding window. The background mean and variance were based on the number of misclassifications for each diagnostic allele. For each sample, chromosome and subspecies in each 5 Mb sliding window, I summed the diagnostic values for each matching allele and divided by the total number of diagnostic alleles to derive the three-variable observation matrix. I then used the HMM to assign the subspecific state to each window based on this matrix. MegaMUGA did a reasonably good job of recapitulating previous results based

on MDA data. All wild-caught animals were classified as completely pure representatives of their predicted subspecies. For inbred strains, most small introgressions ($<$ 0.5 Mb) were not identified. Large introgressions were identified as a deviation from the predicted subspecies, but the subspecific origin was often assigned incorrectly (usually as a mixture of two subspecies). Those results were expected due to the relatively low density of *M. m. musculus* and *M. m. castaneus* diagnostic alleles.

While the genotypes for Chr Y and mitochondrial markers were robust, the relatively small number of markers made phylogenetic analysis problematic. I examined the intensity data and found a large number of additional sample clusters (i.e., VINOs) that were important to producing correct phylogenies. Since the VINO identification method has not yet been extended to MegaMUGA, I performed supervised clustering of 44 mitochondrial and 38 Chr Y probes that were both unique and had multiple distinct clusters. I randomly assigned non-allelic genotypes to the clusters beyond the two expected alleles. I created parsimony trees using the `DNAPARS` program in the `Phylip` package [84]. The Chr Y phylogeny yielded a single best tree, while the mitochondrial phylogeny yielded multiple best trees. I analyzed each SNP independently using a test for leaf node proximity [85] and found that 26 of the mitochondrial markers (59%) showed evidence of homoplasy; only 9 Chr Y markers were homoplasic (24%). This high level of homoplasy in the mitochondrial tree was expected because the majority of the mitochondrial SNPs on MegaMUGA are located in the D-loop region, which has an extremely high mutation rate [86].

Significant increases or decreases in intensity across consecutive probes are indicative of copy number variation (CNV). MUGA and MegaMUGA have been important for us and others as a tool to study multiple types of CNV. In creating the Chr Y phylogeny, I uncovered intra-specific variation in the pseudo-autosomal region (PAR) of the Chr Y. The PAR is a $\sim 700kb$ region of homology between Chrs X and Y where the two chromosome pairs and undergo recombination during male meiosis. It was previously reported that the region of homology in the CAST/EiJ strain (derived from *M. m. castaneus* mice from Thailand) is 430kb

longer than in other mouse strains. However, intensity data showed that, in *M. m. castaneus* mice from Taiwan, the region of homology extended less than 100 kb beyond the ancestral boundary. Furthermore, recombination data from the CC showed that 90% of recombinations involving CAST/EiJ PARs were within the ancestral PAR, 10% were within the 100kb proximal to the ancestral PAR, and no recombinations occurred in the other 330kb of the CAST/EiJ region of homology. This finding suggests that the CAST/EiJ PAR evolved through two separate events: first a duplication of $\sim 100kb$ of Chr X sequence in the ancestral *M. m. castaneus* lineage, followed by a second duplication event exclusive to some subset of southeast Asian mice that happened after they diverged with the Taiwanese population. This lends further support to the finding that *M. m. castaneus* is polytypic [87].

In another study, I surveyed 100 mouse cell lines using the GAP algorithm [88] and found that a large fraction of lines (15%) had evidence of whole-chromosome loss or gain for at least one chromosome.

I also developed a method to distinguish male and female samples based on their X- and Y-chromosome intensity profiles, and simultaneously detect sub-chromosomal CNV. Briefly, I used a supervised method based on predicted sex to identify sex-specific intensity distributions for each marker. I then determined the probability that a sample belongs to each distribution within a moving SNP window, and I identified intervals of consistent copy number prediction. I predicted the baseline chromosome copy number from the relative local copy-number rates. I used this method to predict the sex of approximately 5,000 MegaMUGA arrays in our database. I identified 27 samples with obviously incorrect reported sexes. I also identified 33 females having a single Chr X (XO, Figure 2.11). Interestingly, the frequency of XOs is much higher in the DO population than in other laboratory strains or wild mice. This finding is the subject of ongoing investigation.

Finally, we recently used MegaMUGA to prove the existence of a segregating ~ 250 kb duplication on Chr 12 in the CAST/EiJ inbred line, which was initially predicted from allele-specific analysis of RNA sequencing data (Crowley *et al.* submitted).

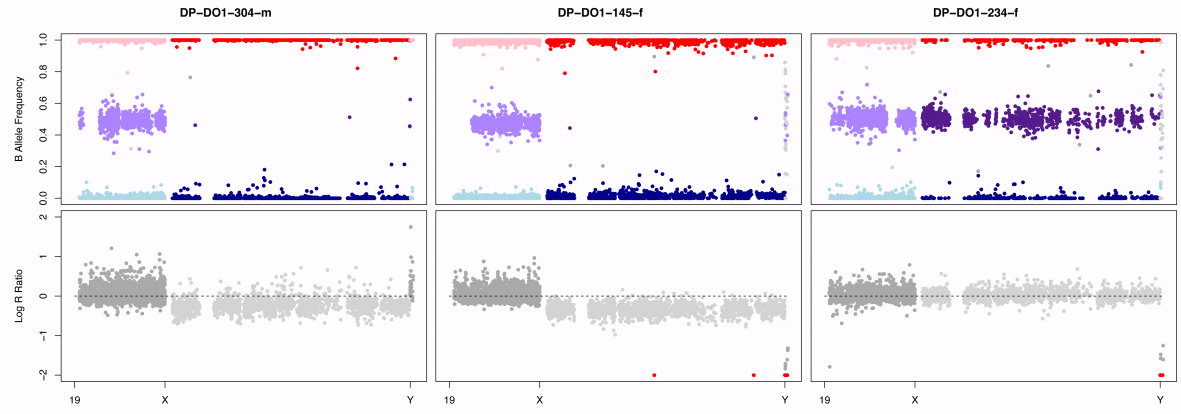


Figure 2.11: MegaMUGA can identify chromosome loss. Intensity profiles of A) a normal male, B) an XO female and C) a normal female. Each panel shows B-allele frequency (BAF, top) and Log-R ratio (LRR, bottom). BAF is a measure of the ratio of the A and B alleles for a SNP; points near 0.0 indicate the A allele, points near 1.0 indicate the B allele, and points near 0.5 indicate a heterozygous genotype. LRR is a measure of the sum intensity of a SNP relative to a reference distribution; values above zero indicate greater intensity than the reference, and values below zero indicate lower intensity than the reference. Values below zero on Chr X are expected for the male, who has only one Chr X, and thus half as much Chr X DNA to hybridize to the array. Similarly, values below zero on the Chr Y are expected for the female. An XO female is detected by values below zero on both Chrs X and Y.

2.12 Future work

Traditional genotyping generally involves preprocessing and normalization procedures where hybridization intensities are converted into a small set of discrete genotypes. We have found that analyzing hybridization intensities directly produces highly repeatable results comparable to those obtained from discrete genotypes. Additionally, intensity analysis captures subtle variations (e.g. CNVs, indels, VINOs) that are not detectable through traditional discrete genotypes. This allows for a multiallelic model of regional variations, where the samples with highly similar intensity profiles within a region are considered to have the same allele. The UNC Computational Genetics group is currently developing a genotyping method for MegaMUGA that will recognize multiallelic variants.

The methods described in this section are currently being applied and extended in several projects to better characterize genetic variation in wild mice, including the study of several previously uncharacterized mouse populations. I have been supervising an undergraduate stu-

dent, Kevin Rocker, on a senior thesis project to identify diagnostic markers in wild mice genotyped on the MegaMUGA array. Kevin will then use that information to conduct a more detailed study of the Farallon Island and Floreana Island mice using the new collection of samples described above. We are also collaborating with Elodie Gazave and Jeremy Searle to identify haplotypes in wild mice. Such a collection of haplotypes will be crucial for genotype phasing, as well as studies of ancestry and population genetics. We are collaborating with François Bonhomme to characterize mice from the central regions (the Middle East and central Asia). Those mice were previously identified as a distinct subspecies, *M. m. gentilulus*; however, preliminary analysis has shown them to harbor diagnostic alleles of all three subspecies. This suggests that the central populations are either hybrids that have resulted from secondary contact, or that they continue to represent the genetic diversity present in ancestral *M. musculus* population prior to speciation. Finally, we are collaborating with Dr. Yung-Hao Ching (Tzu Chi University, Taiwan) to characterize mice wild from the island of Taiwan.

Chapter 3

GENETIC DETERMINANTS OF MEIOTIC DRIVE IN CHROMOSOMAL RACES OF THE HOUSE MOUSE[1]

3.1 The chromosomal races of *M. m. domesticus*

The normal mouse karyotype consists of 40 acrocentric chromosomes (i.e., 2N = 40, 19 pairs of autosomes and two sex chromosomes). In 1869, a European naturalist identified a mouse-like rodent in the Poschiavo valley (Switzerland) with a "tobacco-colored" coat. Believing this a new species, he gave it the taxanomic designation *Mus poschiavinus*. Based on morphological and molecular evaluation, Poschiavo mice belong the *M. m. domesticus* subspecies; however, cytogenetic analysis revealed that the mice have only 26 chromosomes [90]. We now know that the Poschiavo mice are only one of over one hundred karyotypically-abnormal populations of *M. m. domesticus* found throughout western Europe, northern Africa and Turkey (reviewed in [91]). Each population is homozygous for a different set of metacentric chromosomes that have arisen by Rb translocation. In biological classification, these populations constitute different races of the house mouse, commonly referred to as chromosomal races (CRs).

[1]The work described in this chapter was accomplished in collaboration with Jeremy Searle. Dr. Searle helped to provide the vision for the experimental design and was instrumental in coordinating the collection of samples many additional collaborators: François Bonhomme, Nina Bulatova, Pierre Boursot, Janice Britton-Davidian, Ricardo Castiglia, Eva Giagia-Athan-asopoulou, Sofia Gabriel, Silvia Garagna, Sofia Grize, Islam Gündüz, Bettina Harr, Heidi Hauffe, Jerry Herman, Leonas Kontrimavicius, Anna Lindholm, Maria de Luz Mathias, George Mitsainas, Jaroslav Pialek, Priscilla Tucker, Jacint Ventura, Jan Wojcik. A rotation student, Scott Yourstone, developed the algorithm I used for analyzing heterozygosity in wild mice. A manuscript on this work is currently in preparation. In addition, a subset of the data was used in mapping the X-chromosome inactivation controlling element (*Xce*) in Calaway *et. al.* 2013 [89].

A CR is defined as "a group of geographically contiguous or recently separated populations which share the same set of metacentrics and acrocentrics by descent" [92]. As of this writing, we know of 103 distinct CRs (Table 3.1) in 15 countries (Figure 3.1). These include the 97 CRs listed in [91], four CRs reported since that review [93] and two unpublished CRs: CHWE (Anna Lindholm and Sofia Grize) and FNAN (Janice Britton-Davidian). These CRs tend to be very restricted in distribution, often occupying only tens or hundreds of square kilometers. They tend to be found (though not exclusively so) in isolated areas, such as mountain valleys and islands.

CRs interact with each other and with standard-karyotype populations (STs) in varied and complex ways, which has complicated the question of how CRs first arose. Hybrid zones often exist in regions where karyotypically divergent populations overlap. These hybrid zones are characterized by individuals with heterozygous combinations of the metacentric and acrocentric chromosomes found in the source populations. The effect of karyotypic heterozygosity on non-reproductive fitness of the F1 hybrids is negligible [94]; however, the effect on fertility is highly variable and depends on the number of metacentrics [95], the presence of monobrachial homology [96, 97, 98] (a single chromosome arm shared between two different fusions), and the genetic background [99]. In experiments using heterozygotes derived from wild populations, the presence of a small number (1-3) of trivalents at meiosis has little effect on fertility [100, 101, 102, 94]. On the other hand, monobrachial homology can lead to the formation of complex heterozygotes, involving chains or rings of metacentrics, that cause partial or complete infertility [103, 104, 98, 102, 99]. New homozygous karyotypes have been found (and classified as new CRs) in hybrid zones (reviewed in [91]). The process of new CRs forming from interactions of existing CRs with each other and with STs is called "zonal raciation" [105].

Table 3.1: Chromosomal races of the house mouse. 2n = diploid number of race; Pialek 2005 = appears in the most recent survey of CRs; Available = present in the archive of a collaborator; Collected = present in our archive; Genotyped = at least one individual has been genotyped on one of three arrays (MDA, MUGA or MegaMUGA).

	Fused Chromosomes																				
Code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	2n	Country	Pialek 2005	Available	Collected	Genotyped
ADOR		5	6	12			15	17		14	13						26	AT	1			
BBEL				12	10												36	BE	1	1	1	1
BNIV				12													38	BE	1	1	1	1
CHBA					7				16								36	CH	1			
CHBO	10																38	CH	1			
CHBU	18	5	6	12			15	16	14	17	13						22	CH	1	1	1	1
CHCG	3	14		12						11			16				30	CH	1			
CHCH		4	8		15	7			14	12	13					17	24	CH	1			
CHCL					13	7		15	14	12	18					17	26	CH	1			
CHDO				12						15							36	CH	1			
CHE1	3				7					11							34	CH	1			
CHE2	3			12	7					11							32	CH	1			
CHEL	3	14		12	7	11		17	10				16				24	CH	1			
CHGU	3	14		12	7	15		16	13	11							24	CH	1			
CHHN	3	8		12	7	15			14	11			16				24	CH	1		1	1
CHKU	3	14		12	7	15		17		11			16				24	CH	1			
CHMA	3	14		12		17	8			11			16				26	CH	1	1	1	1
CHMU				12													38	CH	1			
CHOL	3	14		12	7	15		9		11			16				24	CH	1			
CHPR					13	7		15	14	12						17	28	CH	1			
CHRM	3			12	7	15			14	11	w		16				26	CH	1			
CHRV	3	14		12			8			11			16				28	CH	1		1	1
CHSU	3			12			8			15							32	CH	1			
CHVI										11							38	CH	1			
CHWE			6	12			8		14	11			16				30	CH		1	1	1
CHPO	3			6	15			12	14		13					17	26	CH	1	1	1	1
DAIL		5	6	12					14		13						30	DE	1			
DBAI			6	12				17									34	DE	1			
DBAR				12							13						36	DE	1			
DCON	18	5	6	12			15	17	14	11			16				22	DE	1	1		
DDUD				12	14												36	DE	1			
DEYB			6	12				17		14	13						30	DE	1			
DFEL				12									14				36	DE	1			
DHOF				12							16		14				34	DE	1			
DHOL				12	15												36	DE	1			
DHUG				12										17			36	DE	1			
DIND				12				10					14				34	DE	1			
DLAN		5	6	12				17			13						30	DE	1			
DMOC		5	6	12				17		14	13						28	DE	1			
DNEU				12	14	10											34	DE	1			
DROT	5			12													36	DE	1			
DSTJ				12	15								14				34	DE	1			
DTIE				12	7	15				14							32	DE	1			
DWEN				12													38	DE	1			
DKEA		5	8			9											34	DK	1	1	1	1
DKOD		5	8														36	DK	1	1	1	1
DKSL			8														38	DK	1	1	1	1
DKWE		5	8						13								34	DK	1		1	1
EAVI				14	15	10			11			13					30	ES	1	1	1	1
EBAR				14	15				11			13					32	ES	1	1	1	1
EGAR			8	14	15	10			11			13					28	ES	1	1	1	1
EIBI													16				38	ES	1			

	Fused Chromosomes																				
Code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	2n	Country	Pialek 2005	Available	Collected	Genotyped
FSAN			14	12	10			13									32	ES	1	1		
FBIS				12	7					14							34	FR	1	1	1	1
FNAN				12													38	FR		1	1	1
GRAM			10		12											17	34	GR	1	1		
GRKA			6					12	16	14	17		15				28	GR	1	1	1	1
GRME										14							38	GR	1	1	1	1
GROL	3	5		6				12	16	14	17		15				24	GR	1	1	1	1
GRP1								12	16	14	17		15				30	GR	1	1	1	1
GRP2								12	16	14			15				32	GR	1	1		
GRPY																17	38	GR	1	1		
GRT1	11	15		14	12	9		17		13							26	GR	1	1		
GRT2		15		14	12	9		17		13							28	GR	1	1	1	1
HRSP					15	12		17	13	14							30	HR	1			
HRZA	11				15	12		17	13	14							28	HR	1			
IACR	2		9	17	13	16		14		12	15						24	IT	1	1	1	1
IALC					14			12		15							34	IT		1	1	1
IBIN		8	4		15	7			14	12	13					17	24	IT	1	1		
ICAN									14							17	36	IT	1			
ICAS		15	4		13	12	11	17	16	14							24	IT	1	1		
ICBO	18	17	13	11	15	7		14	16	12							22	IT	1	1	1	1
ICCH																17	38	IT	1	1	1	1
ICDE	7	18	8	15	17	13			16	11		14					22	IT	1	1	1	1
ICHI				12						11						17	34	IT	1	1		
ICOL									14	12						17	34	IT	1	1		
ICRE	6	8	4		15		18		14	12	13					17	22	IT	1	1	1	1
IFOR										12						17	36	IT	1	1		
IGAL		4	6		15		8		14	12	13					17	24	IT	1			
ILIP	2		9	13	14	16		12		15							26	IT	1	1	1	1
ILUI		4	8		13	7			14	12	18					17	24	IT	1	1		
ILVA	3	8		6	15		18		14	12	13					17	22	IT	1	1		
ILVC	18	4	8		15	7			14	12	13					17	22	IT		1	1	1
IMVA	3			6	15		18	12	14		13					17	24	IT	1	1		
IPAL		15	4		13	12		17	16	14							26	IT	1	1		
IPAN		15*	4		13*												36	IT		1	1	1
ISEV		12	4		15	7		11	14	13						17	24	IT	1			
IUVA	3	8		6	15				14	12	13					17	24	IT	1	1	1	1
IVUL	2		9	13	14			12		16						17	26	IT		1	1	1
PADC		4	14		18		15	11	12	16			17				24	PT	1	1	1	1
PASJ			8	16	14	7			10		12		17		18		24	PT	1	1	1	1
PEDC		4	14		18	7		11	12	16			17				24	PT	1	1	1	1
PLDB		4	14			7		11	12	16			17		18		24	PT	1	1	1	1
PPOD			8	5					18	16	12			17			28	PT	1	1	1	1
PSAN		19	8	16	14	7			10		12		17		18		22	PT	1	1	1	1
PSVI		4	8		18	7			12	16	19		17				24	PT	1	1	1	1
TUMO	11	16	12	6	14		18	9		17			15				22	TN	1	1	1	1
TRDE	4				16												36	TU	1	1	1	1
UKCA				10		13			12								34	UK	1	1		
UKED			14	10					12								34	UK	1	1	1	1
UKJO				10		13			12		14						32	UK	1	1	1	1
UKLY				10					12								36	UK	1	1		
UKWE					14				12								36	UK	1			

Zonal raciation may also occur due to the appearance of novel metacentrics. In several hybrid zones, mutations that have occurred by whole-arm reciprocal translocations (WARTs) have been observed. A WART is a type of translocation in which a whole chromosome arm is exchanged between a metacentric and at least one other chromosome (either metacentric or acrocentric). WARTs cause karyotypic rearrangement but have no effect on 2N. WARTs were first described in crosses between wild-derived mice [106, 107] and later found in natural populations [108]. The possibility of zonal raciation involving complex heterozygotes (for example, caused by WARTs) was initially rejected [109]; however, in at least one case a new CR (ISEV) appears to have been established through a WART in a very short timeframe (~ 20 years) [110]. Chromosomal phylogenetic studies have also suggested that WARTs can provide the most parsimonious explanation of the relationship between geographically proximate CRs [111, 112]. New metacentrics that arise within an existing karyotypically variable population can reproductively isolate carriers, which may then establish a new CR. A molecular study of two parapatric CRs in central Italy indicated the absence of gene flow between them despite the lack of any physical barrier to their interaction [113]. It was found that hybrids had a high degree of structural heterozygosity, and also that there was a high degree of genic variation between the CRs. Gene flow was especially reduced in the most proximal markers of metacentric chromosomes (pericentric regions) [114]. Reduced pericentric gene flow is likely associated with suppressed recombination near the centromere [115, 101]. It is difficult to know whether only one of those factors causes reproductive isolation, or if the two factors progressively reinforce each other. The negative effect of genic differentiation on the fitness of hybrids has been shown in other hybrid populations as well [116].

The karyotype composition of CRs may also be affected by differing fitness of certain chromosomal fusions, or differing susceptibilities of chromosomes to Rb translocation. Certain fusion pairs are repeated in several, sometimes geographically distant, populations while other pairs are never observed. Of the 342 possible autosomal metacentric pairs, only 103 have been discovered in at least one CR (Table 3.1). Three fusion pairs (4.12, 9.14 and 5.15)

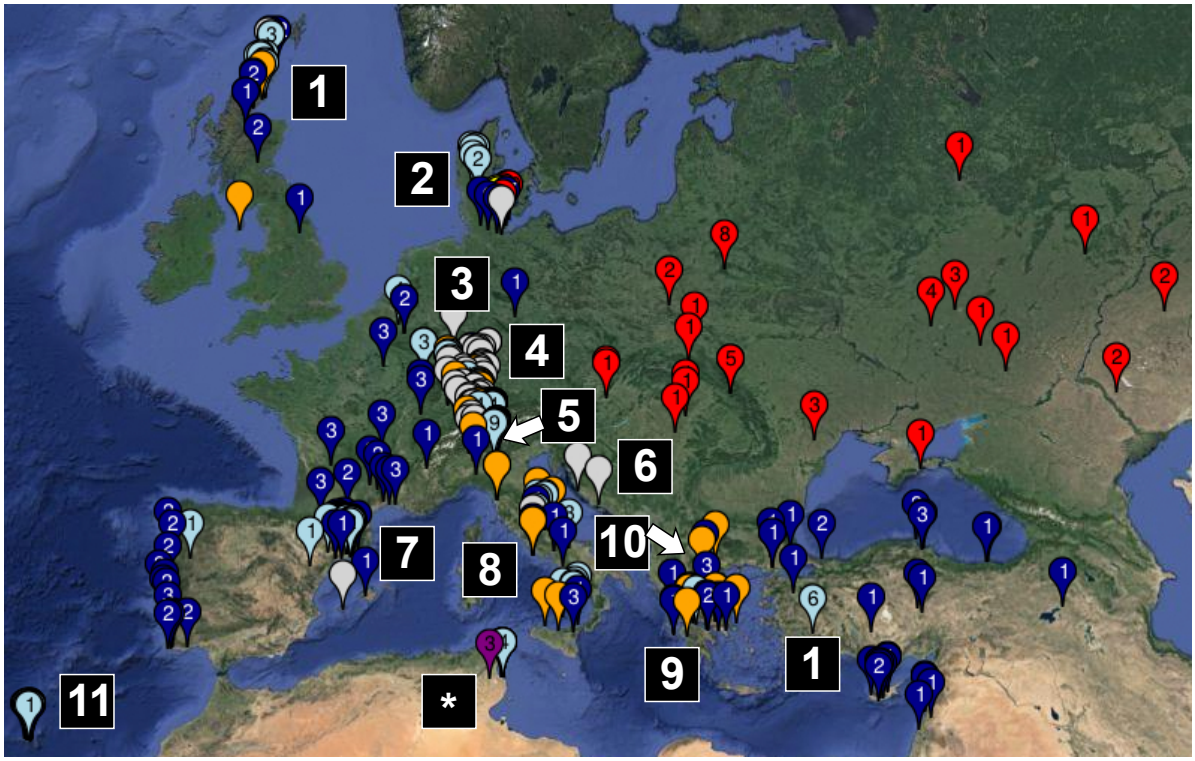


Figure 3.1: Geographic locations of chromosomal races and collected samples. Colored flags represent collected samples. Dark blue: standard population, Light blue: chromosomal race; Purple: standard and non-standard mice trapped together; Orange: races that may be collected in the future; Gray: remaining races; Red: *M. m. musculus* samples. Numbered boxes indicate the locations of chromosomal race systems identified in [91]. 1: Scotland; 2: Denmark; 3: Northern Europe/Northern Switzerland; 4: Southern Switzerland; 5: Northern Italy; 6: Croatia; 7: Barcelona; 8: Central/Southern Italy; 9: Peloponnesus (Greece); 10: Mainland Greece; 11: Madiera; asterisk: TUMO, a race that we have collected but does not belong to a system.

account for 15% of the 510 metacentrics found in all CRs (Figure 3.2). Chromosomes 4 and 12 alone account for 16% of all chromosomes that appear in fusion pairs, while Chr 19 is only found in two metacentrics. There is evidence that the underrepresentation of Chr 19 is due to pericentric enrichment for critical or dosage-dependent genes [117]. If instead the non-uniform distribution of fusion pairs is due to shared fusions between related CRs, the question still remains of whether those fusions proliferate successfully by chance or due to some selective advantage.

The complex interactions between karyotypically different populations have often made it difficult to draw clear boundaries between CRs. In staggered hybrid zones with geographically

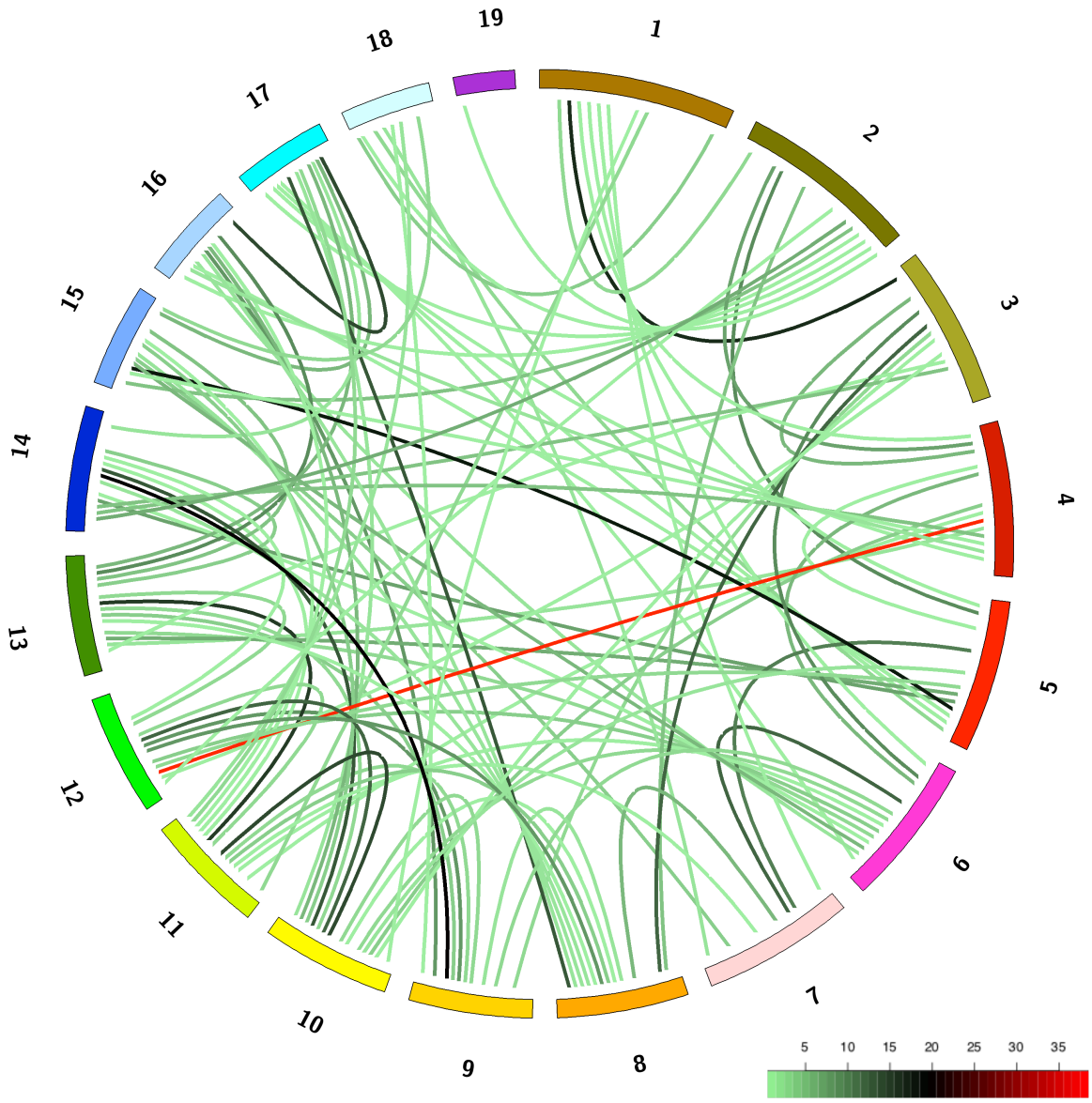


Figure 3.2: Distribution of Rb fusion pairs is non-random. Each line represents a Rb translocation. Line color indicates the number of Rb races fixed for that metacentric, from green (low) to red (high).

separated clines of multiple metacentrics [103], it is difficult to decide whether the mice with an intermediate number of metacentrics are a product of hybridization between a CR and a ST or are the result of the gradual accumulation of new Rb fusions. For example, in the CRs near Barcelona (Spain), no CR is completely fixed for any combination of metacentrics. Instead, chromosomal clines exist in which certain metacentrics are more frequent [118]. Capanna *et al.* (1974) first used the word "system" to describe a group of CRs that appear to share a common origin based on geographic proximity and shared metacentrics. Pialek *et al.* (2005) [91] described 11 such systems (and noted CRs that do not fit into a system, Figure 3.1).

It is worth noting that a few metacentric populations have been described in regions where *M. m. domesticus* is not the dominant subspecies: India [119], Japan [120] and the Czech Republic (Zima 1989). The sample sizes in all cases were small (3, 1 and 3, respectively). Interestingly, the single metacentric found in the Japanese mouse, 9.15, is not found in any CR of *M. m. domesticus*. Since Rb translocations are relatively common and only mildly deleterious, these few observations are not unexpected, even in light of the selection against metacentrics during meiosis in wild-type *M. musculus*. While the possibility that CRs exist in other *M. musculus* subspecies cannot be ruled out, all of my analyses make the assumption that, within *M. musculus*, extreme karyotypic diversity is exclusive to the *M. m. domesticus* subspecies [121].

3.2 Introduction to the study

There has been significant study and debate regarding the following questions: 1) Was there a single mutation event that enabled the genesis and proliferation of the CRs, or have there been multiple independent founder events? 2) How can mutations with presumably deleterious effects on viability and fertility be present (and, apparently, under positive selection) in many populations of *M. m. domesticus*? and 3) By what mechanism(s) are Rb translocations selected for and fixed within a CR? These questions must be considered in light of the fact that the karyotypic diversity of *M. m. domesticus* has arisen within an evolutionarily brief period

of time (10,000 years or less [122, 108]). Several mechanisms have been invoked to answer each of these questions (reviewed in [123]); however, they may be reduced to three general scenarios: 1) a high frequency of Robertsonian translocations in *M. m. domesticus*, some of which are fixed by chance in isolated populations; 2) a selective advantage associated with the metacentric chromosomes strong enough to overcome their deleterious fitness effects; and 3) a change in the direction of meiotic drive, such that metacentrics are under positive selection during female meiosis in CRs, as opposed to the known selection in favor of acrocentric chromosomes in *M. musculus sensu lato* [16]. Importantly, all three of these scenarios necessitate either environmental or genetic differences affecting the CRs relative to ST mice. Environmental differences have been proposed that result in a high mutation rate, such as mutagens (radiation, radon) associated with natural disasters [124], or that favor some phenotypic difference in the CRs [123]. While those explanations cannot be ruled out, they are currently only theoretical. The only evidence of selection acting on the fitness of metacentrics in the house mouse has shown that they are disfavored [95]. Furthermore, a greater ability of CRs to adapt to changed environmental conditions would likely be due to underlying genetic causes.

We set out to test the prediction that the ability of CRs to accumulate Rb translocations has a genetic basis. If preferential transmission of metacentrics is due to meiotic drive, then the centromeres themselves are the *responder* loci; we were instead concerned with identifying possible *distorter* loci. We determined that the most feasible test of our hypothesis was a case-control genome-wide association study (GWAS) based on a broad sampling of CRs and STs. A study using wild mice had some advantages over one using a laboratory population. First, there is direct evidence of selection in favor of metacentrics in the wild, whereas all previous attempts to create laboratory hybrids involving wild-derived mice with Rb translocations have displayed the wild-type preference for acrocentric chromosomes [95]. Second, large collections of tissues and DNAs from wild-caught mice existed in the hands of researchers that have been studying the CRs for the past 40 years. We determined that obtaining samples from those collections was more time- and cost-effective than trapping, importing, karyotyping and

breeding wild mice. Finally, we had no way of knowing the frequency of the *distorter* allele (or alleles) driving the accumulation of metacentrics. If both environmental and genetic factors favor the accumulation of metacentrics in CRs, then the causal allele might only be needed at a moderate frequency to give rise to the observed karyotypic variation. If we had sampled the wrong population or individuals, our study would have failed before it began. A potential limitation of our study design was that we could not directly determine the phenotype of each individual, as that would have required the live animal for cytogenetic or breeding experiments. However, those types of experiments would have been infeasible in any case due to time and cost constraints. Instead, we used diploid number (2N) as a proxy phenotype. In most cases, the 2N was known or could be predicted based on where the individual was trapped. We expected 2N to be positively correlated with the presence of the mutation. Therefore, we treated 2N as a continuous trait rather than a binary one (i.e., standard/non-standard karyotype).

3.3 GWAS design

A review of GWAS designs [125] supported the use of a two-staged approach [126] (Figure 3.3) as most effective in terms of cost and results. Therefore, sample collection, genotyping and analysis were organized into two stages. The first stage focused on an overall broad survey of populations, but with deep coverage in a small number of populations. The first stage used the MDA to provide the highest resolution for identifying candidate markers. In stage two, the sample size and genotyping depth will be expanded using MegaMUGA, which includes the most highly associated candidate markers from phase one. The scope of the project I will describe in this chapter encompasses all of stage one.

We assumed a "simple" genetic model (i.e., a single locus or small number of loci) by Occam's Razor; it is far less likely for a complex trait to have either spread long distances by gene flow or arisen multiple times independently. In addition, the apparently frequent changes in mammalian karyotype compositions over evolutionary time [16] suggest a recurring mu-

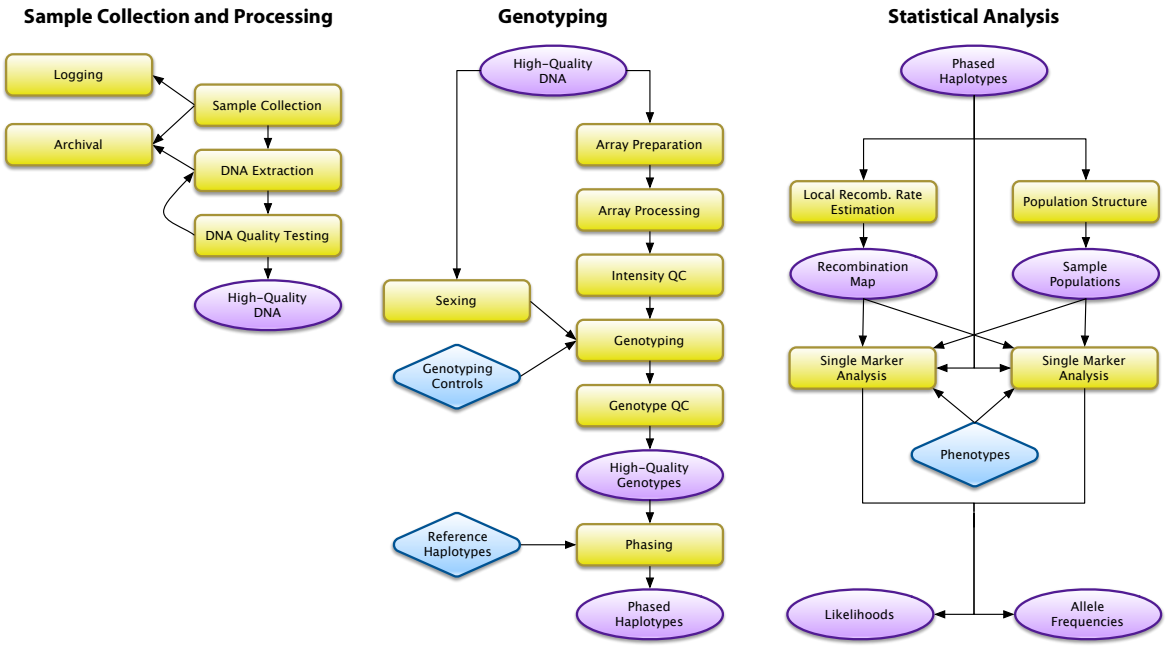


Figure 3.3: Design of our two-stage GWAS. Yellow rectangles represent manual or computational tasks; purple ovals represent resources or data generated by our study; blue diamonds represent data from external sources.

tation in a gene that plays key role in female meiosis. We expected the allele frequencies at *distorter* locus (or loci) to be very different in CRs (cases) and STs (controls). Females of all laboratory strains (which have a primarily *M. m. domesticus* genome [40]) that have been tested for meiotic drive have the standard phenotype (centromeric drive in favor of acrocentrics), including F1s between normal-karyotype strains and Rb chromosomal substitution strains [127]. We predicted the frequency of the causal allele to be very high in cases and less than 0.5 in controls (probably much less). We expected the frequency of the causal allele to be further bolstered by founder effect and drift in small, semi-isolated populations.

3.4 Results

3.4.1 The Wild Mouse Genetic Survey is a rich resource for mouse genetics

For the past several years, we have been contacting collaborators to request samples from CRs and STs. As of early 2014, we have collected 1,260 total samples from all three sub-

species (although 90% of samples are *M. m. domesticus*) and 44 CRs (Figure 3.1 and Table 3.1). Each investigator provided us with annotations for each sample, including a unique ID, sex, collection date and location (latitude and longitude from GPS when available), chromosome count (i.e., 2N), race code and detailed karyotype. Race codes are four letter identifiers, in which the first one or two characters identifies the country and the remaining characters identify the local population. Karyotypes were determined by a variety of methods, but primarily chromosome spread. Chromosome staining (aka G-banding) was always used for samples that were characterized during the initial discovery and description of a new CR. In some cases, samples obtained from STs or from well-characterized CRs were not karyotyped; instead, an inferred karyotype was reported.

We used several criteria to select for genotyping a subset of the samples that we collected. Our criteria were intended to maximize the amount of genetic diversity that we could capture. Our priorities were to select:

1. A minimum of four samples from each of the races we sampled;
2. A greater proportion of males than females, so that we could construct a phylogenetic tree of Chr Y;
3. Trapping sites separated by the greatest possible genetic distance;
4. Trapping dates separated by the greatest amount of time;
5. Samples with karyotypes that had been confirmed cytogenetically, and that were homozygous for all metacentrics; and
6. Races with lower 2N, based on our hypothesis that the genetic variants we were interested in discovering would be more highly correlated with races that had fixed a larger number of metacentrics.

Additionally, for population genetic analyses we chose two CRs, ICRE in northern Italy and UKJO in Scotland, to sample deeply (14 and 10 individuals, respectively). We also sampled a large number of related CRs within three different systems: northern Italy, Barcelona

(Spain), and the Island of Madeira (Portugal). We gave particular focus to the northern Italian system for several reasons: 1) It contained many CRs in relatively close proximity yet harboring substantial karyotypic diversity, enabling us to effectively study the relationship between genetic and chromosomal variation; 2) The majority of races had fixed nine Rb translocations (the maximum possible), meaning the frequency of the causal allele among cases was likely to be high; and (3) There are available two live and several additional cryopreserved wild-derived laboratory strains established from the northern Italian system (three of which have been genotyped on MDA [40]), and it is important to understand the relationship between the natural and laboratory populations if those strains prove useful for later experiments.

In total, we genotyped 385 unique *M. m. domesticus* samples across the three platforms. I created a database called the Wild Mouse Genetic Survey (WMGS) that provides multiple views of our data set. First, it provides access to the high-density MDA genotypes from 103 samples that I used to conduct the first-stage GWAS (described later). Second, it provides access to genotypes from all samples for only the 1,163 markers that were common and well-performing across all platforms and were segregating within the *M. m. domesticus* samples. Although the later data set sacrifices depth for breadth, it will be more useful than the former for population genetic analyses.

3.4.2 Substantial population structure exists in *M. musculus* mice genotyped on MDA

The MDA was designed with SNPs ascertained in classical inbred strains and a small number of wild-derived strains [60]. Whole-genome sequencing revealed that approximately 75% of sites that are polymorphic in classical strains are also polymorphic in wild-derived strains [56]. This suggests that the majority of markers on MDA (and MUGA and MegaMUGA) should be polymorphic in wild mice. I directly tested this hypothesis by computing the minor allele frequency (MAF) distribution in our samples.

Of 547,782 well-performing MDA SNPs, only 36,735 (6.7%) were monomorphic when considering all wild mouse samples. An additional 107,036 SNPs (19.5%) had a rare allele

(defined as MAF < 0.05). Monomorphic markers are by definition uninformative (since there is no genetic variability to be associated with phenotypic variability). Markers with rare alleles are typically excluded from GWAS studies unless sample sizes are very large (ten thousand or more), because strong statistical power is required for associations with rare alleles to exceed significance thresholds. Therefore, the fraction of MDA markers useful for mapping (73.5%) closely matched the estimation from sequence data. MAFs were slightly biased toward lower values, with approximately half of informative markers having a MAF between 0.05 and 0.2 (Figure 3.4 A).

The three *M. musculus* subspecies are monophyletic and have been reproductively isolated for 250,000 years or more (except at hybrid zones, see Chapter 2); therefore, I expected a large number of alleles to be private to each subspecies, i.e., fully diagnostic. I tested this hypothesis using principal component analysis (PCA), a statistical procedure that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. When applied to sequence data, the variables correspond to markers or sets of markers. PCA quantifies the amount of variation observed in the sequence data that is explained by each variable. The relationship between markers in a set may be inferred from SNP annotations (for example, functionally or ancestrally related markers), or they may be inferred by visual examination of sample clustering. PCA provides a means to uncover (sometimes cryptic) relatedness between samples. I used `EIGENSTRAT` [128] to perform PCA on MDA genotypes. When all samples were included in the data set, 53.5% of the variance was explained by subspecific origin. I conclude that the majority of the genetic variation present in *M. musculus* is due to intersubspecific differences.

While fully diagnostic alleles are important for assigning ancestry [40], they are not useful for association mapping a trait (such as ours) that is only present in one subspecies. When considering only *M. m. domesticus* samples, 187,771 SNPs (34.3%) were monomorphic and 101,635 SNPs (18.6%) had a rare minor allele. Therefore, 258,191 SNP markers (47.1%) were available for our GWAS. Markers informative for *M. m. domesticus* were also slightly

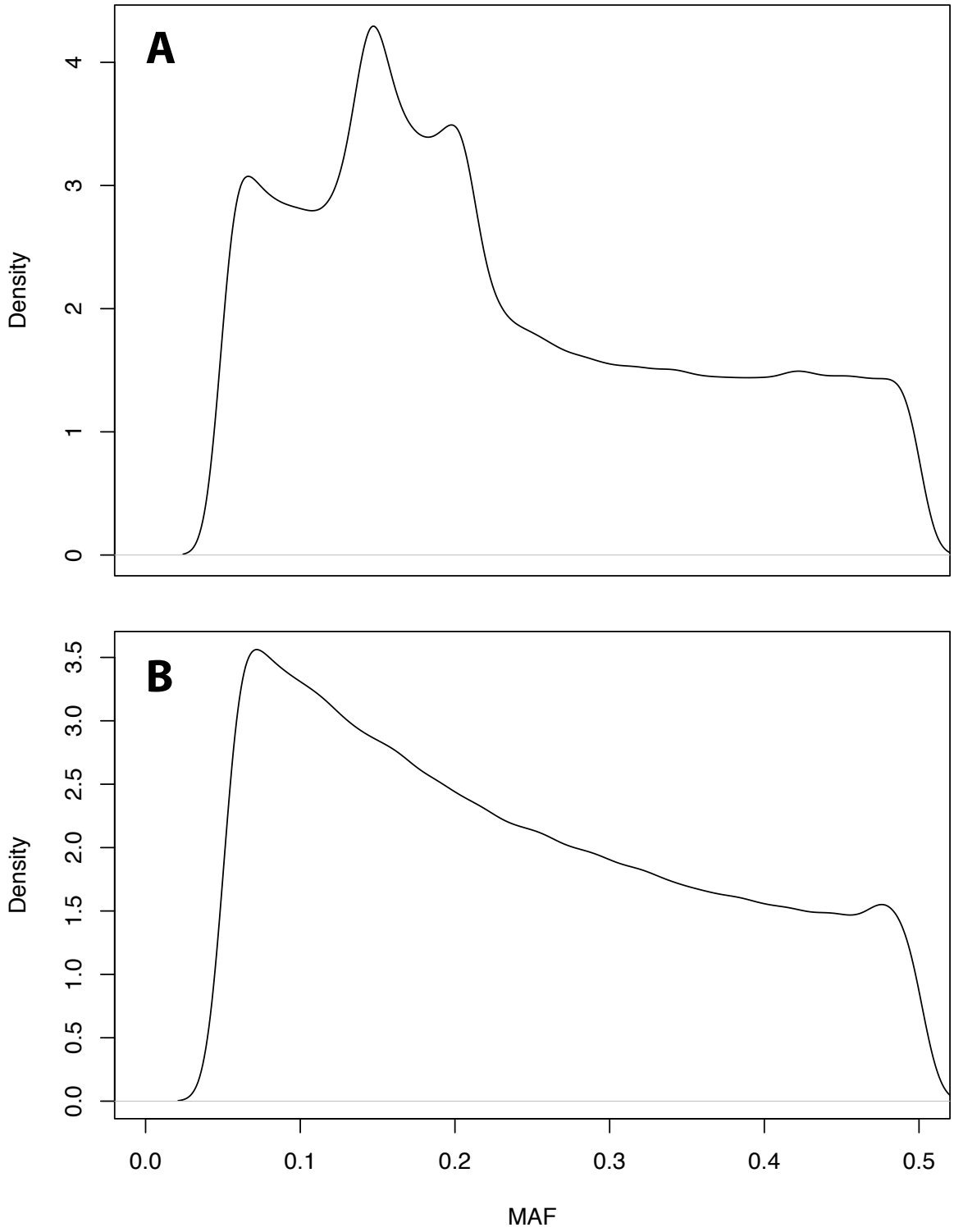


Figure 3.4: MAFs at MDA markers are weighted toward lower values in wild mice sampled from A) all three *M. musculus* subspecies and B) only *M. m. domesticus*. MAFs are only shown for common alleles (i.e., MAF $>$0.05).

biased toward lower MAFs (Figure 3.4 B). I conducted a PCA using only *M. m. domesticus* samples (Figure 3.5). The first principal component explained 26% of variance in the data and separated samples by gross geography. The second component separated sub-populations. In most cases, that meant that one or more CRs were differentiated from STs; although, as expected, the Barcelona CRs and STs were not differentiated by the second component due to the clinal structure of that population. These results suggest that genetic variation between CRs and STs should be sufficient to identify associations (should they exist), but not so great as to present an intractable number of candidate associations to validate.

VINOs may also be used in association mapping by treating them as additional markers in tight linkage to the SNP marker [41]. There were 20,794 and 12,064 VINOs with 5% or greater V calls in the full set of samples and in *M. m. domesticus* samples, respectively.

Another important consideration for association studies is whether, for each SNP, the frequency of each genotype is not significantly different from the expected value, which may be determined from allele frequencies. The relationship between allele frequency and genotype frequency is known as the Hardy-Weinberg principle [129]. Briefly, the fraction of samples with each homozygous genotype is expected to be the square of the frequency of each allele, and the fraction of samples with the heterozygous genotype is expected to be twice the product of the allele frequencies (assuming a bi-allelic marker). I used an exact test of Hardy-Weinberg equilibrium (HWE) to identify markers that deviated from this expectation [130]. After correcting for multiple testing, I found that 30.8% of *M. m. domesticus* informative markers deviated significantly from HWE ($p < 0.05$). Of the markers that deviated from HWE, 96.4% were due to a dearth of heterozygotes. Although troubling for our study, this observation was expected based on previous studies of heterozygosity within local populations of mice, which showed that observed heterozygosity is consistently and substantially lower than expected [131].

In association studies, subpopulations within a sample that are more closely related to each other than to other members of the sample (population structure) can lead to false-positive as-

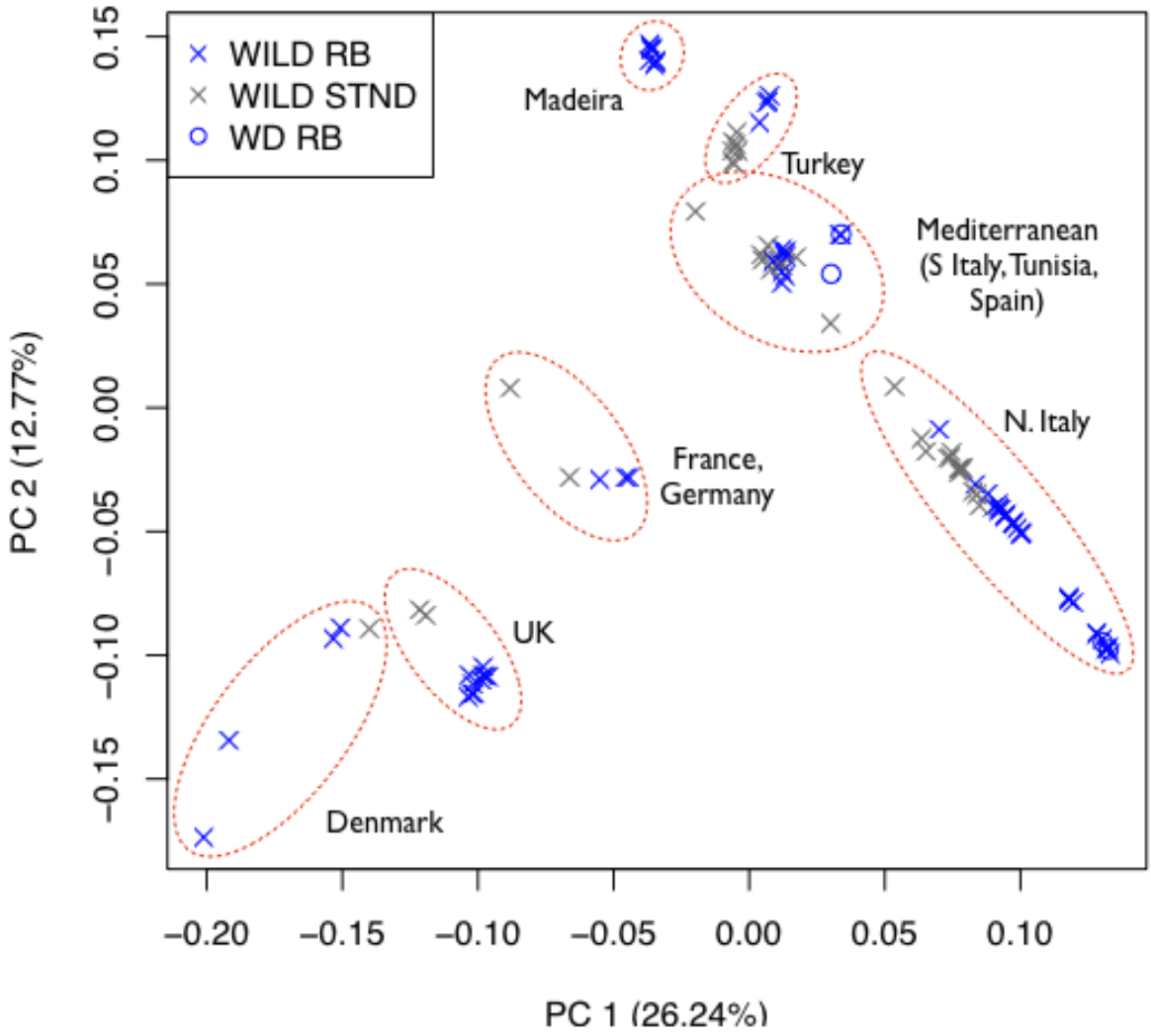


Figure 3.5: Principal component analysis of wild *M. m. domesticus* mice. The x- and y- axis show the first and second principal components, respectively. Shapes and colors represent sample type. Blue X: wild-caught mice from chromosomal races; Gray X: wild-caught mice from standard populations; Blue circle: inbred mice derived from wild-caught chromosomal race mice. The first component differentiates geographic populations (enclosed in red dotted circles), while the second component differentiates sub-populations.

sociations. I used the `FineStructure` algorithm [73] to identify population structure in our data set. `FineStructure` is a two step method, in which chromosomes are first "painted" and then individuals are clustered using a model-based Bayesian method. Computational chromosome painting is similar to the method used in [40] to assign subspecific origin; the genome is broken into blocks in which the number of haplotypes is much smaller than the number of samples, and samples with the same haplotype are assumed to be IBD within that region. The clustering step determines whether pairs of samples are IBD across enough of their genome to be considered as part of the same population.

To use the `FineStructure` algorithm, it is recommended to first have phased genotype data. I used the `fastPHASE` algorithm [132] to accomplish that task. Phasing was problematic in our dataset since a comprehensive set of reference haplotypes do not exist for wild mice. Therefore, I used the simplified method of conditioning each sample on all other samples. Conditional phasing basically treats each sample, one at a time, and predicts the phase (i.e., which alleles are co-located on the same chromosome) from the local haplotypes observed in the other samples. I compared the phased genotypes from `fastPHASE` with another program, `IMPUTE2` [133] and found them to be generally consistent.

The result of `FineStructure` was a set of 83 independent populations, of which 64 were *M. m. domesticus* (Figure 3.6). I computed allele frequencies by sampling one individual from each *M. m. domesticus* population and re-tested the markers for deviations from HWE. I found that, after accounting for population structure, only 7.7% of markers failed. In summary, our MDA data set contains substantial population structure that can largely be corrected either by subsampling from independent populations or by including a relationship matrix as a covariate in the model of association.

3.4.3 Heterozygosity varies widely in *M. m. domesticus* populations

Heterozygosity is an important metric when considering whether a population is suitable for GWAS. The more inbred a population is, the fewer segregating QTLs there will be. In our

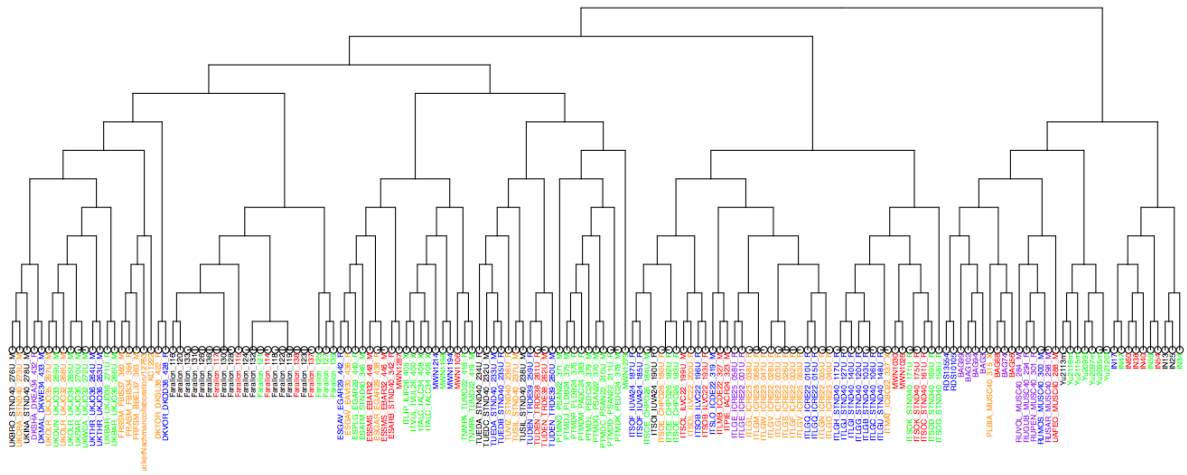


Figure 3.6: Populations within wild mouse samples. Dendrogram of 160 wild mouse samples based on autosomal MDA genotypes. For each population, one of six colors was chosen arbitrarily. Colors are reused and are only used to differentiate consecutive populations in the tree.

experiment, we know that the phenotype is segregating in wild mice, and so a highly level of inbreeding is actually desirable. On the other hand, exploratory GWAS studies (with no prior expectation of phenotype divergence within the population) would prefer a population with low inbreeding [134].

The simplest measure of individual heterozygosity is the fraction of markers with a heterozygous genotype. However, heterozygosity is most often determined using a specific application of F-statistics, called an inbreeding coefficient [135]. Inbreeding coefficients range from -1.0 (completely heterozygous, such as with an F1 intercross) to 1.0 (completely inbred), with a value of 0.0 indicating that the sample's observed heterozygosity is not significantly different from the Mendelian expectation of 50%. The inbreeding coefficient is inversely proportional to the fraction of heterozygous sites.

While an inbreeding coefficient is a property of an individual, the heterozygosity of a population can be determined from the distribution of inbreeding coefficients. I selected one individual from each of the independent *M. m. domesticus* populations identified in the previous section and computed their inbreeding coefficients using `PLINK`. Coefficients ranged widely (0.19 – 0.95) with a mean of 0.52 $\pm$ 0.23 (Figure 3.7). Rather than exhibit random

variation, heterozygosity and inbreeding coefficients appeared to be consistent within populations. I was interested to know if heterozygosity was a result of population-specific factors (such as inbreeding or founder effects) or if it was causally related to the accumulation of metacentrics. I grouped all individuals by CR and created a boxplot to show the mean and variance of heterozygosity within each race (Figure 3.8, green boxes). The correlation between a CR's mean heterozygosity and its 2N was not significant. Qualitatively, CRs with higher 2Ns tended to have higher heterozygosity, but CRs with low 2Ns varied widely in their levels of heterozygosity. I next tested whether heterozygosity levels differed between metacentric and acrocentric chromosomes. For each CR, I computed the distribution of heterozygosities for metacentric and acrocentric chromosomes independently (Figure 3.8, gray and orange boxes, respectively). There was a general trend of metacentric chromosomes being less heterozygous than acrocentric chromosomes (17 of 25 races), however none of the differences were significant after correcting for multiple testing. I conclude that heterozygosity is generally a feature of systems of related CRs. The fact that we are using populations with such widely different levels of heterozygosity in our GWAS suggests that our power to detect associations may be limited. While this is not necessarily a hinderance for our study, exploratory association studies in wild mice will require careful sample selection. The availability of the WMGS will help investigators to identify appropriate populations for their studies.

3.4.4 Linkage disequilibrium decays rapidly in wild mice

While the statistical power of an association study is determined by the sample size and the allele frequencies, resolution is determined by linkage disequilibrium (LD). LD is the non-random association of alleles at two or more loci that descend from the same ancestral chromosome. In outbreeding populations, LD is expected to be low due to, on average, many generations of recombination separating unrelated individuals [136]. LD may create false associations in GWAS studies, and therefore it is common to filter SNP data so that the correlation between adjacent markers is below some threshold. The standard measure of LD

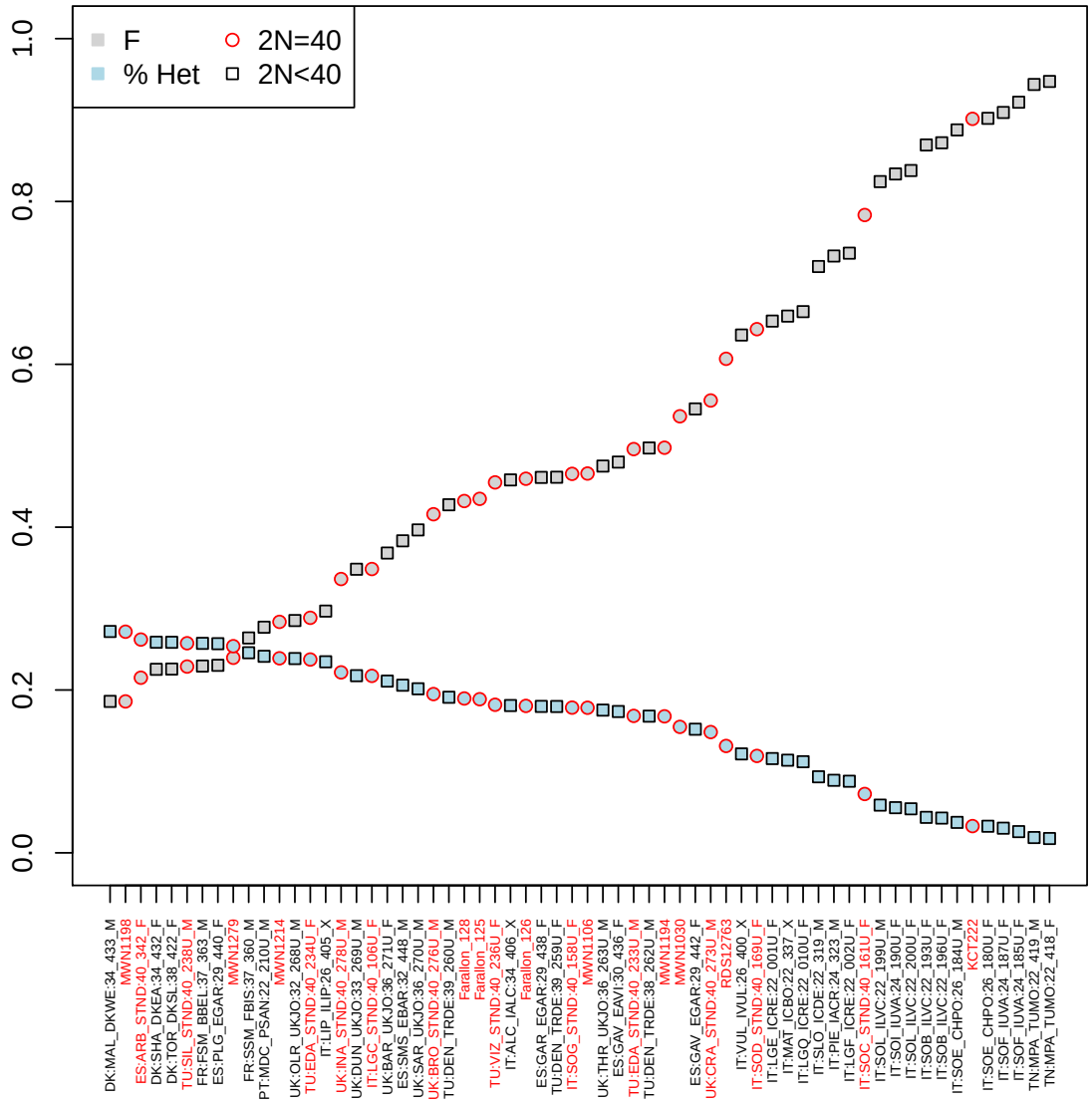


Figure 3.7: Inbreeding is variable in wild *M. m. domesticus* mice. Inbreeding coefficient (F, gray points) and the fraction of SNPs that have a heterozygous genotype call (blue points) are shown for individuals sampled from independent *M. m. domesticus* populations. Individuals with a standard karyotype are labeled in red and shown as circular points with red borders, while individuals with 2N <40 are labeled in black and shown as square points with black borders.

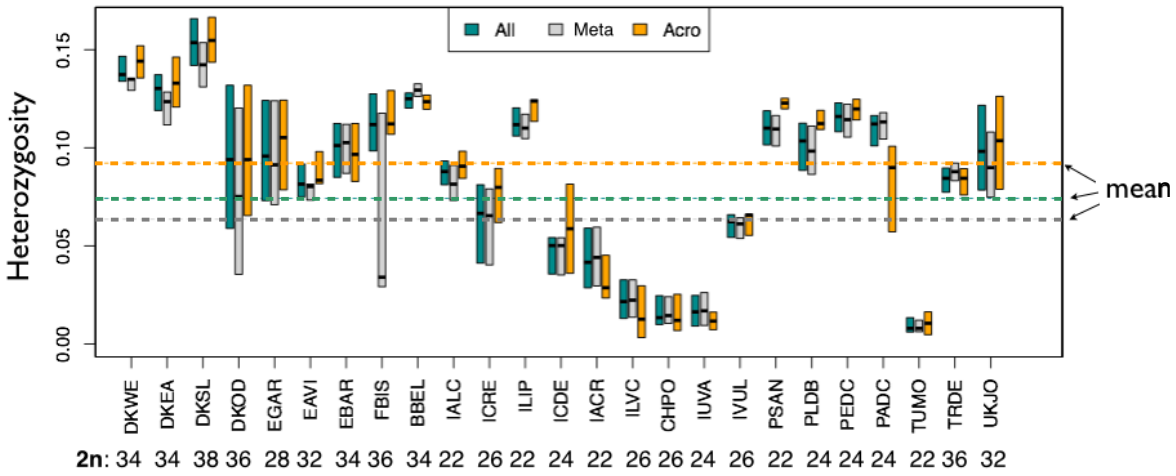


Figure 3.8: Heterozygosity is variable in chromosomal races. For each race, boxplots of heterozygosity, measured as the fraction of heterozygous SNPs, is shown for only acrocentric chromosomes (orange), only metacentric chromosomes (gray), and all chromosomes combined (green). The 2N is shown for each race below its four-letter code. Dashed lines show the mean heterozygosity for the three different chromosomal classes.

between a given pair of markers is the squared correlation coefficient, r^2, which is the normalized deviation of the observed haplotype frequencies from their expected frequencies. The choice of cutoff to determine which markers are in LD is somewhat arbitrary; commonly used values range from 0.2 - 0.7.

I used `PLINK` [137] to compute LD between all pairs of informative markers in the full MDA data set. I first examined the overall pattern of LD by binning the genome into 500 kb windows and computing the 95th-percentile r^2 value for each pair of bins (Figure 3.9 A). Most of the genome exhibits moderate-to-high LD that is reflective of the strong population structure that exists between the three *M. musculus* subspecies. As expected, LD on Chr X is higher than the rest of the genome due to a lower recombination rate. Interestingly, several regions (dark blue bands) exhibit little-to-no LD with the rest of the genome. These regions may be of potential interest for future study because they are expected to have a different inheritance pattern than than the rest of the genome. For example, the largest such region is on Chromosome 17 and coincides with the mouse *t*-haplotype region, which is subject to male-specific non-random transmission due to the lethality of certain genotypes. When considering

only unrelated *M. m. domesticus* mice, LD is extremely low genome-wide (Figure 3.9 B).

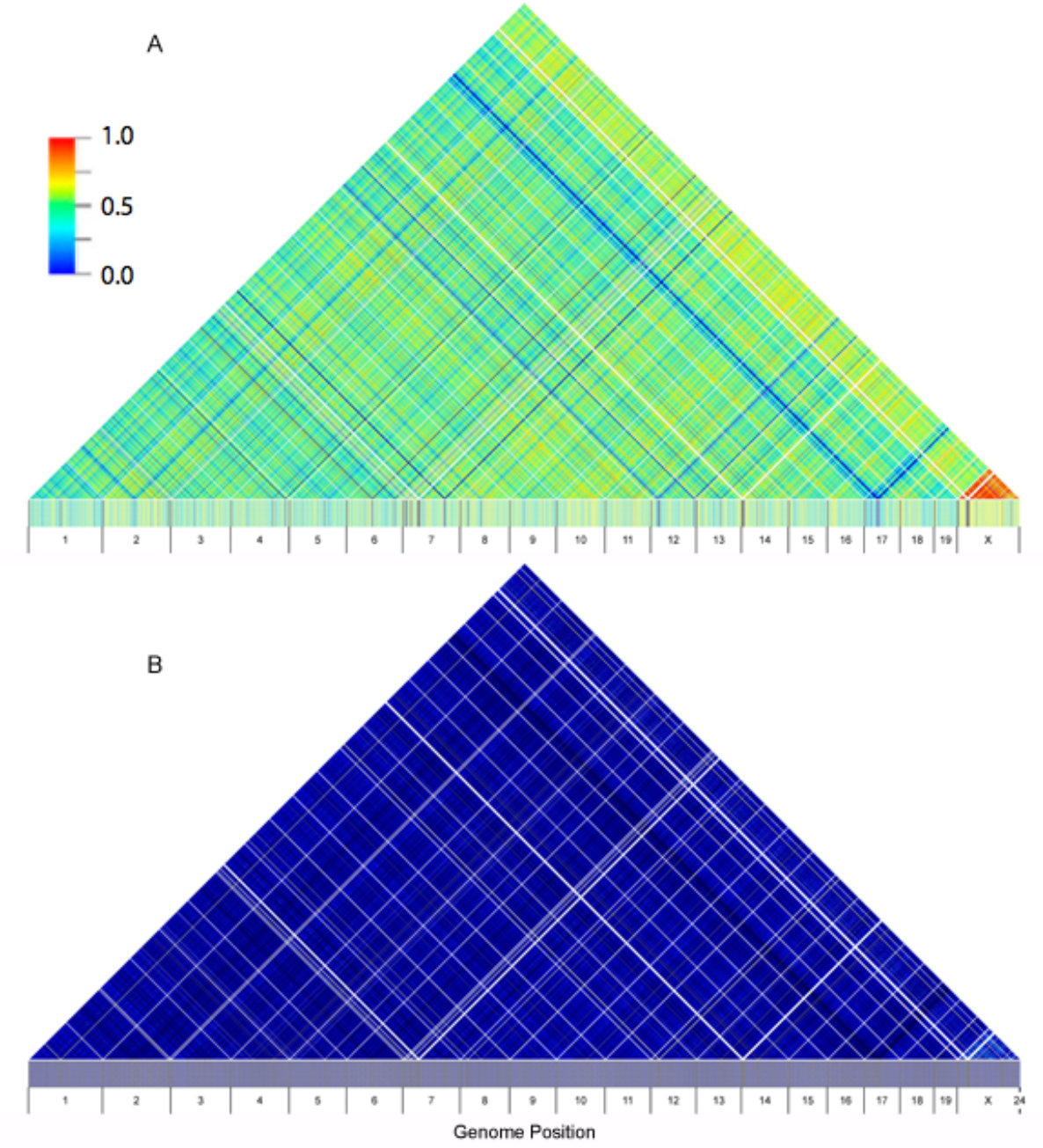


Figure 3.9: Linkage disequilibrium is minimal in wild *M. m. domesticus* mice. Heatmaps of r^2 values for mice from A) all three *M. musculus* subpecies and B) only *M. m. domesticus* show that LD is moderate-to-high (green-to-red pixels) in the former and very low in the latter (blue pixels). Each pixel represents the squared correlation coefficient (r^2) between the two markers diagonally adjacent to it. White bands represent regions lacking markers.

The resolution of an association study is primarily determined by the rate at which LD

decays, i.e., the extent of the region surrounding a marker that is expected to be in linkage with that marker. A population with slow LD decay decreases the number of markers required for mapping but increases the size of the average candidate interval, which in turn increases the difficulty of identifying a causal variant. Due to the availability of high-density genotyping arrays such as the MDA, rapid LD decay is almost always preferred to the alternative. I used `PLINK` to compute the r^2 for all markers within 5 Mb of each informative marker in unrelated *M. m. domesticus* mice. I then generated a histogram of both the mean and the 95th percentile values using a bin size of 1 kb (Figure 3.10). LD in both wild mice and humans was found in another study to fall below $r^2 = 0.4$ after 100 kb [136]; in our samples, equivalent LD decay occurred within 80 kb. Since the mouse genome has, on average, about one gene per 100 kb, association mapping in wild mice can potentially offer single-gene resolution.

3.4.5 First-stage GWAS identifies a significant association between genotype and 2N

After our lab transitioned away from using the MDA array in favor of the lower cost Illumina platforms, we decided to end the first stage of the study and conduct the first stage of our GWAS. For this study, I only used *M. m. domesticus* samples. From the populations defined by `FineStructure`, I selected one individual at random from each population and excluded the rest. In the second-stage GWAS we will use the more robust method of regressing the phenotype on covariates (such as population structure) using standard multiple linear regression software, and then using the residuals from the regression as the phenotype values. The final study sample consisted of 39 cases and 25 controls.

I restricted my study to the autosomes. In addition filtering for missingness and MAF, I used `PLINK` to prune SNPs in local LD. The final data set consisted of 49,600 SNPs, with an average pairwise spacing of 50 kb. For the chosen samples and SNPs, I created a filtered set of haplotypes from those I generated for the population structure analysis. The use of haplotypes for association studies is recommended [138].

I used Bayesian association mapping software, `BIMBAM` [139]. `BIMBAM` computes

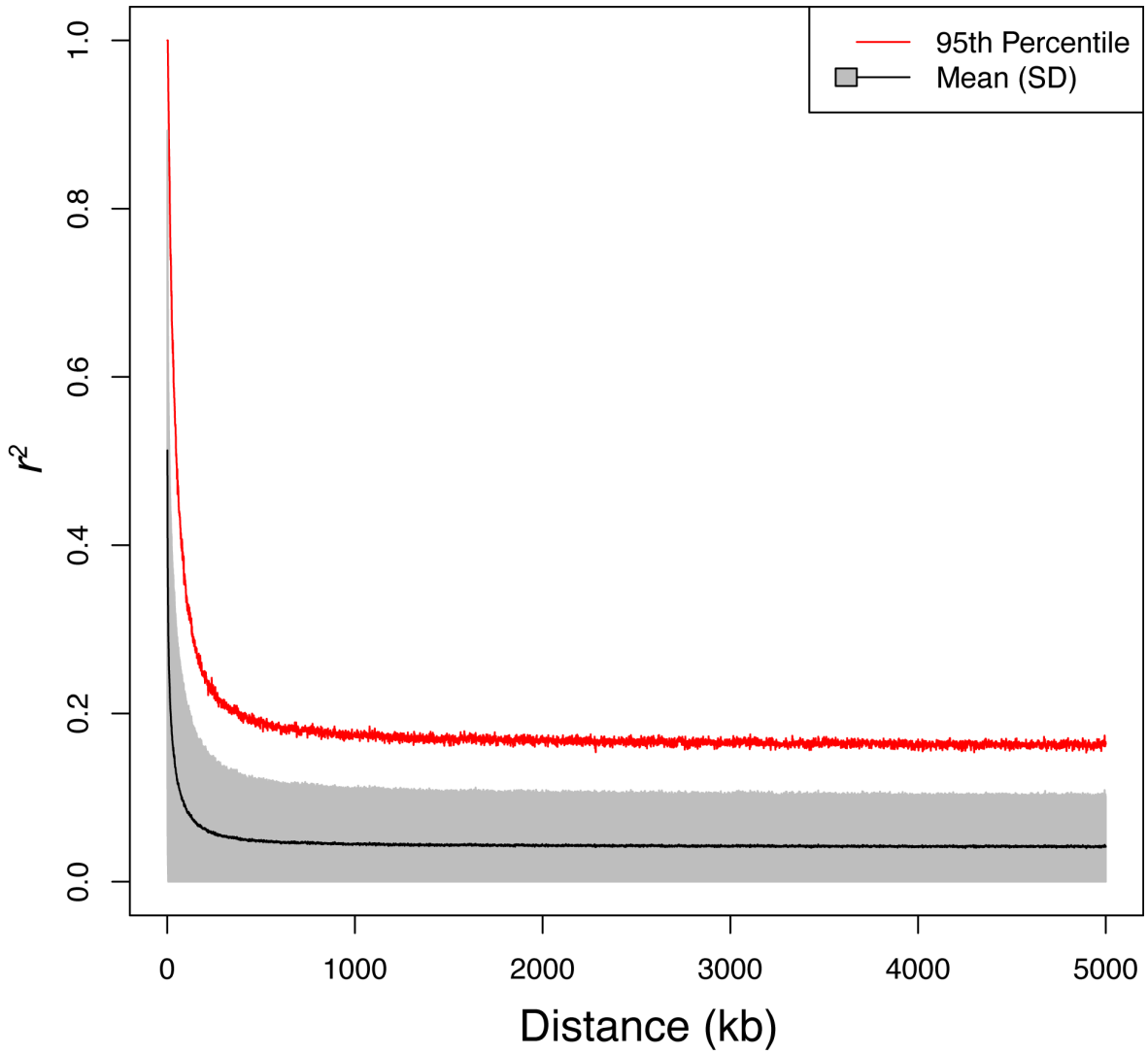


Figure 3.10: LD decays rapidly in a sample of unrelated *M. m. domesticus* mice. The mean (black line) and 95th percentile (red line) r^2 values are shown for inter-SNP distances up to 5 Mb, in 1 kb bins. The gray region shows ±1 standard deviation. Commonly used r^2 thresholds for LD are between 0.3 and 0.7.

Bayes Factors (BF) for each SNP under a model that incorporates both additive and dominant effects. I used the normalized 2N as the phenotype for each sample. I specified an unbiased prior [140, 141] expectation of association, $1/N = 2.0x10^{-5}$, where N is the number of SNPs, to reflect our expectation of only a small number of associations.

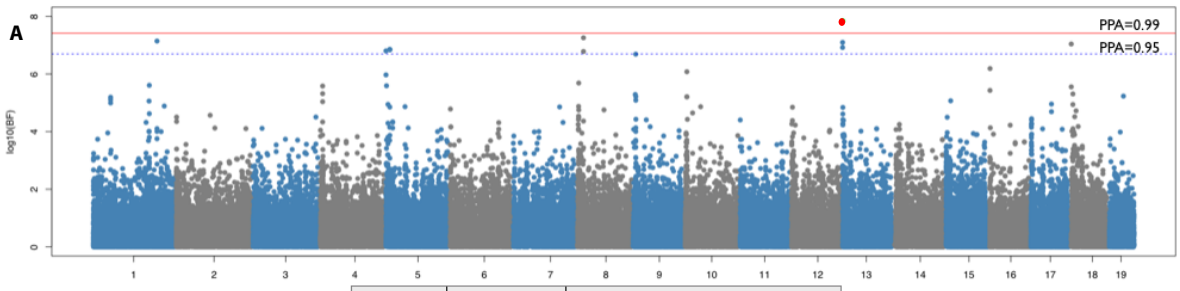
After running `BIM-BAM`, I created a Manhattan plot of the results by plotting $log10(BF)$ for each SNP (Figure 3.11 A). To compute the thresholds for significant and suggestive associations, I used the equations given in [141]:

$$BF = PPA/((1 - PPA) * (\pi/(1 - \pi)))$$

where π is the prior probability of association and PPA is the posterior probability of association. I used PPAs of 0.99 and 0.95 for the significant and suggestive thresholds, respectively. Those were quite conservative values compared to those used or suggested in the literature [141, 142].

I identified one significant and six additional suggestive peaks (Figure 3.11 B). Chi-squared tests of the most highly-associated SNPs under those peaks supported the alternative hypothesis of association between genotype and phenotype with small *p*-values. I conducted a set of checks for spurious association, including whether the marker was an outlier in any QC statistic and consistency of allele frequency across platforms; none of the significant or suggestive SNPs failed these checks.

I attempted fine mapping of the significant association on Chr 13. I obtained genotypes for all *M. m. domesticus* samples genotyped on MDA (both wild-caught and wild-derived) for the 200 kb (14 markers) adjacent to the most highly associated marker and partitioned them into cases and controls (2N $<$ 40 and 2N = 40, respectively). I then colored each allele of each SNP differently, which further partitioned the samples into haplotype groups. I identified 20 unique haplotypes, six of which were only present in cases, and ten of which were only present in controls (Figure 3.12). The most common haplotype in the cases (55/81 = 68% of individuals) was also present at low frequency in controls (7/79 = 9% of individuals). A Chi-squared test



B

		All Samples		Unrelated Samples				
Chr	Pos	BF	PPA	BF	PPA	Rb:ST	X^2 (2df)	p-value
1	153.92M	10.579	0.037	7.144	0.982	5.415	26.449	1.8E-06
1	171.18M	16.113	1.000	4.887	0.227	3.514	22.931	1.0E-05
4	3.79M	13.64	0.978	2.891	0.003	1.714	37.756	6.3E-09
5	3.09M	6.309	0.000	6.803	0.960	6.564	58.214	2.3E-13
5	12.29M	14.401	0.996	6.853	0.965	3.815	41.742	8.6E-10
6	4.29M	13.874	0.987	3.136	0.005	1.607	6.161	4.6E-02
8	17.03M	9.141	0.001	7.257	0.986	4.128	34.292	3.6E-08
9	9.37M	10.89	0.072	6.688	0.949	4.231	34.462	3.3E-08
10	6.62M	13.507	0.970	6.079	0.821	3.692	24.231	5.5E-06
10	76.92M	13.433	0.964	2.059	0.000	1.207	11.758	2.8E-03
11	12.77M	13.621	0.977	2.288	0.001	1.658	15.206	5.0E-04
11	89.52M	14.46	0.997	2.506	0.001	1.789	23.558	7.7E-06
13	3.01M	17.706	1.000	7.843	0.996	18.455	71.813	2.2E-16
13	5.02M	13.636	0.977	3.809	0.024	1.651	9.780	7.5E-03
15	93.07M	13.723	0.981	3.017	0.004	2.248	30.364	2.6E-07
16	4.02M	11.088	0.109	6.19	0.855	4.077	31.289	1.6E-07
17	4.49M	13.667	0.979	3.913	0.030	2.769	23.306	8.7E-06
18	3.27M	14.207	0.994	7.041	0.977	10.770	23.736	7.0E-06

Significant
Suggestive
Non-negligible

Figure 3.11: GWAS identifies a significant association on Chr 13. A) Manhattan plot of Bayes Factors (BF, log10 scale) for each SNP. Chromosomes are given alternating colors. Blue dotted line and red solid line represent thresholds for suggestive and significant associations, respectively. Red point above Chr 13 marks the significant association. B) Table of significant (red), suggestive (yellow) and non-negligible (green) associations. BFs are given for the study run before and after population structure correction. Chi-squared values and *p*-values are given for single marker association tests.

showed the first five SNPs to have allele frequencies that were significantly different (after correction) between cases and controls. Assuming that linkage extends up to 50 kb, the causal variant would be between Chr 2 2.95 – 3.10 Mb. However, the true proximal boundary of the candidate interval is unknown due to the lack of markers proximal to 3.0 Mb. There is a high likelihood that the causal variant is the centromere itself, lies within the pericentric region (proximal to 3 Mb), or is novel with respect to the reference sequence, because there are no annotated genes within the first 100 kb of Chr 13, and there is only a single (non-coding) gene within the first 500 kb of Chr 13.

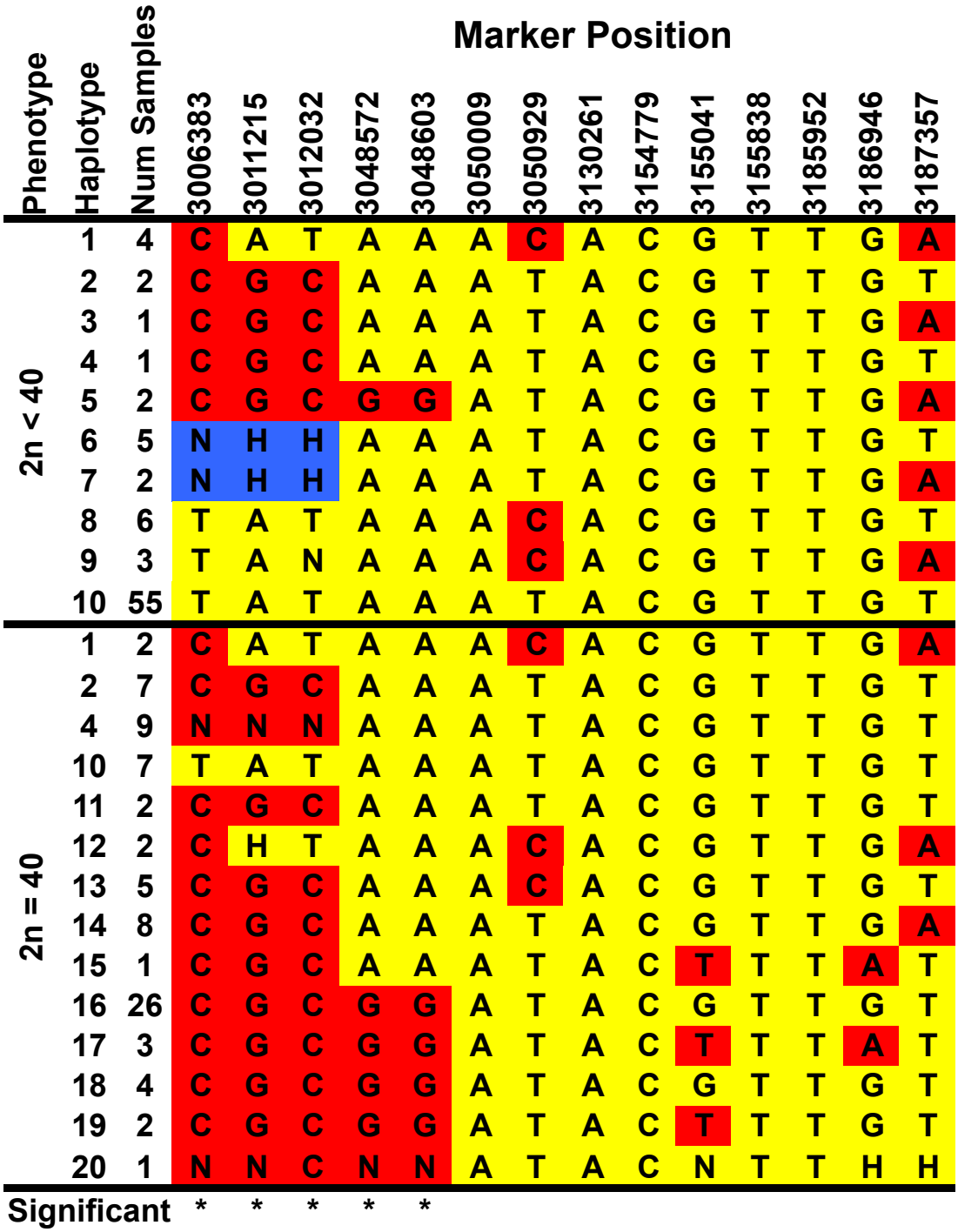


Figure 3.12: Haplotype analysis of significant assocation on Chr 13. Each row is a different haplotype, with cases on top and controls on the bottom. The position and genotypes for each marker are given for each haplotype; alleles are colored arbitrarily as red or yellow for homozygous genotypes and blue for ambiguous (heterozygous or missing) genotypes. Markers for which allele frequencies differ significantly between cases and controls are identified by an asterisk.

3.4.6 Pericentric regions have reduced genetic variation in CRs

The majority of peaks identified in our GWAS occurred in the vicinity of the centromere (Figure 3.11 A). This is an important result in light of our expectations about the genetic architecture of centromeric drive and also previous findings regarding gene flow in the vicinity of genomic regions involved in chromosomal rearrangements. A genetic element that influences segregation by promoting some types of centromeres over others would benefit by being tightly linked to a centromere. Selection acting on a pericentric locus is expected to reduce the rate of recombination near the centromere due to the preferential transmission of non-recombinant chromosomes. Therefore, gene flow is also expected to be reduced in the vicinity of the centromere of metacentric chromosomes [143]. That hypothesis is supported by several previous studies [115, 114, 144].

Chromosome-wide heterozygosity was not significantly different between metacentric and acrocentric chromosomes (Figure 3.8). I tested whether this was also true for pericentric regions, or whether our data was consistent with reduced gene flow in pericentric regions of metacentric chromosomes. I computed the fraction of heterozygous SNPs in the 200 markers proximal to each centromere in each ST and CR individual, and then separated them into three chromosomal classes: standard acrocentric, acrocentric in a CR indivdiual, and metacentric in a CR individual. I found that metacentric pericentromeres had markedly lower heterozygosity than CR or standard acrocentrics (0.8%, 5.2% and 7.6%, respectively). I tested the significance of this finding by comparing the distributions of heterozygosity in the three different chromosome sets using a Kolmogoriv-Smirnov test. I generated random combinations of metacentrics from the standard acrocentrics and from the CR acrocentrics in order to make the comparisons more equivalent. All three tests were significant, although the metacentric group was much more significantly different from either of the other two (both tests had *p*-values lower than the detection limit of $2x10^{-16}$) than the acrocentric sets were from each other ($p = 0.0005$, Figure 3.13 A). I next examined whether this was a general feature of metacentric chromosomes, or whether reduced heterozygosity was specific to certain metacentrics.

I found that on all autosomes except Chr 19, the mean heterozygosity of metacentric pericentromeres was lower than those of acrocentric chromosomes (Figure 3.13 B). Chr 19 is only found in two metacentrics, both in Madieran races (which have high heterozygosity, Figure 3.8).

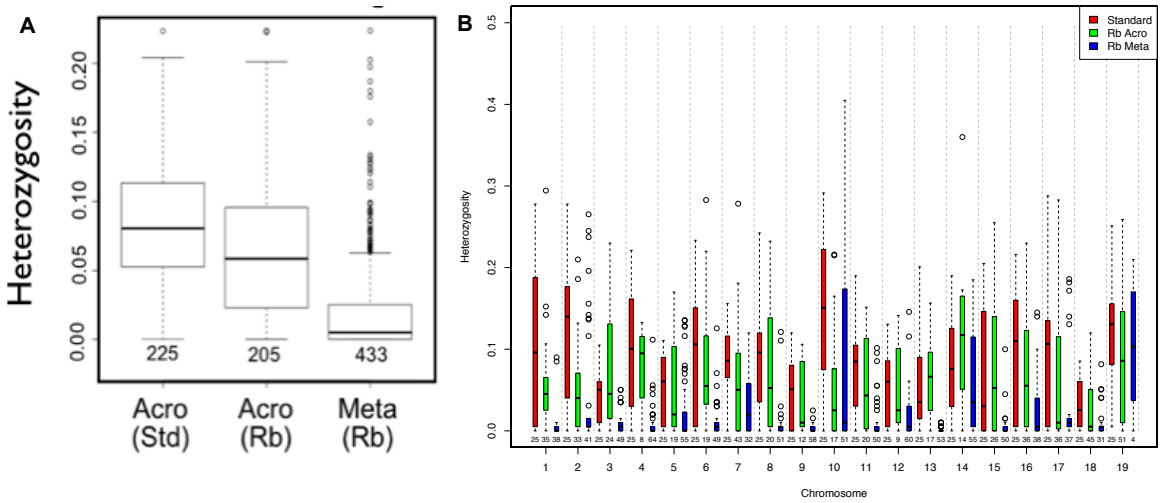


Figure 3.13: Heterozygosity is significantly different between chromosome types. Box plot of heterozygosity, measured as the fraction of markers with a heterozygous genotype, partitioned by A) chromosome type and by B) chromosome. Acro (Stnd)/red boxes: acrocentric chromosomes in standard karyotype individuals; Acro (Rb)/green boxes: acrocentric chromosomes in individuals from chromosomal races; Meta (Rb)/blue boxes: metacentric chromosomes in individuals from chromosomal races.

While reduced heterozygosity in metacentric pericentromeres indicates a greater degree of fixation of those regions within local populations, it does not necessarily mean that CRs share a common haplotype in those regions. A better measure of whether pericentric regions of metacentrics are more closely related than are other chromosomal regions is to compute allele frequencies along the chromosome. For each chromosome, I partitioned samples as described above and computed mean MAF in 100-SNP windows along the chromosome. MAFs were generally lower in metacentric compared to acrocentric chromosomes (Figure 3.14). Pericentric regions (most proximal 200 SNPs) had significantly lower MAFs in metacentric chromosomes compared to acrocentric chromosomes in both CRs and STs (0.08 *vs.* 0.12 and 0.14, p = $1.9x10^{-6}$ and $3.6x10^{-10}$, respectively). In metacentric chromosomes, pericentric regions

were only slightly more homozygous than telocentric regions (0.08 vs. 0.10, $p = 0.04$), but were not significantly more different (after correction for multiple testing) than regions chosen at random ($p = 0.003$, 100 iterations). I conclude that pericentric regions of metacentric chromosomes in CRs are less variable than pericentric regions of acrocentric chromosomes. This is consistent with a hypothesis of locally shared ancestry between CRs; however, higher resolution data (such as whole-genome sequence) will be required to adequately test that hypothesis.

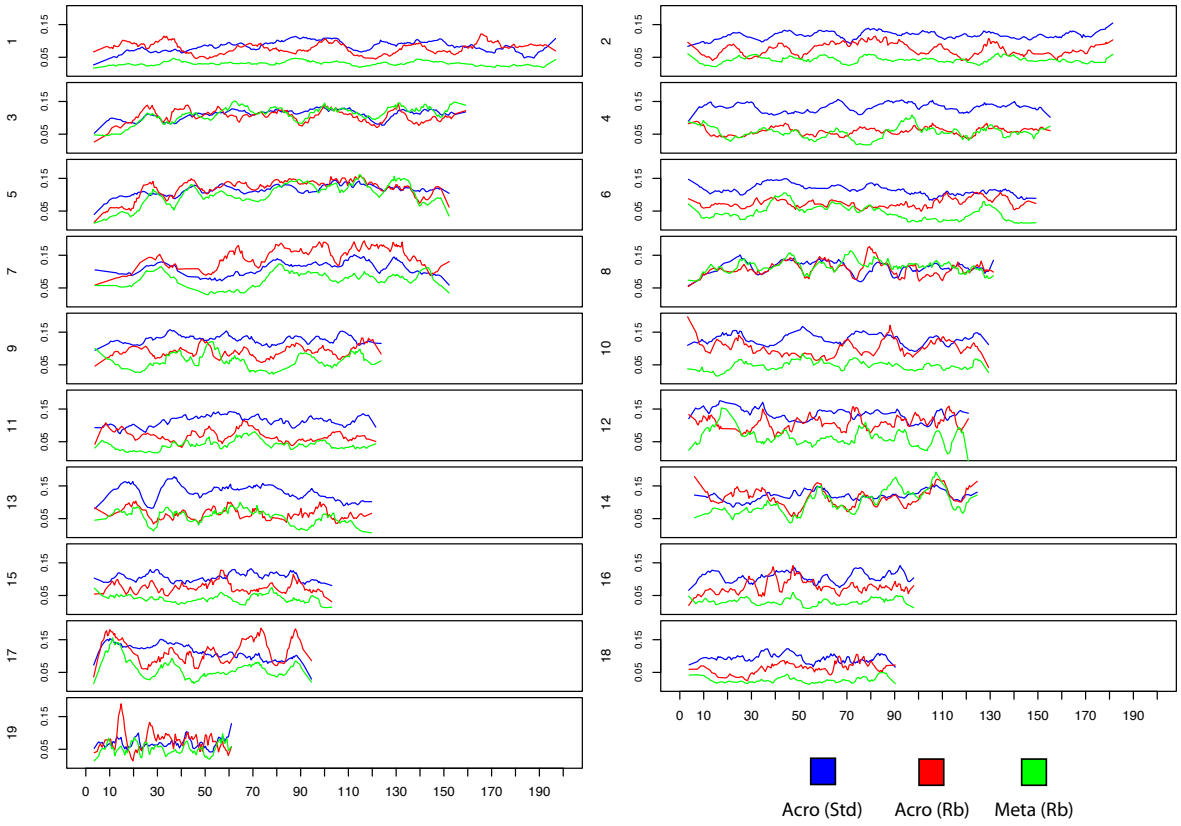


Figure 3.14: Metacentric chromosomes have reduced MAF in pericentric regions. For each chromosome (1-19), mean MAFs in 100-SNP windows are shown for ST individuals (blue), CR individuals that are acrocentric for that chromosome (red) and CR individuals for whom that chromosome is involved in a metacentric (green).

3.4.7 Chromosomal races are enriched for loci under positive selection

A classic (or "hard") selective sweep describes the process of a novel, major effect mutation arising on a single haplotype in a population and ultimately reaching fixation. The concept was later expanded to include "soft" selective sweeps in which selection acts on standing variation in the advent of a change in environment [145, 146]. Selective sweeps are observed as a reduction or absence of genetic diversity within a region of LD encompassing the mutation. The lower the recombination rate, or the more recently the sweep occurred, the larger the size of the LD block. Most authors make an implicit or explicit assumption that the mutation must have a strongly positive effect on fitness, while few have recognized that meiotic drive and other mechanisms employed by selfish genetic elements may yield identical results [147, 148, 31].

In the context of *M. m. domesticus*, a selective sweep that is present only in the chromosomal races may be associated with a *responder* or with a *distorter* locus that helps to promote fixation of metacentrics. Association studies may or may not identify loci under selection, depending on the mode of selection and the age of the sweep. To find loci under positive selection on metacentric chromosomes that may have been missed by the association study, I used the `SweepFinder` software [149] to look for evidence of selective sweeps in the CRs that were absent in STs.

Using all individuals regardless of karyotype, I found a single region on Chr 13 with a high likelihood (> 10) of having undergone a selective sweep (Figure 3.15 A). There are several other regions with moderate likelihoods (> 5), including Chr 19. In contrast, there are at least nine high-likelihood peaks when the samples are restricted to the CRs (Figure 3.15 B). There are again peaks on Chrs 13 and 19 at the same locations as in the previous scan; there are additional peaks on Chrs 1, 4, 8, 11, 14, 15 and 16. The CR-only peaks are candidates for regions that may harbor loci associated with the accumulation of metacentrics.

There is evidence that selective sweeps are common in *M. m. domesticus* [72]. Furthermore, in previous studies many of the identified sweeps are specific to certain populations,

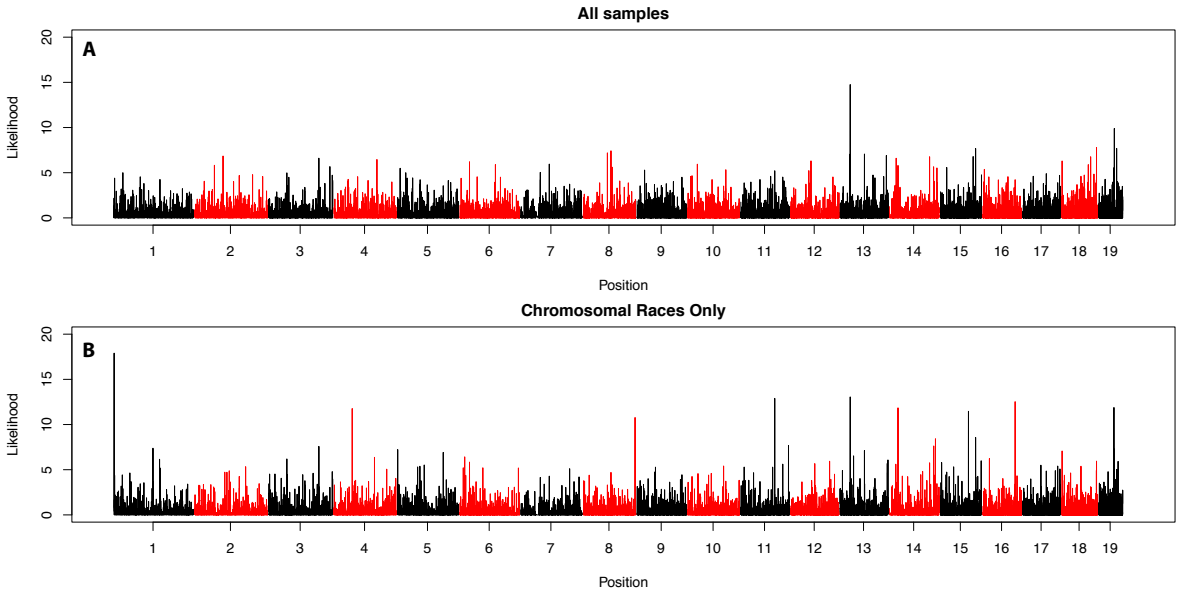


Figure 3.15: Enrichment of evidence of selective sweeps in CRs. Likelihoods are shown across the genome for the presence of a selective sweep A) in all samples and B) in CRs only. Chromosomes are given alternating colors.

and few to none are shared across all populations. Interestingly, none of the sweeps identified in our study overlapped those identified by Staubach *et al.* (2012). This suggests either that ancient sweeps cannot be identified in the mouse due to limitations of the available data or methods, or that most sweeps in the mouse are recent and occur at the level of local populations. The latter explanation is consistent with the recent range expansions of *M. m. domesticus*, which presumably would have been associated both with genetic bottlenecks and with adaptation to new environments.

3.5 Discussion

3.5.1 Are wild mice suitable for association studies?

In this work, we have created important new resources for the study of genetic variation in wild house mice. First, we have established an extensive archive of DNAs from natural mouse populations that cover the entire extent of the native range of *M. m. domesticus* as well as several *M. m. musculus*, *M. m. castaneus* and new-world populations. In conjunction with our

previous work to develop a cost-efficient genotyping array (MegaMUGA) that included a substantial number of wild-mouse specific markers, we will be able to rapidly genotype additional samples for any population requiring further study, either in our lab or with collaborators. Second, we have already genotyped a large subset of samples, including 44 difference CRs. In conjunction with the methods and tools that we have developed, that genetic data has enabled unprecedented genome-wide characterization of natural mouse populations; previously only limited sequences were available.

Although it has been suggested that wild mice could be a source for novel phenotypes [51, 136], no previous study (to our knowledge) has attempted association mapping in wild mice. Therefore, our study is a test-case for the suitability of wild mice to GWAS. First, it must be recognized that population structure exists in mice just as it does in humans. One method to avoid population structure is to obtain a homogeneous sample from a relatively local population [131, 136]. While such a strategy provides a high degree of power to detect QTL, it is limited by the fact that relatively few QTL will be segregating within a local population. A second strategy is to sample more broadly. This strategy also provides substantial power and potentially more phenotypic diversity than the first method, but the effects of cryptic relatedness are likely to be stronger. The second method is also logistically more difficult, since it requires trapping from widely separated sites. In either case, it is advisable to ascertain the structure of a sample and correct for it if necessary. We were in the fortunate position of being able to pursue a hybrid strategy, whereby we sampled broadly from many different CRs, and deeply within a few CRs. We found that population structure was likely to have a strong effect, and we thus took subsamples of independent populations in order to obtain a data set in which individuals were uniformly related. Another potential strategy that we did not pursue was an association study using only a highly related population that is segregating for our phenotype of interest, such as the northern Italian system of CRs and nearby ST populations. We believe that the flexibility for multiple approaches to sample selection make the WMGS a powerful resource that should be further expanded, both in breadth and depth.

Second, wild mouse genomes are not uniformly heterozygous, and, at a population level, have less heterozygosity than is expected from allele frequencies. This latter fact is probably more reflective of mouse social behavior than of any selection against heterozygosity [150]. Heterozygosity is also non-uniform at a population level. Some CRs have a higher level of heterozygosity than the average ST population, while others are nearly inbred. The most substantial difference in our study was between the CRs of northern Italy and Tunisia, which generally had low heterozygosity (mean 0.036), and the other CRs (0.091). TUMO appears to be almost fully inbred, which is consistent with previous observations [151]. The high level of heterozygosity in the Madeiran CRs was somewhat unexpected because typically there are bottlenecks associated with island colonization that result in low genetic diversity in island populations. However, our results were consistent with microsatellite data [144], and previous studies have observed that other Mediterranean island mouse populations also show no reduction in genetic variation relative to mainland populations [152]. Heterozygosity has substantial implications for both potential phenotypic diversity and statistical power, and so it is important for an investigator to select a mapping population that is well matched to the design and goals of their study. We believe the WMGS is a powerful tools that will help investigators to make such choices.

Third, wild mice offer great potential for fine-scale mapping. Compared to inbred strains and mouse genetic reference populations [39], outbred mice [134] and humans [136], a broad sampling of wild mice such as the one we used in our study exhibits extremely low LD. For example, commercial outbred populations have LD decay radii (defined as the distance beyond which r^2 falls below 0.5) of 0.5 – 4 Mb [134], compared to 50 kb (0.05 Mb) in wild *M. m. domesticus* mice. This means that wild mice essentially offer single-gene mapping resolution.

In summary, wild mice offer great potential for identifying the genetic basis of phenotypes. However, such studies will generally require greater effort in planning, phenotyping and data analysis than do studies that use commercial laboratory populations. We hope that resources

developed in this study will help to lessen the burden and make wild mice more accessible an experimental system.

3.5.2 Genetic variants associated with Rb fixation

Our study is among the first to attempt a GWAS in a natural, non-human population. Although there are undoubtedly methodological improvements we can make in the second stage of the GWAS, the fact that we were able to identify a significant association from a relatively small sample size lends support to our approach. Equally promising is the identification of several loci that appear to be under positive selection only in the CRs. Thus far our investigations of the significant associations have been cursory and have not yielded promising candidate genes. However, there is no requirement that the *distorter* be a coding variant; a structural variant or regulatory variant (such as a long non-coding RNA) are equally likely. We recognize the challenge that this presents, and we have whole-genome sequenced mice from two different CRs to facilitate identification of *distorter* loci.

The marker under the significant hit on Chr 13 was the first SNP adjacent to the centromeric region on the array. In addition, one suggestive and one non-negligible association were also within 1 Mb of the centromere of their respective chromosomes (5 and 16). These associations are attractive candidates for the *distorter*, which may benefit from close association with a centromere. A cursory search for likely candidate genes under these peaks yielded few results. The pericentric region of Chr 13 is gene poor. The single gene within the first 500kb of the beginning of the sequence (all mouse autosome sequences begin at 3Mb to reflect that the centromeric sequences are uncharacterized) was *Speer6-ps1*, an uncharacterized, testes-specific protein. The single pericentric gene on Chr 5 was *Vmn1r79*, an olfactory receptor gene. Possible candidates on Chr 16 included *Slx4*, a subunit of an endonuclease involved in DNA repair and recombination.

We note that there were several reasons why our use of a proxy phenotype might have generated false associations: 1) Because we expected the causative allele(s) to be present

at low frequency in the general population, by chance we may have genotyped individuals possessing a causative allele that have 2N = 40; 2) We may have genotyped hybrids that carried a causative allele but no Rb translocations, or carried Rb translocations but also the wild-type allele; or 3) we incorrectly phenotyped individuals. A power calculation showed that the effect of phenotypic classification would be unlikely to significantly affect the outcome of our GWAS. Our sample selection procedure minimized the chances of incorrect phenotypes. We were conservative in removing related samples and in our choice of significance threshold, so it was unlikely that the identified associations were spurious due to unidentified populations structure.

3.5.3 What mechanisms may enable the fixation of metacentric karyotypes?

There is little doubt that Rb translocations play the primary role in the formation of chromosomal variants in the CRs. Rb translocations are common in mice and humans, as both *de novo* and inherited variants [153]. Rb translocations are thought to play a primary role in karyotype evolution of some taxa, including mammals [154]. Rb translocations arise from breakage of two chromosomes within their minor satellite region followed by fusion, presumably during homologous non-allelic recombination, although environmental mutagens cannot be ruled out as causal in at least some cases [124, 110]. Rb translocation formation is accompanied by the loss of several hundred kb of minor satellite DNA [155]. This may contribute to metacentric fixation since fissions of those chromosomes would yield acrocentric chromosomes with greatly reduced fitness, and minor satellite DNA is important for centromere function [156].

The role of WARTs must also be recognized in generating new fusion combinations. Although there are only a small number of cases in which WARTs have been observed directly, several studies have constructed phylogenies of the CRs using chromosomes as the characters and found that the inclusion of both Rb translocations and WARTs resulted in more parsimonious trees [103, 111, 91]. Most of the WARTs reported so far have involved the exchange

of chromosome arms between two metacentrics, although logic suggests that the simpler exchange between a metacentric and an acrocentric must occur at least as frequently.

Four factors could influence the probability of fixation of a new chromosomal variant: genetic drift, selective advantage of homozygotes for newly arisen rearrangements, inbreeding and meiotic drive [157, 154, 91]. Mathematical models suggest that drift alone is insufficient to explain the abundance of CRs [157]. A selective advantage of homozygous metacentrics that is strong enough to drive some populations to fixation should also be strong enough to drive the entirety of *M. m. domesticus* to fixation. It may be that the reason we do not see more widespread karyotypic change is due to the relatively recent origin of the chromosomal races, however in at least two cases it has been shown that a race can colonize a new area very quickly [110, 158]. Particularly important is the case of the Isle of May. Over a four period after 77 UKED mice from Eday were released onto the Isle of May, which previously had only ST mice, the metacentric chromosomes spread throughout the population and reached a stable equilibrium frequency [158]. If metacentrics were truly superior, they should have gone to complete fixation. It is worth noting that there is limited evidence of differences in mate choice and aggression between standard and CR mice [19,34], so phenotypic differences may play some role in the spread of the CRs. As for inbreeding, our data show no evidence that CRs are more inbred than STs. Heterozygosity and MAF were only significantly different between CRs and STs within the regions immediately proximal to centromeres (Figure 3.13, Figure 3.14). A previous study examined inbreeding with a set of CRs and STs in Belgium and found no evidence of inbreeding [159]. Finally, meiotic drive has been shown to have the strongest potential to fix new chromosomal variants and will be discussed separately below.

Although it has also been suggested that an elevated mutation rate in *M. m. domesticus* may explain the large number of chromosomal races, there is no direct evidence to support that theory. The expectation of that theory is that Rb translocations should be observed in a substantial fraction of individuals from both natural populations of mice and wild-derived strains, however neither has been reported. Instead, all evidence points to the prevalence of

Rb translocations being specific to the chromosomal races.

There is evidence that certain chromosomes (or certain fusion pairs) are selected for more strongly than others. First, the distribution of fusion pairs is non-random (Figure 3.2). Second, the level of selection against metacentrics is variable in experimental crosses between wild-caught mice and laboratory strains [95, 160]. Castiglia & Capanna [102] studied the relationship between chromosome length and transmission rate of metacentrics in a hybrid zone between a CR and ST. They found that large metacentrics have a lower transmission rate than small metacentrics. This finding is consistent with the trend toward intermediate-sized metacentrics in CRs (Figure 3.16), although the mechanism by which larger chromosomes might experience a reduced segregation advantage is unclear. The non-random distribution of fusion pairs may also be driven by differences in zonal raciation following initial founder events; however, certain fusions seem to be recurrent across different systems of the CRs. This may reflect greater homology between centromeres of some pairs of chromosomes than others, or it may indicate that certain chromosomes are more prone to fission than others. In the latter case, breakage in pericentric regions with housekeeping or other dosage-dependent genes near to the centromere appears to be selected against [117].

3.5.4 The centromeric drive model of chromosomal race evolution

As discussed in chapter 1, centromeric drive is a specific type of meiotic drive in which certain types of centromeres are selectively favored over others. The centromeric drive theory was independently proposed by two groups at about the same time. The two groups described parsimonious models, but focused on different evidence. Henikoff *et al.* [161] invoked centromeric drive to explain the observation that both centromeric sequences and centromeric proteins evolve rapidly even though most mechanisms related to chromosomal segregation are highly conserved. On the other hand, Pardo-Manuel de Villena and Sapienza [16] invoked centromeric drive to explain the observations of selection acting on centromere number, such as for B chromosomes in plants, insects and rodents, Rb translocations in mice and humans,

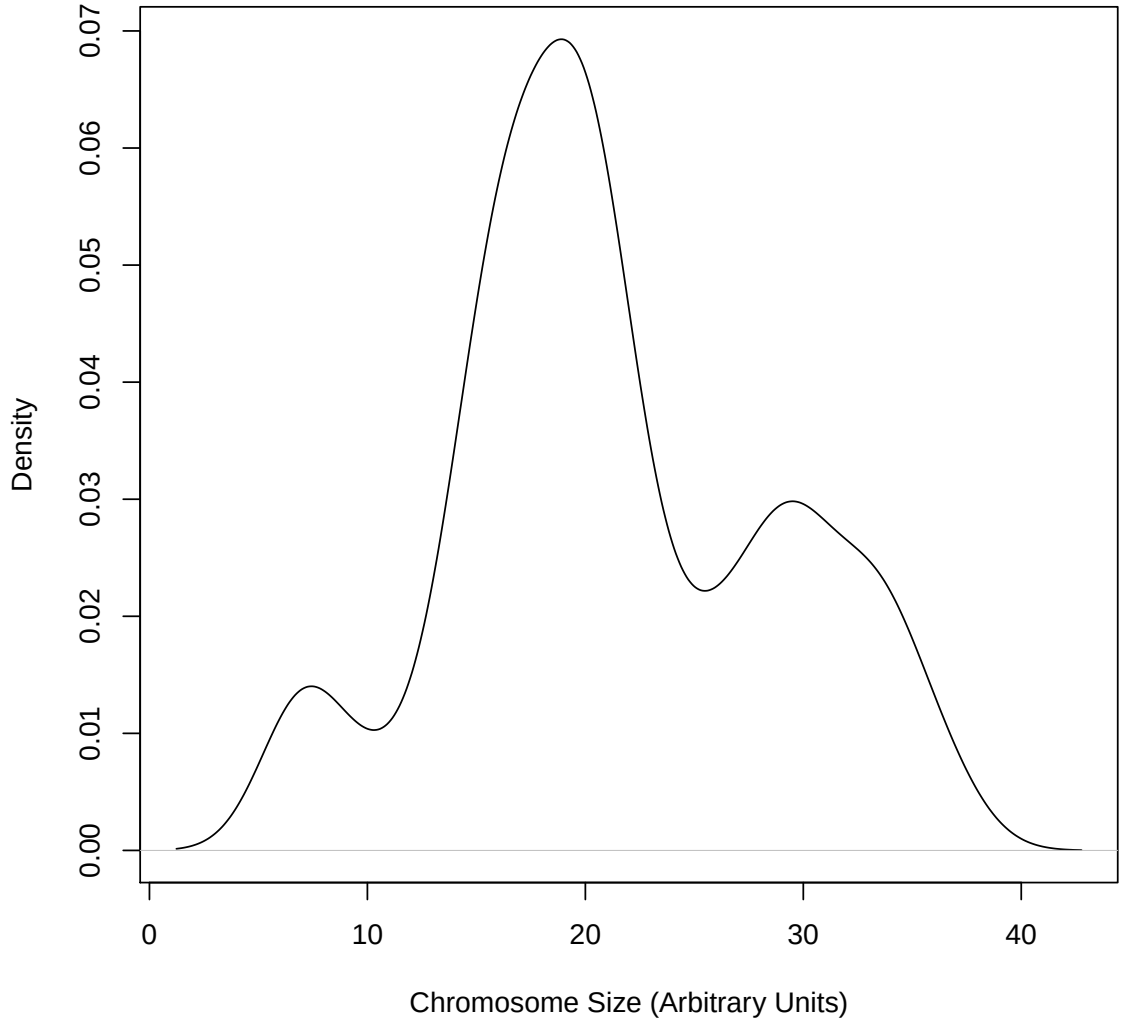


Figure 3.16: The majority of metacentrics are of intermediate size. Metacentric size is estimated as the sum of the indices of the two fused chromosomes, where index ranges from 1 (smallest chromosome) to 19 (largest chromosome); therefore the largest possible length is 37.

and fission products in chickens. They proposed that the two poles of the meiotic spindle could have different preferences for greater or lesser centromere number. Furthermore, they showed that mammalian karyotypes are highly biased toward extreme compositions (all-acrocentric or all-metacentric), suggesting a non-random pattern in karyotype evolution that could best be explained by centromeric drive that changes direction (i.e., preference for chromosome morphology) frequently over evolutionary time.

Importantly, the centromeric drive model makes the same prediction about genetic diversity in the CRs regardless of their origin: a centromeric drive locus will exist in a region of reduced gene flow such that it is preserved despite introgression. A previous study showed that gene flow is reduced in the pericentric regions of hybrids between adjacent CRs [114]. Our heterozygosity data suggests that the pericentric regions of metacentric chromosomes has a greater level of fixation than acrocentric chromosomes in either CRs or STs, however our data on nucleotide diversity showed that diversity was only reduced extremely close to the centromere. This situation may result from pericentric regions that are segregating among CRs, but that go to fixation within a race once they become linked to a driving centromere.

The fact that we identified several loci having a high degree of association with the accumulation of Rb translocations lends support to our choice of method. The fact that several of those markers were tightly linked to centromeres suggests candidate regions to examine for structural or regulatory variants that may play a role in the interaction between the centromere and the meiotic spindle, or between the spindle and the cellular cortex. Alternatively, the wild-type allele may be a suppressor of drive in favor of the metacentric that is ancestral to *M. musculus* but has a mutant allele in *M. m. domesticus* and is present at low to intermediate frequency. Future work will be to test evolutionary models under these two scenarios to see which is more likely.

3.6 Materials and Methods

3.6.1 Genotyping

Samples were selected for genotyping in periodic batches. The genotyping protocols are given in Appendix A. I used MouseDivGeno [40, 41] to call genotypes and VINOs for approximately 1000 samples that had been genotyped at the UNC Genotyping Core, including all of the samples selected for this study. The use of such a large reference sample set was instrumental to insuring accurate genotype calls.

Genotypes for MUGA and MegaMUGA were called by staff at Neogen using the Illumina `BeadStudio` software. We received the exported data, which included genotype calls and intensity values. The UNC Computational Genetics group has developed a pipeline for importing all results into a database. The have also developed a suite of tools for visualizing and manipulating the data. All of my analyses used a combination of those tools and programs I wrote that directly accessed the database. I performed the standard QC procedures (described in Appendix A) on all samples.

Our final set of MDA arrays consisted of 103 *M. m. domesticus* and 24 *M. m. musculus* unique samples. Our final set of MUGA arrays consisted of 160 *M. m. domesticus* and 21 *M. m. musculus* unique samples. As of this writing, we have genotyped 199 *M. m. domesticus* and 9 *M. m. musculus* unique samples on MegaMUGA. In total, there were 425 unique, high-quality arrays across all three platforms (385 *M. m. domesticus*, 30 *M. m. musculus* and 10 *M. m. castaneus*).

In addition to the genotypes available from [40], we selected 11 *M. m. musculus* samples from three additional populations (Poland, Russia, Ukraine). These, along with previously genotyped *M. m. musculus* and *M. m. castaneus* mice [40] were intended to serve as outgroups in phylogenetic analyses.

There were 1,845 autosomal markers that were common across all platforms. I excluded Chr X since different expectations of zygosity in males vs. females complicate analyses of that

chromosome. I also excluded Chr Y and mitochondrial (mtDNA) markers since so few were overlapping between the three data sets. I filtered the markers such that only those segregating within *M. m. domesticus* were retained. There were 62 arrays that were technical replicates on multiple platform. I reduced the combined marker set further by excluding any markers that were not consistent across replicates. The final combined data set consisted of 1,221 markers for 425 unique samples.

We are currently developing methods for imputing genotypes between the three different platforms, but we expect that imputed genotypes will have a high error rate due to the relatively small number of haplotypes we have sampled, in comparison to the large effective population size of the mouse [162].

3.7 Future Directions

3.7.1 Stage-two GWAS

The stage-one GWAS was done using 64 samples, with mice from 26 different CRs. Since that time, we have genotyped 294 additional individuals, and we have sampled from an additional 18 CRs. Shortly, we expect to have a combined data set of ~ 500 *M. m. domesticus* mice for use in stage two of our GWAS.

As described in the previous chapter we developed the MegaMUGA array in part to enable better characterization in wild mice. Importantly, I contributed SNPs for the 60 hits with the highest PPAs from my association study, including redundant markers for all significant and suggestive hits. A substantial portion of markers on MegaMUGA were drawn from SNPs discovered by whole-genome sequencing, many of which were not available when the MDA was designed, and so the marker profiles of the two platforms will be substantially different. This may result in additional significant associations. We are currently developing new methods to exploit the capabilities of wild mouse studies, and extending the VINOs method previously developed for MDA.

The stage-two GWAS will implement several important improvements over stage one.

First, covariates for relatedness and sex will be considered. Second, we will consider epistatic (i.e., two-locus) interactions. Conditional search [163] is an attractive approach because it will discover SNPs that are associated only through epistatic interaction. We will calculate measures describing the distribution of basic case/control association results, e.g. overall genomic inflation factor [164] and whether the top associated SNPs show marked deviations for frequency and genotyping tests. We will also "sanity-check" our results using a more traditional frequentist method.

Our preliminary study focused on SNP associations, and we will expand to also consider VINOs. We have previously had success in treating VINOs as synthetic, binary markers (presence or absence). Alternatively, `BIM-BAM` has the ability to operate on continuous (intensity) values for markers, rather than discrete genotypes.

3.7.2 Characterization of candidate loci

Our strategy will differ depending on the type of locus. Coding and regulatory variants can be characterized using established databases and bioinformatic tools to determine where the mutation occurs (exon, intron or regulatory region), the effect of the mutation (e.g. synonymous, missense or nonsense substitution, binding enhancement or suppression) and implications on protein structure and function (if known). For non-coding loci such as RNA, we will attempt to determine the structural implication of the mutation using RNA structure prediction software. Next, we will conduct a search for homologues using a comparative genome analysis tool such as the UCSC genome browser [165]. We will determine if equivalent mutations exist in other species and whether those species have been tested for centromeric drive.

Pericentric associations present a challenge because the locus of interest may lie within the uncharacterized centromeric regions. Identification of such loci can be accomplished by BAC fishing [166] or chromosome-walking backward from known sequence [167] until a gene-like sequence is discovered.

The available data on our candidate loci are likely to be highly variable in terms of com-

pleteness and accuracy. Some of these loci may be novel in the mouse, or may have acquired a different function since diverging from other mammals. We expect our study to be a stepping-stone to molecular characterization of the centromeric drive locus, and thus our results are intended to guide further efforts rather than be exhaustive or definitively conclusive. We may not be able to identify appropriate laboratory models for experimental validation, in which case we will propose the most appropriate wild populations that should be sampled to develop new models.

3.7.3 Sequence analysis of wild mice

While genotyping arrays are a cost-effective method of screening large numbers of samples for genotype-phenotype associations, the probability that a SNP included on the array will prove to be the causal variant is low. The centromeric drive causal variant may also be harbored within a structural variant, or the causal variant may be structural rather than a single-base mutation or small indel; structural variants are nearly impossible to detect *de novo* using microarrays. To facilitate discovery of causal variants and other projects, we have used Illumina technology to sequence the genomes of two wild-caught and two wild-derived mice (Figure 3.17). Among other criteria, we selected the two CR samples sample most likely to be carrying the causal *distorter* allele(s). We identified one such wild-caught specimen (ES446) that was trapped in Tarragona, Spain, karyotyped as canonical EBAR (2N=32), and was homozygous for the Rb-associated allele in 12/18 candidate SNPs identified in our preliminary GWAS and was heterozygous at another 5 SNPs. The second CR mouse was a wild-derived inbred strain, ZALENDE/EiJ. We wanted to characterize an inbred strain with Rb translocations, since that could be an important laboratory model for the study of candidate genes. The two ST mice selected were a wild mouse from northern Italy (near to the original trapping site of the mice from which ZALENDE/EiJ was derived) and LEWES/EiJ, an inbred strain derived from mice originally trapped in Lewes, Delaware. The combination of mice we selected will allow us to make a number of informative pairwise comparisons.

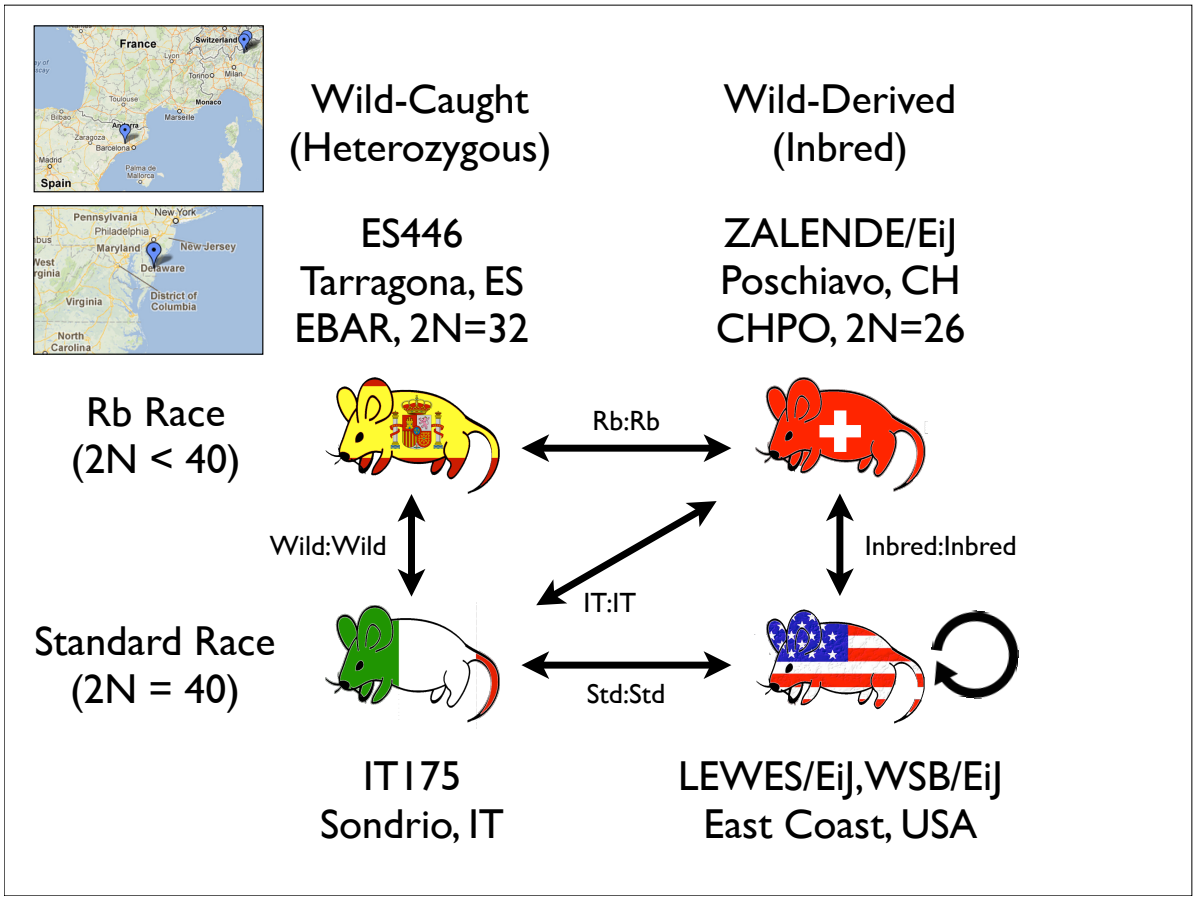


Figure 3.17: Wild mice selected for whole-genome sequencing.

The traditional approach to whole-genome sequence assembly is to align reads to a reference genome. We and others have shown that the laboratory mouse genome is polyphyletic [50, 55, 71, 40], and thus aligning to a single reference sequence may not be the best approach. This is especially true for wild-caught mice whose genome structure is quite different from that of the standard reference strain, C57BL/6J. Using 17 Sanger genomes, we have constructed pseudo-genomes that are better suited for the assembly of our wild mouse samples. A pseudo-genome is created by stitching together pieces of the previously assembled genomes based on local levels of identity between MDA and Sanger genotypes. We only used *M. m. domesticus*-diagnostic SNPs to test for identity.

While we expect this assembly to be very accurate over most of the genome, structural differences (e.g. CNV and inversion) between wild mice and the reference will present an

important challenge. To address this challenge, we will perform supervised local realignment [168] in our candidate intervals. Because many of our candidate intervals encompass the extreme proximal ends of accessible chromosome sequences, we will also attempt to extend our sequence as far as possible into the pericentric regions using guided assembly. A disadvantage of reference-guided assembly is that large differences between the reference and the target are often inaccessible. For example, segments that are duplicated in the target relative to the reference are aligned to a single place in the reference. These large differences are of particular interest to us because it is possible that one or more will be associated with centromeric drive. *De novo* assembly may enable us to access these sequences because it does not depend on a reference. *De novo* assembly has traditionally been time- and resource-prohibitive, however recent advances have simplified the process [169, 170]. We will attempt *de novo* assembly and look for deviations from the reference-guided assembly to identify potential structural variants that should be investigated further. We will also attempt to identify translocation breakpoints in the fused chromosomes as these may provide valuable information about sequences associated with chromosome breakage [171].

When alignment is completed, some portion of reads remain that cannot be mapped. There are three classes of unmapped reads: (1) Short contigs are sets of overlapping reads that can be aligned into a longer sequence but cannot be aligned to the reference genome; (2) Non-unique reads typically consist of tandem repeats or commonly occurring sequences (e.g. transposable elements) that map with equal likelihood to multiple places in the genome; (3) Unmappable reads cannot be confidently aligned to anything else; these will include low-quality reads, mtDNA and sequencing artifacts (primer or vector sequences). Recent sequencing of humans [172, 173] showed that several additional Mb of category 1 sequence could be identified using *de novo* assembly, though most mapped to some known sequence. A small set of category 3 reads from our samples will encompass a chromosomal breakpoint of a Rb translocation. These may be mappable using the brute-force approach of generating all possible partitions (e.g. there are 99 possible partitions of a 100 bp read) and then aligning each end separately.

Category 2 reads are of particular interest because they will include pericentric satellite repeats. We will perform statistical analyses of these sequences (e.g. nucleotide frequencies) and compare them to known repeats using `RepBase` [174]. We will compare unmapped reads in our sequences and the all-acrocentric laboratory mice to determine if any sequences are preferentially lost from Rb translocations.

Chapter 4

GENETIC CHARACTERIZATION OF A NOVEL MEIOTIC DRIVE SYSTEM IN THE MOUSE[1]

4.1 Genetic Reference Populations

Genetic reference populations (GRPs) are sets of individuals with fixed and known genomes that can be replicated indefinitely [39]. GRPs are created from two or more genetic backgrounds (founders), typically well-characterized inbred strains. The founders are randomly bred to generate dozens to hundreds of unique genomes, which are then inbred to create independent lines. GRPs are popular for the study of complex traits and biological systems in both medical and life science applications because their genomes are reproducible, allowing for optimal case/control and gene-by-environment designs. GRPs are expensive to produce, but they have increasing returns as the community of users becomes larger because the phenotypic, genetic, and genomic data associated with each line becomes richer. This makes it possible to integrate data from distinct biological fields (system genetics). The first generation of GRPs were created from crosses between pairs of inbred strains [175, 176]. Genetic and phenotypic diversity are enhanced in the current generation of GRPs due to inclusion of mul-

[1]The work described in this chapter was accomplished in collaboration with the Churchill lab, and with the help of many individuals who provided mice or genotype data. Gary Churchill helped to provide the vision for the study. Petko Petkov and Tim Bell conducted most of the PCR genotyping. Andrew Morgan and Dan Gatti performed QTL mapping. Biological specimens were provided by Daniel Pomp, Jim Crowley, Lucy Rowe and David Threadgill. Data sets were provided by: David Aylor, Elissa Chesler, Nigel Crawford, Jef French, Allison Harrill, Kent Hunter, Yi Liu, Debbie O'Brien and Karen Svenson. The projects described under future work are being done in collaboration with David Threadgill, Michael Lampson, Amanda Chunco and Island Conservation. A manuscript on this work has been submitted to PLoS Biology. In addition, we published the initial reports of TRD in Aylor *et. al.* 2011 [38] and Collaborative Cross Consortium 2012 [39].

tiple strain and subspecific backgrounds [70, 177, 39]. Advanced intercross lines and outbred stocks have been created as companion resources to GRPs to enable fine-mapping of QTLs [178, 179, 180].

4.2 The Collaborative Cross

The Collaborative Cross (CC) is a GRP established from eight laboratory inbred strains [181, 182, 39] (Figure 4.1). Five of the strains – A/J, C57BL/6J (B6), 129S1/SvImJ, NOD/ShiLtJ and NZO/HlLtJ – are classical inbred lines. The other three strains are wild-derived: WSB/EiJ was derived from mice trapped in Centreville, MD (USA); PWK/PhJ was derived from mice trapped near Prague, Czech Republic; and CAST/EiJ was derived from mice trapped near Bangkok, Thailand. A comparison of these three strains to wild mice sampled from the native range of each subspecies revealed that WSB/EiJ is a "pure" representative of *M. m. domesticus*; PWK/PhJ and CAST/EiJ are primarily of *M. m. musculus* and *M. m. castaneus* origins, respectively, but also contain introgressions from *M. m. domesticus* [40]. The eight founder strains capture a much greater level of genetic diversity than existing existing resources.

The eight founder strains were crossed together using a funnel breeding scheme, in which a funnel is determined by the initial ordering of lines (Figure 4.1). Approximately 2,000 funnels were initiated in 2004 at three different breeding facilities (Oak Ridge National Laboratories, Tennessee, USA; International Livestock Research Institute, Kenya, later relocated to Tel-Aviv University, Israel; Geniad Ltd., Australia). In each funnel, the genome is randomized by three generations of outbreeding followed by repeated generations of inbreeding by brother-sister mating. In the first generation of each funnel (G1), females of strains at the odd-numbered positions are mated to males at the subsequent even-numbered position (i.e., 1F x 2M, 3F x 4M, etc). G2 animals are created by crossing the first two and the second two G1s. Finally, a female of the first G2 is bred to a male of the second G2 to create the G2F1 generation. Inbreeding begins in the G2F1 generation and continues until the line is

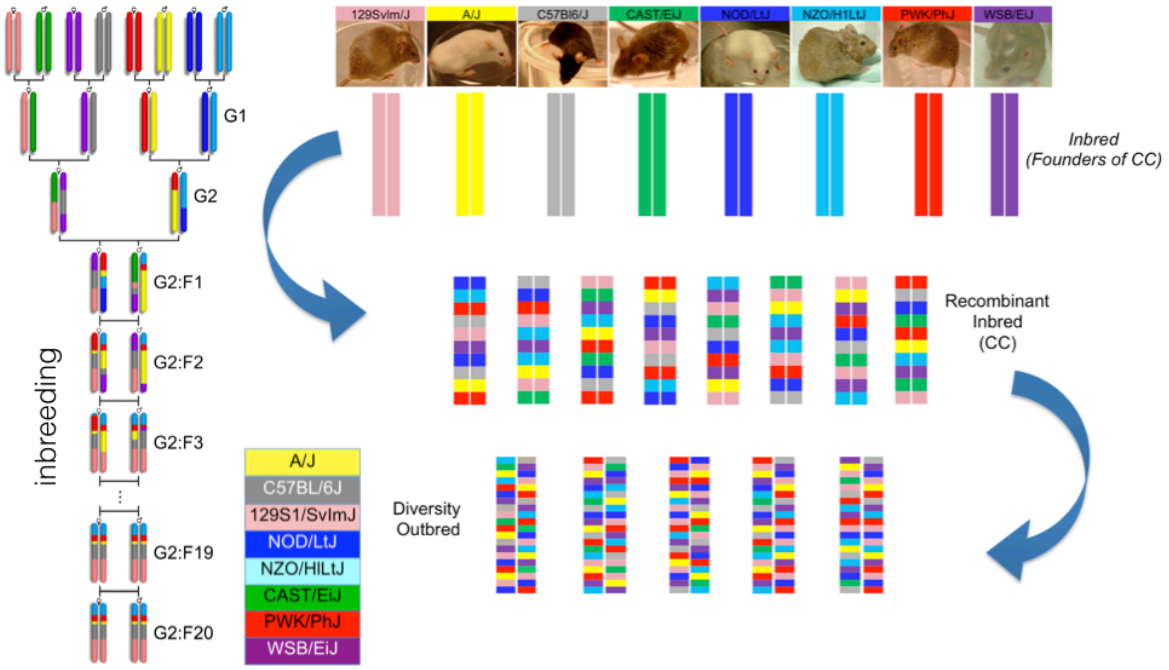


Figure 4.1: The Collaborative Cross and Diversity Outbred. Each founder strain has a consistent letter code and color (legend). A) Schematic of funnel breeding scheme. B) The genomes of the CC lines are unique, but each individual within a line has an identical genome to all others. The DO is an outbred population, and so each individual has a unique genome. The genomes of CC and DO individuals may be completely reconstructed from founder haplotypes. Courtesy of Karen Svenson.

homozygous in at least 95% of the genome. At that point, the lines are re-derived from cryopreserved embryos in a clean facility and are made available to the research community. Each CC line represents a unique genome with approximately equal contributions from the eight founder strains (12.5% from each). Currently, there are 58 finished lines available for use by researchers. A large proportion of the initiated lines became "extinct" due to infertility. It is predicted that the high rate of infertility is due to Bateson-Dobzhansky-Muller incompatibilities between the lines of different subspecies [39], as has been shown in other intersubspecific crosses [183].

Extensive simulations, statistical models and computational tools have been developed to maximize the utility of this resource [184, 185, 186, 187, 188, 83].

4.3 The Diversity Outbred

The goal of the CC was to create a panel of reproducible laboratory strains with a high degree of genetic and phenotypic variation. However, inbreeding also has the effect of fixing a single haplotype for the majority of the genome within as few as 20 generations [83]. This means that the ability of recombination to create new allelic combinations quickly diminishes. On average, CC lines were estimated to have 92 recombination segments, yielding a mean haplotype block size of $\sim 28Mb$, although those estimates may be 30-50% lower than actual values due to the resolution of the MUGA array [39]. That resolution is sufficient for several types of studies, but not for identifying causative variants. For that reason, a complementary population called the Diversity Outbred (DO) was created at the Jackson Laboratory [180]. Individuals from 144 different CC lines in the early stages of inbreeding (generations 4-12) were bred randomly to create the first DO generation (G1). Breeding was randomized in subsequent generations to avoid sib mating. This planned randomization minimizes the effects of both genetic drift and positive selection. As of this writing, DO G16 is in production. DO generations are synchronized, and there are roughly four per year. As of DO G10, there were roughly twice as many informative recombinations per individual than in the CC [180]. The number of informative recombinations is expected to increase, and thus the average haplotype block size is expected to decrease, linearly with increasing generations. The relentless increase in recombination density was one reason for the development of the more dense MegaMUGA array.

4.4 Introduction to the study

Deviations from the expected Mendelian allele frequencies are commonly observed in GRPs [36, 21, 20, 38, 39]. For example, despite the expectation of equal contributions from each of the eight founder strains to each CC line, CC lines are in fact highly variable in the fractions of their genomes that are attributable to each founder [39]. In nearly all cases,

however, those deviations are not significant or reproducible, but instead are due to breeding errors or random effects. The small fraction of observations that are both reproducible and significant are classified as TRD.

We previously reported significant overrepresentation of one of the eight possible alleles of the CC mouse recombinant inbred panel across a locus that spanned tens of Mbs on Chr 2 [38, 39]. We first reported TRD for loci on Chrs 2 and 15 in an analysis of partially inbred CC lines [38]. Subsequent studies on larger and mostly orthogonal sets of CC lines reproduced the finding of TRD across a large region of Chr 2 (73.25–124.85 Mb) [189, 39]. The WSB/EiJ allele appeared at a frequency of ~ 0.22, almost twice the expected level of 0.125. In the largest study, the finding held across all three breeding populations [39]. We concluded that TRD of Chr 2 in favor of WSB/EiJ was a common feature of the CC rather than a simple chance event, and that the size and shape of the distorted region suggested the involvement of multiple loci. Across the region, the frequencies of the other founder strains were reduced at a roughly equal rate. It was difficult to determine the true level of TRD or the underlying mechanism solely from the CC data. The randomization of funnel orders complicated the determination of sex-specific mechanisms; inbreeding limited the maximum observable level of distortion by fixing alleles within lines; and the high rate of extinction suggested that (perhaps multiple, independent) epistatic interactions were involved.

The overrepresented allele was inherited from Watkins Star Line B (WSB), a laboratory inbred strain that was originally derived from wild-caught house mice (*Mus musculus domesticus*) trapped near Centreville, MD in 1976. That line was selectively bred for dark agouti coat color with white head blaze. The Jackson Laboratory acquired the line in 1986 and established the commercially available WSB/EiJ inbred strain. WSB/EiJ is one of the most commonly used wild-derived inbred strains, has undergone extensive genetic and phenotypic characterization [40, 56], and has been utilized in many intra- and inter-specific intercrosses [190]; however, to our knowledge, there have not been reports of TRD in favor of a WSB/EiJ locus in any non-CC background.

The observation of TRD in favor of WSB/EiJ was important for several reasons. First, observations of TRD across multiple, largely independent populations indicated that it was a general feature of the CC rather than a random or population-specific effect [39]. Non-random chromosome segregation has implications for linkage and association mapping studies. The presence of an allele under selection may confound the identification of true associations for phenotypes unrelated to TRD. Second, if TRD was not just a feature of the CC, but instead a feature of the CC genetic background (i.e., the specific mixture of the eight CC founder strains), then the implications for the DO were severe since the WSB/EiJ allele was segregating in the whole population rather than in specific lines. A mathematical model of meiotic drive indicated that even weak selection in favor of the WSB/EiJ allele would lead to fixation over a relatively small number of generations [157]. Significantly reducing or eliminating allelic diversity across a large region of Chr 2 would have negatively impacted the value of the DO as a resource for fine-mapping. Third, observations of TRD often indicate interesting underlying biological mechanisms. In the recent past, further exploration of TRD in other model systems has uncovered novel meiotic drive systems [191, 20], and has led to a fundamental change in our understanding of the role of female meiosis in karyotype evolution [16].

TRD in Chr 2 has also been reported in two other strains. First, in interspecific backcrosses between C57BL/6J (a classical inbred strain, primarily of *M. m. domesticus* origin) and *Mus spretus* [192, 193, 194], the *M. spretus* allele was overrepresented across a ~ 140 Mb region in the middle of Chr 2 with a maximum transmission frequency of 0.66 [192]. Second, in an F2 cross between two selection lines (for low and high body weight) derived from the Hsd:ICR outbred population [195], the ICR allele was present in 59% of offspring. In both crosses, the overrepresented allele was transmitted through the mother.

We designed a study to identify the mechanism underlying the observation of TRD in the CC. We used a step-wise design in which we first devised a test that, regardless of its outcome, would eliminate a large number of possible mechanisms; subsequent tests were designed based on the results of previous ones. Finally, we concluded that meiotic drive is

required to explain the observed TRD. Following the nomenclature of referring to loci that exhibit TRD as *responder* elements, we have designated the Chr 2 locus *Responder to drive on Chr 2* (*R2d2*). We mapped *R2d2* to a 9.3 Mb recombination-cold region in the middle of Chr 2 (76.8 – 86.1 Mb) and identified an exceptionally strong candidate for the causative allele. This allele is a 34-fold copy number gain of a 127 kb DNA fragment that is present in the three strains that exhibit TRD on Chr 2 (WSB/EiJ, SPRET/EiJ and ICR). TRD at *R2d2* had a weak but significant inverse correlation with average litter size. However, analysis of the absolute number of progeny inheriting the favored allele and the levels of inferred and observed lethality demonstrated that female meiotic drive is the cause of the distortion. The strength of drive was highly reproducible in *R2d2* heterozygous females of the same genotype. *R2d2* heterozygous females exhibited one of three possible outcomes depending on their genetic background: Mendelian segregation, highly significant but moderate TRD, or complete distortion in favor of the WSB/EiJ allele. The *R2d2* meiotic drive system is unique among mammalian systems because of its extreme levels of distortion and the fact that it is under genetic control of unlinked modifier loci, making it particularly amenable to both genetic and mechanistic dissection.

4.5 Results

4.5.1 Extreme TRD in Chr 2 is present in the DO population

To test whether TRD of the WSB/EiJ allele in Chr 2 is present in the DO, I analyzed 1,175 animals from DO generation 8 (G8) that were genotyped using two compatible genotyping arrays (MUGA or MegaMUGA). I sampled the genotypes of each individual at 1 Mb intervals along Chr 2 and then computed the overall frequencies of the eight founder allele at each position. The WSB/EiJ allele was over-represented relative to the other seven founder alleles across a roughly 100 Mb region in the middle of Chr 2 (Figure 4.2). However, there was a striking difference in the level of distortion observed in the CC and the DO, with the WSB/EiJ allele frequency reaching a maximum of 0.22 in the former compared to 0.55 in the latter, due

to the further outcrossing that occurs in the DO. We conclude that TRD favoring the WSB/EiJ allele is a general feature in crosses in the CC genetic background; however, the level of TRD may vary widely depending on the breeding design of the experimental population (inbred vs. outbred).

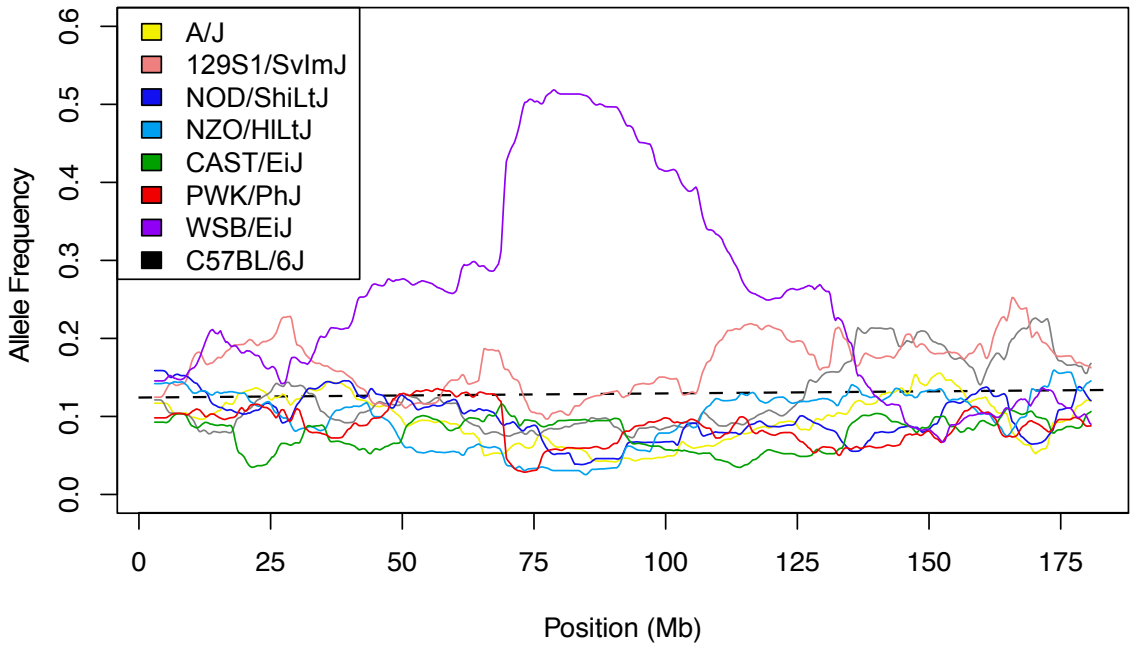


Figure 4.2: Chr 2 allele frequencies in the DO. The mean allele frequencies of the eight CC founder alleles in 1,175 G8 DO individuals are shown at 1 Mb intervals along Chr 2. Dotted line shows expected frequency of 0.125.

4.5.2 TRD is exclusive to heterozygous females

To determine the parental origin of the TRD, we analyzed 5,321 offspring from 18 experimental crosses in which exactly one parent was heterozygous for the WSB/EiJ allele in an interval spanning the region of maximum distortion on Chr 2 (75-90 Mb) [38, 189, 39]. In all cases the heterozygous parent was an F1 hybrid derived either from an intercross between the WSB/EiJ inbred strain and one of eight other inbred strains (the seven founder strains of the CC or PWD/PhJ), or from two CC lines, of which one was homozygous for the WSB/EiJ allele on Chr 2 and the other was homozygous for a non-WSB/EiJ allele. F1 hybrids were mated

to either C57BL/6J or FVB/NJ mice, and their progenies were euthanized at birth and genotyped using genetic markers located in the region of maximum distortion (see Materials and Methods). For each cross, we computed the TR of the WSB/EiJ allele and the non-WSB/EiJ allele using the aggregate genotypes across all litters from parents with identical genotypes (Table 4.1).

TR in the six paternally segregating crosses (rows 1-6 in Table 4.1) were as expected under Mendelian segregation (mean 0.502 ± 0.017, $p = 1.0$). In contrast, the mean TR in maternally segregating crosses (rows 7-18 in Table 1) was 0.667 ± 0.177 and deviated significantly from the null hypothesis ($p = 7.1\text{x}10^{-89}$). We conclude that TRD in favor of the WSB/EiJ allele on Chr 2 is restricted to the progeny of heterozygous females.

The TRs among maternally segregating crosses were significantly different ($p = 6.5\text{x}10^{-151}$), demonstrating that TRD is under genetic control. F1 female hybrids derived from crosses between WSB/EiJ and CAST/EiJ, PWD/PhJ or A/J showed no distortion (crosses 7-11 in Table 1, mean TR = 0.509 ± 0.051, $p = 0.33$). Moderate but significant distortion ($p = 2.6\text{x}10^{-16}$) was present in F1 hybrid females derived from crosses between WSB/EiJ and 129S1/SvImJ, NZO/HILtJ or NOD/ShiLtJ; and in (AU8042xCC001/Unc)F1 hybrid females (crosses 12-15, mean TR = 0.668 ± 0.026, $p = 3.5\text{x}10^{-18}$). Finally, extreme distortion was observed in reciprocal (WSB/EiJxC57BL/6J) F1 hybrid females and in (CC001/UncxOR15155)F1 hybrid females (crosses 16-18, mean TR = 0.929 ± 0.026, $p = 1.3\text{x}10^{-190}$). We conclude that heterozygosity for the WSB/EiJ allele in the central region of Chr 2 is necessary but not sufficient to observe TRD, because TR was consistent with Mendelian inheritance in some females that met that criterion.

We also conclude that the grandparental origin of the WSB/EiJ allele has no influence on TRD because the TR levels were not significantly different between three pairs of reciprocal F1 dams (compare crosses 7 and 8, 9 and 10 and 17 and 18 in Table 1, $p = 0.53$, 0.11 and 0.59, respectively).

Table 1. Segregation ratios in the progeny of *R2d2*$^{WSB/other}$ heterozygous F1 hybrid sires and dams

Cross	Dam	Sire	Informative parent	*R2d2*^{WSB}	*R2d2*^{notWSB}	TR	*p*
1	C57BL/6J	(WSB/EiJxC57BL/6J)F1	sire	132	136	0.493	$8.1x10^{-01}$
2	C57BL/6J	(C57BL/6JxWSB/EiJ)F1	sire	139	128	0.521	$5.0x10^{-01}$
3	FVB/NJ	(PWK/PhJxWSB/EiJ)F1	sire	263	283	0.482	$3.9x10^{-01}$
4	FVB/NJ	(WSB/EiJxPWK/PhJ)F1	sire	188	171	0.524	$3.7x10^{-01}$
5	FVB/NJ	(CAST/EiJxWSB/EiJ)F1	sire	110	112	0.496	$8.9x10^{-01}$
6	FVB/NJ	(WSB/EiJxCAST/EiJ)F1	sire	98	99	0.498	$9.4x10^{-01}$
7	(WSB/EiJ/CAST/EiJ)F1	C57BL/6J	dam	257	274	0.484	$4.6x10^{-01}$
8	(CAST/EiJxWSB/EiJ)F1	C57BL/6J	dam	248	288	0.463	$8.4x10^{-02}$
9	(PWD/PhJxWSB/EiJ)F1	C57BL/6J	dam	127	142	0.472	$3.6x10^{-01}$
10	(WSB/EiJxPWD/PhJ)F1	C57BL/6J	dam	146	122	0.545	$1.4x10^{-01}$
11	(A/JxWSB/EiJ)F1	FVB/NJ	dam	30	22	0.580	$2.7x10^{-01}$
12	(NODShiLtJ/JxWSB/EiJ)F1	FVB/NJ	dam	106	61	0.640	**$5.0x10^{-04}$**
13	(129S1/SvImJxWSB/EiJ)F1	FVB/NJ	dam	136	73	0.650	**$1.3x10^{-05}$**
14	(AU8042xCC001/Unc)F1	FVB/NJ	dam	85	38	0.690	**$2.3x10^{-05}$**
15	(NZO/HlLtJxWSB/EiJ)F1	FVB/NJ	dam	130	59	0.690	**$2.4x10^{-07}$**
16	(CC001/UncxOR15155)F1	FVB/NJ	dam	35	4	0.900	**$6.9x10^{-07}$**
17	(WSB/EiJxC57BL/6J)F1	C57BL/6J	dam	506	33	0.939	**$2.9x10^{-92}$**
18	(C57BL/6JxWSB/EiJ)F1	C57BL/6J	dam	512	28	0.948	**$2.4x10^{-96}$**
Subtotal			sire	930	929	0.500	1.0
Subtotal			dam	2,318	1,144	0.670	**$7.1x10^{-89}$**

4.5.3 *R2d2* maps to a 9.3 Mb interval in the middle of mouse Chr 2

To define the boundaries of the locus subject to TRD, we screened 61 CC lines and 378 DO mice that had been genotyped with MegaMUGA for recombinations involving the WSB/EiJ strain in the 75-90 Mb interval of Chr 2. We identified five DO females (DO-600, DO-681, DO-732, DO-832 and DO-OCA45) and two CC lines (OR15155 and AU8042) that each had at least one informative recombination (Figure 1). Next, we mated four of the DO females (all except DO-OCA45 that was already heterozygous) and the two CC lines to one of two additional CC lines (CC001/Unc and CC005/TauUnc) that had no contribution from WSB/EiJ on Chr 2, to obtain heterozygous F1 hybrid females. Each hybrid female was genotyped with MegaMUGA and mated to FVB/NJ males (total of 35 crosses, Table 4.2).

We found that the Chr 2 interval exhibited significant TRD in eight of the ten chromosomes (mean 0.9 ± 0.094, $p = 9.6\text{x}10^{-54}$, Figure 4.3 A), but not in the other two chromosomes (mean 0.425 ± 0.078, $p = 0.21$, Figure 4.3 B). Females exhibiting Mendelian segregation ratios were not used for mapping because, as shown in the previous section, heterozygosity on Chr 2 is required but not sufficient for TRD. Females with TRD in favor of the WSB/EiJ allele were all heterozygous for a 9.3 Mb interval (boxed in Figure 4.3). The proximal boundary of the interval is defined by the recombination found in the CC strain OR15155 (i.e., the most distal SNP inconsistent with a WSB/EiJ haplotype). The distal boundary of the interval is defined by the recombination found in DO-732 and DO-832 females (i.e., the most proximal SNP inconsistent with a WSB/EiJ haplotype). Those SNPs define the boundaries of the locus subject to TRD, Chr 2 76,860,361 – 86,117,205 (dbSNP IDs rs253778980 and rs33743639, respectively; all positions from NCBI/37 unless otherwise noted). We named this locus *R2d2* for "*Responder to drive on Chr 2*" based on the origin of the TRD (see last section of the Results).

Supplementary Table 2. Segregation ratio and litter size in *R2d2*$^{WSB/other}$ heterozygous DO, CCxDO and CCxCC dams

Dam	Sire	Informative parent	*R2d2*^{WSB}	*R2d2*^{notWSB}	TR	*p*	ALS ± SD	Live/Dead embryos
DO-G13-001	FVB/NJ	dam	22	22	0.5	1	7.3 ± 1.5	5/2
DO-G13-003	FVB/NJ	dam	30	25	0.55	0.59	9.2 ± 1.6	10/0
DO-G13-004	FVB/NJ	dam	16	16	0.5	1	6.8 ± 2.4	5/1
DO-G13-005	FVB/NJ	dam	21	8	0.72	0.024	5.8 ± 2.3	9/1
DO-G13-006	FVB/NJ	dam	24	6	0.8	0.0014	5.0 ± 1.9	7/1
DO-G13-007	FVB/NJ	dam	31	6	0.84	$4.10x10^{-05}$	6.2 ± 2.0	10/0
DO-G13-008	FVB/NJ	dam	19	25	0.43	0.45	8.8 ± 1.9	7/2
DO-G13-009	FVB/NJ	dam	29	30	0.49	1	8.6 ± 0.7	7/0
DO-G13-010	FVB/NJ	dam	39	26	0.6	0.14	9.3 ± 2.0	10/1
DO-G13-011	FVB/NJ	dam	20	4	0.83	0.0015	5.8 ± 1.2	5/2
DO-G13-012	FVB/NJ	dam	17	9	0.65	0.17	8.7 ± 0.5	nd
DO-G13-013	FVB/NJ	dam	34	11	0.76	0.00082	7.5 ± 1.9	4/0
DO-G13-014	FVB/NJ	dam	32	3	0.92	$4.20x10^{-07}$	5.8 ± 1.6	6/1
DO-G13-015	FVB/NJ	dam	19	15	0.56	0.61	6.8 ± 1.5	nd
DO-G13-016	FVB/NJ	dam	15	21	0.42	0.41	9.0 ± 1.6	8/3
DO-G13-017	FVB/NJ	dam	21	16	0.57	0.51	7.4 ± 2.1	9/0
DO-G13-018	FVB/NJ	dam	25	11	0.69	0.029	5.4 ± 1.7	2/7
DO-G13-019	FVB/NJ	dam	23	4	0.85	0.00031	3.9 ± 1.1	4/1
DO-G13-020	FVB/NJ	dam	32	10	0.76	0.00094	7.0 ± 0.6	6/0
DO-G13-021	FVB/NJ	dam	31	11	0.74	0.0029	7.0 ± 1.2	6/3
DO-G13-022	FVB/NJ	dam	37	22	0.64	0.067	9.8 ± 0.9	10/2
DO-G13-023	FVB/NJ	dam	13	9	0.59	0.52	5.5 ± 1.1	5/2
DO-G13-024	FVB/NJ	dam	28	30	0.48	0.9	9.8 ± 2.2	6/4
DO-G13-025	FVB/NJ	dam	43	12	0.78	$3.30x10^{-05}$	6.7 ± 0.7	9/2
DO-G13-026	FVB/NJ	dam	23	2	0.92	$1.90x10^{-05}$	5.0 ± 1.4	7/1
DO-G13-028	FVB/NJ	dam	30	16	0.65	0.054	7.8 ± 1.7	8/0
DO-G13-029	FVB/NJ	dam	21	10	0.68	0.071	6.2 ± 0.7	7/1
DO-G13-033	FVB/NJ	dam	29	11	0.73	0.0064	8.0 ± 1.3	10/0
DO-G13-034	FVB/NJ	dam	32	15	0.68	0.019	6.7 ± 2.1	nd
DO-G13-035	FVB/NJ	dam	34	18	0.65	0.036	7.4 ± 2.2	7/1
DO-G13-036	FVB/NJ	dam	26	15	0.63	0.12	8.2 ± 1.5	8/0
DO-G13-037	FVB/NJ	dam	20	25	0.44	0.55	9.0 ± 1.3	11/0
DO-G13-038	FVB/NJ	dam	30	15	0.67	0.036	7.5 ± 2.3	7/1
DO-G13-039	FVB/NJ	dam	8	9	0.47	1	5.8 ± 0.6	6/1
DO-G13-040	FVB/NJ	dam	14	16	0.47	0.86	7.8 ± 2.9	3/2
DO-G13-041	FVB/NJ	dam	33	38	0.46	0.64	10.1 ± 1.0	11/0
DO-G13-042	FVB/NJ	dam	20	17	0.54	0.74	9.3 ± 0.4	9/0
DO-G13-043	FVB/NJ	dam	23	11	0.68	0.058	5.7 ± 2.4	7/0
DO-G13-045	FVB/NJ	dam	8	6	0.57	0.79	7.0 ± 2.0	nd
DO-G13-046	FVB/NJ	dam	23	0	1	$2.40x10^{-07}$	4.6 ± 0.8	2/1
DO-G13-047	FVB/NJ	dam	17	3	0.85	0.0026	4.0 ± 1.4	2/2
DO-G13-048	FVB/NJ	dam	31	22	0.58	0.27	8.8 ± 1.6	10/1
DO-G13-049	FVB/NJ	dam	16	21	0.43	0.51	7.5 ± 1.3	8/2
DO-G13-050	FVB/NJ	dam	24	1	0.96	$1.50x10^{-06}$	5.0 ± 2.4	1/1
DO-G13-051	FVB/NJ	dam	24	9	0.73	0.014	4.7 ± 3.2	7/4
DO-G13-052	FVB/NJ	dam	28	17	0.62	0.14	7.5 ± 1.4	nd
DO-G13-054	FVB/NJ	dam	21	16	0.57	0.51	8.0 ± 1.9	11/0
DO-G13-056	FVB/NJ	dam	20	15	0.57	0.5	7.0 ± 1.7	nd
DO-G13-057	FVB/NJ	dam	17	20	0.46	0.74	6.2 ± 1.6	nd
DO-G13-059	FVB/NJ	dam	15	4	0.79	0.019	4.0 ± 2.1	8/0
DO-G13-061	FVB/NJ	dam	11	13	0.46	0.84	6.0 ± 2.5	9/0
DO-G13-063	FVB/NJ	dam	23	9	0.72	0.02	6.6 ± 1.4	7/2
DO-G13-064	FVB/NJ	dam	18	7	0.72	0.043	5.0 ± 2.3	9/1
DO-G13-065	FVB/NJ	dam	10	3	0.77	0.092	2.8 ± 2.1	7/1
DO-123	FVB/NJ	dam	22	17	0.56	0.52	7.8 ± 1.5	nd
OLA-45	FVB/NJ	dam	16	14	0.53	0.86	7.4 ± 1.5	8/1
(DO-681xCC001/Unc)F1-018	FVB/NJ	dam	22	24	0.48	0.88	9.4 ± 2.7	nd
(DO-681xCC001/Unc)F1-019	FVB/NJ	dam	27	30	0.47	0.79	11.8 ± 1.8	nd
(DO-832xCC001/Unc)F1-001	FVB/NJ	dam	21	5	0.81	0.0025	5.2 ± 2.3	nd
(DO-832xCC001/Unc)F1-002	FVB/NJ	dam	23	5	0.82	0.00091	7.3 ± 1.5	nd
(DO-832xCC001/Unc)F1-003	FVB/NJ	dam	11	0	1	0.00098	6.0 ± 1.0	nd
(DO-832xCC001/Unc)F1-004	FVB/NJ	dam	34	0	1	$1.20x10^{-10}$	7.7 ± 1.2	nd
(DO-832xCC001/Unc)F1-005	FVB/NJ	dam	23	0	1	$2.40x10^{-07}$	8.0 ± 0.8	nd
(DO-832xCC001/Unc)F1-008	FVB/NJ	dam	10	1	0.91	0.012	5.5 ± 0.5	nd
(DO-832xCC001/Unc)F1-009	FVB/NJ	dam	8	4	0.67	0.39	6.0 ± 1.0	nd
(DO-600xCC005/TauUnc)F1-005	FVB/NJ	dam	21	2	0.91	$6.60x10^{-05}$	5.8 ± 1.9	nd
(DO-600xCC005/TauUnc)F1-007	FVB/NJ	dam	17	0	1	$1.50x10^{-05}$	4.5 ± 1.7	nd
(DO-600xCC005/TauUnc)F1-009	FVB/NJ	dam	5	0	1	0.062	5.0 ± 0.0	nd
(DO-600xCC005/TauUnc)F1-010	FVB/NJ	dam	11	2	0.85	0.022	6.5 ± 1.5	nd
(DO-600xCC005/TauUnc)F1-011	FVB/NJ	dam	9	0	1	0.0039	5.0 ± 2.7	nd
(DO-732xCC005/TauUnc)F1-001	FVB/NJ	dam	13	0	1	0.00024	3.3 ± 2.5	nd
(DO-732xCC005/TauUnc)F1-002	FVB/NJ	dam	18	1	0.95	$7.60x10^{-05}$	6.7 ± 3.1	nd
(DO-732xCC005/TauUnc)F1-003	FVB/NJ	dam	21	1	0.95	$1.10x10^{-05}$	5.5 ± 1.1	nd
(DO-732xCC005/TauUnc)F1-008	FVB/NJ	dam	7	0	1	0.016	7	nd
(DO-732xCC005/TauUnc)F1-009	FVB/NJ	dam	9	1	0.9	0.021	5.0 ± 0.0	nd
(DO-732xCC005/TauUnc)F1-010	FVB/NJ	dam	5	0	1	0.062	5	nd
(DO-732xCC005/TauUnc)F1-011	FVB/NJ	dam	2	0	1	0.5	2	nd
(DO-732xCC005/TauUnc)F1-012	FVB/NJ	dam	5	0	1	0.062	5	nd
(AU8042xCC001/Unc)F1-004	FVB/NJ	dam	3	3	0.5	1	6	nd
(AU8042xCC001/Unc)F1-005	FVB/NJ	dam	8	8	0.5	1	8.5 ± 0.5	nd
(AU8042xCC001/Unc)F1-006	FVB/NJ	dam	9	6	0.6	0.61	7.5 ± 1.5	nd
(AU8042xCC001/Unc)F1-007	FVB/NJ	dam	11	3	0.79	0.057	7.0 ± 0.0	nd
(AU8042xCC001/Unc)F1-008	FVB/NJ	dam	11	7	0.61	0.48	9.0 ± 0.0	nd
(AU8042xCC001/Unc)F1-009	FVB/NJ	dam	14	3	0.82	0.013	8.5 ± 0.5	nd
(AU8042xCC001/Unc)F1-010	FVB/NJ	dam	6	4	0.6	0.75	10	nd
(AU8042xCC001/Unc)F1-012	FVB/NJ	dam	8	1	0.89	0.039	9	nd
(AU8042xCC001/Unc)F1-013	FVB/NJ	dam	5	1	0.83	0.22	6	nd
(AU8042xCC001/Unc)F1-014	FVB/NJ	dam	10	2	0.83	0.039	12	nd
(CC001/UncxOR15155)F1-005	FVB/NJ	dam	11	0	1	0.00098	6.0 ± 1.0	nd
(CC001/UncxOR15155)F1-006	FVB/NJ	dam	13	4	0.76	0.049	8.5 ± 0.5	nd
(CC001/UncxOR15155)F1-007	FVB/NJ	dam	11	0	1	0.00098	6.5 ± 10.5	nd

4.5.4 A 4.3 Mb-long expansion is the causative allele at *R2d2*

Among the eight CC founder strains, the *R2d2* candidate interval has 8,195 SNPs, 2,224 small insertions/deletions and 32 structural variants (SV) that are private to the WSB/EiJ strain [56, 196]. Although this very large number of variants would typically make it difficult to confidently identify and prioritize candidates, one large SV has several unique features that made it an exceptionally strong candidate causative allele for the TRD phenotype. That SV is best described as a copy number gain of a 127 kb-long genomic DNA sequence. As a unit, that sequence is unique (i.e., is present as a single copy) in the C57BL/6J reference genome. The 127kb unit is composed of nine non-contiguous sections that, in total, span 158 kb of the reference genome (Chr 2 77,707,014 – 77,865,265, Figure 4.4).

We used the normalized per-base read depth from whole-genome sequence alignments generated by the Sanger Mouse Genomes Project [56] and in [197] to estimate the number of copies of the 127 kb unit in 20 inbred strains (see Materials and Methods). Similar to C57BL/6J, 15 of the 20 strains, including 5 CC founder strains (A/J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HlLtJ and PWK/PhJ) were copy number 1 (i.e., a single haploid copy), and CAST/EiJ was copy number 2. In contrast, WSB/EiJ was copy number 34 and SPRET/EiJ was copy number 36, resulting in ~ 4.4 Mb of additional DNA in those strains (Figure 2). The two ICR lines (HR3 and HR7) had an intermediate copy number (~ 10). We hypothesize that the large copy number gains in WSB/EiJ, SPRET/EiJ and the ICR lines are causative of TRD in those strains.

Many structural variants identified from whole-genome sequencing reads have uncertain genomic positions due to the challenge of mapping large variants that are absent from the reference genome. To determine the position of the copy number gain, we mapped the WSB/EiJ and CAST/EiJ alleles using segregating populations that have been genotyped at medium (MegaMUGA) or high (MDA) density [198]. In the CC founder strains, probes located in the 127 kb unit (three probes in MegaMUGA and 68 probes in MDA) have hybridization intensities correlated with the copy numbers estimated from aligned read depth. MDA provides

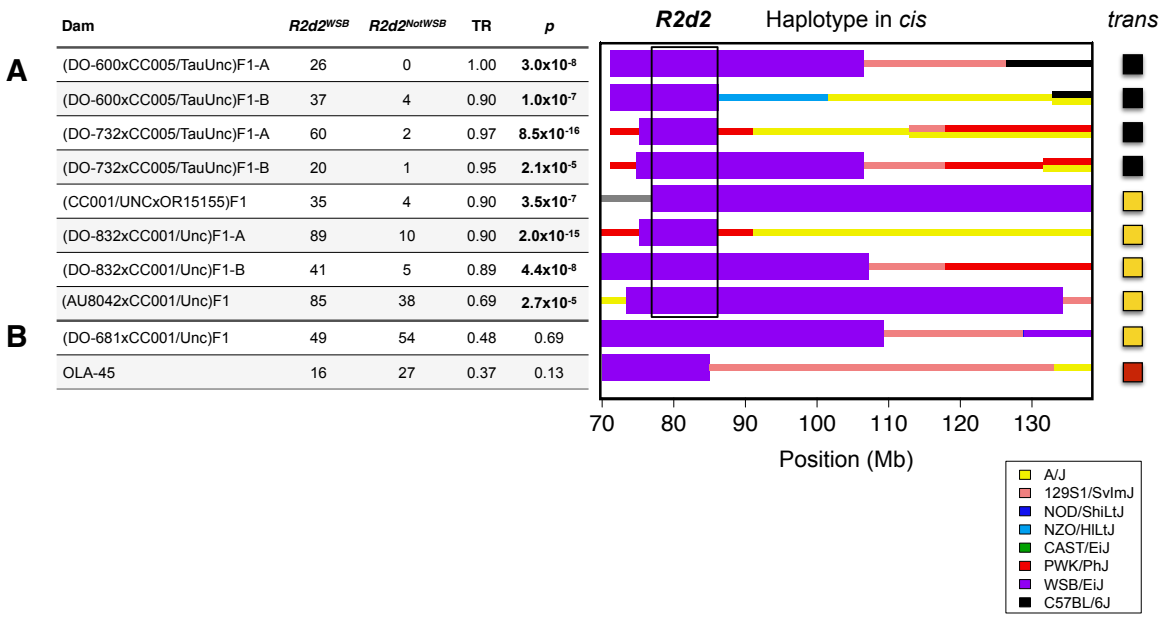


Figure 4.3: *R2d2* maps to a 9.3 Mb candidate interval. F1 females from two CCxCC and four DOxCC crosses were test crossed to FVB/NJ sires. Because CC lines are inbred, they have a single haplotype; therefore, there were ten unique haplotypes among the F1 females. The aggregate number of WSB/EiJ and Non-WSB/EiJ alleles transmitted by females of each haplotype (left panel) and the corresponding haplotypes in *cis* (middle panel) and *trans* (right panel) to the WSB/EiJ allele are shown for females A) with TRD and B) without TRD. Significance of deviation of TR from Mendelian expectation of 0.5 was computed using a binomial exact test (*p*-value). Thick purple bars indicate the extent of WSB/EiJ contributions, and thin bars indicate the extent of contributions from all other strains. The black box indicates the boundaries of the candidate interval as determined by the region that is WSB/EiJ in all females with TRD.

robust discrimination between the reference (1 copy), CAST/EiJ (2 copies) and WSB/EiJ alleles (34 copies) (Figure 4.5 A). MegaMUGA is able to discriminate mice carrying the WSB/EiJ allele with little ambiguity (Figure 4.5 B). Using the sum intensities of the informative probes, we mapped the WSB/EiJ and CAST/EiJ copy number gains in two independent populations and platforms. QTL mapping identified a single, broad, highly significant peak on mouse Chr 2 in each population, and those peaks overlap with each other and with the initial candidate interval (Figure 4.5 C-E). We conclude that the copy number gain is closely linked to the single copy found in the reference genome. This location is consistent with the large copy number gain being the causative allele.

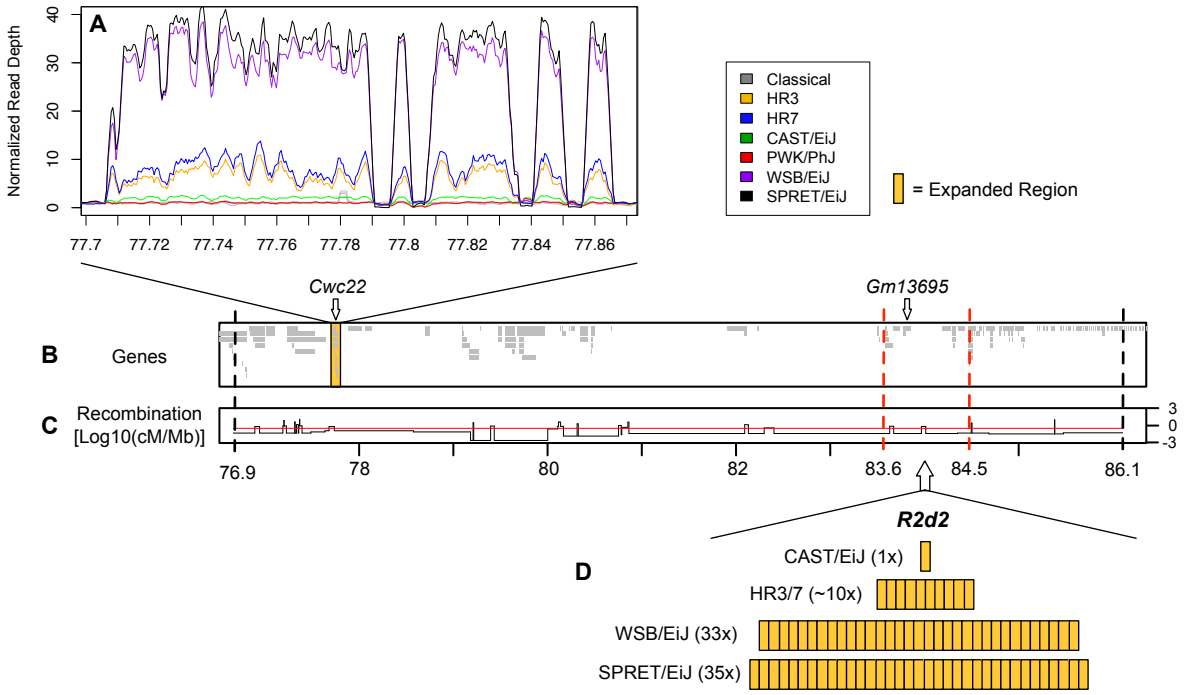


Figure 4.4: *R2d2* is a copy number gain that is novel with respect to the reference sequence. A) Within the 9.3 Mb candidate interval defined in Figure 4.3, we identified a 158 kb region with substantially enriched sequence read depth in some strains. Read depth in 100 bp windows, normalized by the genome-wide mean read depth for that strain, is shown for the three strains with TRD, WSB/EiJ (purple), SPRET/EiJ (black) and two sublines of ICR, HR3 (orange) and HR7 (blue), for CAST/EiJ (green), which has copy number expansion but no TRD, and PWK/PhJ (red) and the five classical CC founders (gray), which have neither the expansion nor TRD. Gaps in the copy number gain in represent regions that are present in the reference sequence but absent from some or all sequenced strains. B-C) The candidate interval is both gene-poor and recombination-cold. The yellow box shows the location of the 158 kb region that is expanded in (A). Vertical dashed lines indicate the boundaries of the candidate interval (black) and mapped insertion site (red). Gene tracks (B) show the locations of Ensembl genes in the NCBI/37 reference genome within the interval, with arrows indicating the location of *Cwc22* and its pseudogenes (*Gm13695*). The recombination track (c) shows the recombination frequency based on Liu et al. (2014), normalized by physical distance (Mb) and log10-transformed. The red line on the bottom track indicates the mean recombination frequency for Chr 2. D) Number of additional copies of the 127 kb unit that are present in CAST/EiJ, WSB/EiJ, the two ICR lines and SPRET/EiJ, and map somewhere within the highlighted 900 kb region.

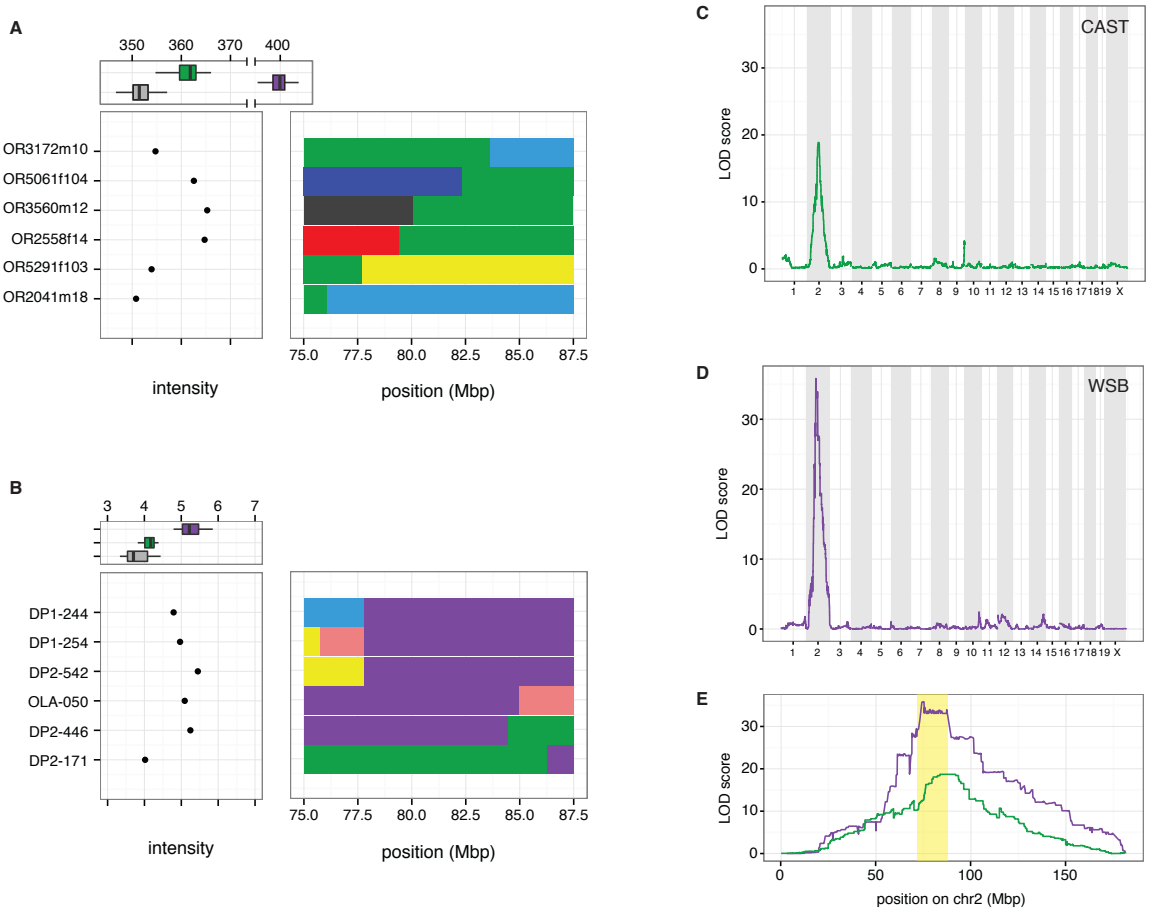


Figure 4.5: Linkage mapping localizes *R2d2* to a 900 kb region in Chr 2. A) Critical recombinants defining the CAST/EiJ copy-number gain (CNG) in the Collaborative Cross (CC) G2:F1 population. Sum-intensity across 38 probes in the *R2d2* locus on the MDA are plotted at left, and haplotypes of corresponding recombinant CAST/EiJ chromosomes shown at right. Distribution of sum-intensity for animals with a non-recombinant CAST/EiJ haplotype (green) or no CAST/EiJ haplotype (grey) in the interval is shown in the lower-left panel. B) Critical recombinants defining the WSB/EiJ copy number gain in the DO. Sum-intensity across 3 probes in the *R2d2* locus on the MegaMUGA array are plotted at left, and haplotypes of corresponding recombinant WSB/EiJ chromosomes shown at right. Distribution of sum-intensity for animals with a non-recombinant WSB/EiJ haplotype (green) or no WSB/EiJ haplotype (grey) in the interval is shown in the lower-left panel. C) LOD plot from a single-locus QTL scan in 330 CC G2:F1 animals, using MDA sum-intensity as the phenotype. D) LOD plot from a single-locus QTL scan in 96 (FVB/NJx(WSB/EiJxPWK/PhJ)F1)G2 offspring, using MegaMUGA sum-intensity as the phenotype. E) Superposition of LOD curves from panels (A) and (B) on Chr 2. The *R2d2* candidate interval is shaded in yellow.

Analysis of individual mice with recombinant chromosomes in the candidate interval revealed that the copy number gain maps to a 900 kb interval (Chr 2 83,631,096 – 84,541,308, rs28066428 and rs243209871, respectively, Figure 4.4 and Figure 4.5 A,B). Specifically, the CAST/EiJ copy number gain is located distal to the transition from the CAST/EiJ to the NZO/HILtJ haplotypes found in mice OR3172m10 and OR3172f9, because both mice have low hybridization intensity consistent with a single copy the absence of the CAST/EiJ copy number gain (2 copies, Figure 4.5 A). Similarly, the WSB/EiJ copy number gain is located proximal to the transition from the WSB/EiJ to the CAST/EiJ haplotype found on DO mouse DP2-446, because it had high hybridization intensity consistent with the presence the WSB/EiJ copy number gain (34 copies, Figure 4.5 B). These results demonstrate that the copy number gain is not located immediately adjacent to the single copy present in reference genome but approximately 6 Mb distal to it. This copy number gain is most likely the causative allele for TRD at *R2d2*.

We also identified nine DO females with apparent recombination inside of the minimum region. Out of these, a single female had a transition from $R2d2^{NotWSB}$ to $R2d2^{WSB}$ and also exhibited TRD. This allowed me to slightly narrow the proximal boundary of *R2d2* to 77,138,390 (rs33371061).

4.5.5 Meiotic drive causes maternal TRD at *R2d2*

The results above demonstrate that TRD at *R2d2* is only observed in the progeny of heterozygous females. This restricts the plausible causes of TRD to meiotic drive, genotype-dependent embryonic lethality or a combination of both. To identify the cause of TRD, we first determined whether TR levels were correlated with litter size in 56 DO females (these females are a random sample from an outbred population, Figure 4.6 and Table 4.2). We observed a strong inverse correlation between average litter size and TR at *R2d2* (r^2 = -0.65, $p = 7.2x10^{-8}$, Figure 4.7 A). We conclude that the presence and the strength of TRD is significantly associated with reduced litter sizes and thus with some type of embryonic lethality.

However, the inferred level of lethality is insufficient to explain the observed level of TRD. For example, in females with Mendelian segregation the average litter size was 7.8, while in DO females with extreme TRD (>0.92) the average litter size was 5.2. In other words, embryonic death could only account for a fraction of the "missing" progeny with *R2d2*^{NotWSB} genotype. We directly determined levels of embryonic lethality by euthanizing and dissecting pregnant DO females at mid-gestation (see Materials and Methods). We observed that females with TRD had slightly, but not significantly, higher numbers of resorbed embryos present in utero than did females with Mendelian segregation (1.3 and 1.1 resorbed embryos, respectively, p = 0.66, Figure 4.7 B). We conclude that embryonic lethality alone is insufficient to explain TRD at *R2d2*.

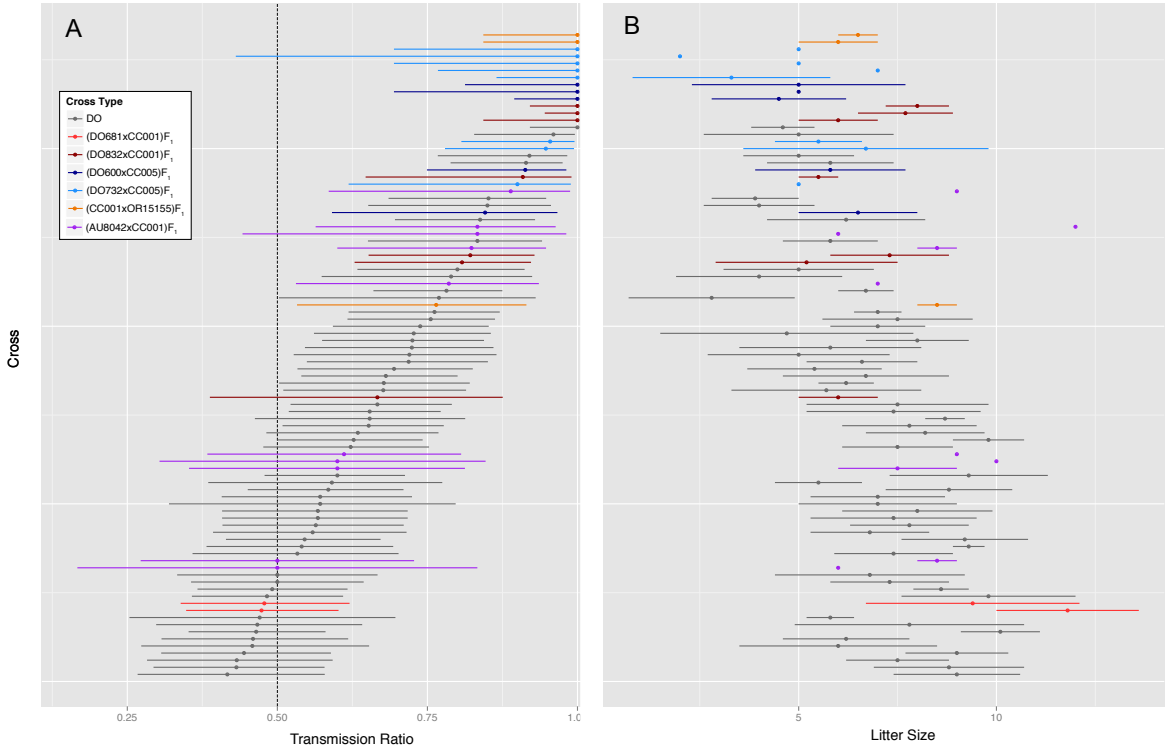


Figure 4.6: TR and Litter Size are variable in DO and CC crosses. A) TRs (points) and 95% confidence intervals (lines) for the different types of crosses indicated in the legend. Gray points represent test-crosses between heterozygous DO females and FVB/NJ males. All other crosses are those that appear in Figure 1. Dotted line shows Mendelian expectation of 0.5. B) Corresponding average litter size (points) and standard deviation (lines) for the crosses in panel (A). Crosses represented only by a point had a single litter.

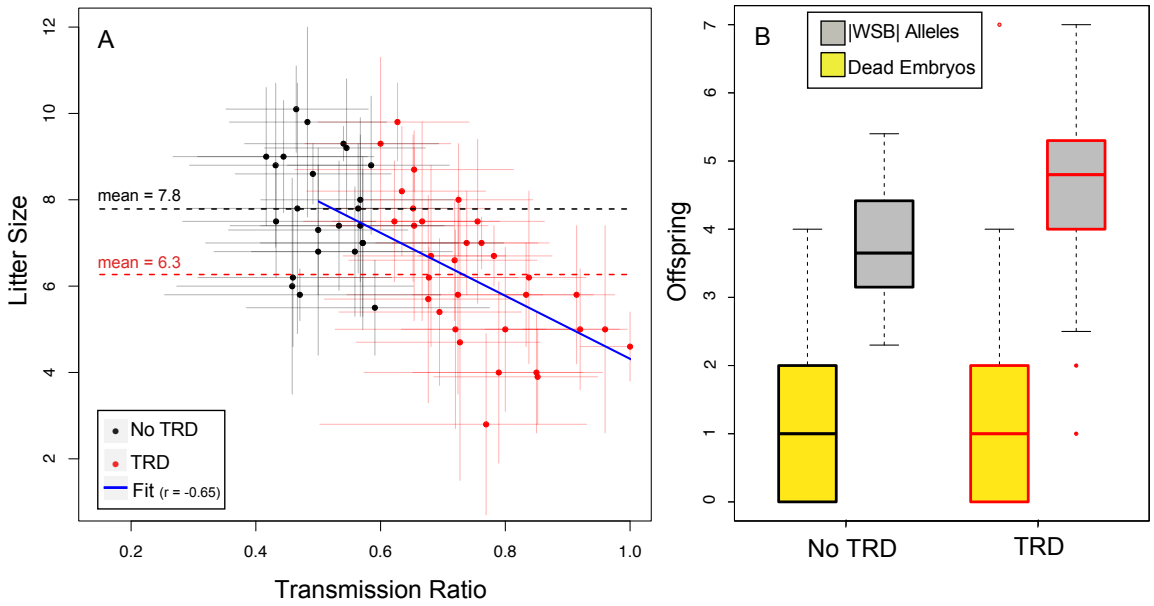


Figure 4.7: TRD at *R2d2* is explained by both meiotic drive and embryonic lethality. A) Litter size is shown in relation to TR for test crosses between DO females and FVB/NJ males without TRD (black dots) and with TRD (red dots). Horizontal bars show TR standard error and vertical bars show litter size standard error. Dotted lines show mean litter sizes for females without and with TRD (black and red, respectively). Blue line shows a linear fit to TR and average litter size. B) Boxplots of mean absolute number of offspring bearing WSB/EiJ alleles per litter (excluding the final litter, gray) and number of dead embryos in the final litter (yellow) for females without TRD (left boxes, black outlines) and with TRD (right boxes, red outlines). There are significantly more WSB/EiJ-carrying offspring of females with TRD than without, which is evidence of meiotic drive. Embryonic lethality is not different between females with and without TRD.

Although embryonic lethality can change the proportion of progeny inheriting alternative alleles at *R2d2*, only meiotic drive can lead to an increase in the absolute number of progeny inheriting the $R2d2^{WSB}$ allele per litter in females with TRD compared to females with Mendelian segregation. To directly test whether meiotic drive was responsible for TRD, we determined the average absolute number of offspring per litter that inherited the $R2d2^{WSB}$ and $R2d2^{NotWSB}$ alleles in the progenies of the DO females with either TRD or Mendelian segregation (Figure 4B). In females with Mendelian segregation, the average number of pups per litter that inherited either allele was the same (3.80 $R2d2^{WSB}$ and 3.96 $R2d2^{NotWSB}$), and the sum

of both was equal to the average litter size of those females. In contrast, in the progenies of females with TRD the average of the number of offspring per litter that inherited the $R2d2^{WSB}$ allele (4.51) was significantly greater than the absolute number of either allele in the offspring of females without distortion ($p = 0.0064$ and $p = 0.049$ for the $R2d2^{WSB}$ and $R2d2^{NotWSB}$ alleles, respectively). The same result holds true for viable embryos at mid-gestation: the average number of offspring that inherited $R2d2^{WSB}$ and $R2d2^{NotWSB}$ alleles was 5.1 and 1.6 for females with distortion and 4.0 and 3.4 for females without distortion. From the consistent and significant excess of $R2d2^{WSB}$ alleles in the offspring of females with TRD, we conclude that meiotic drive is required to explain TRD at *R2d2*.

4.6 Discussion

4.6.1 How do meiotic drive and embryonic lethality contribute to TRD at *R2d2?*

A major focus of our study was to discriminate among the many mechanisms that could give rise to TRD at *R2d2*, and to rule out as many as possible. First, the fact that TRD is only observed through the female germline rules out both spermatogenesis-mediated processes and sperm competition. Second, the presence of TRD at birth rules out differential postnatal survival. Third, the fact that distortion was independent of the maternal granddam precludes cytoplasmic effects. The remaining plausible explanations are differential fertilization based on the oocyte genotype, embryonic lethality and/or meiotic drive. The first two mechanisms should reduce the average litter size (ALS) proportionally to the level of distortion ($ALS^{observed} = ALS^{expected}$ x TRD) while not changing the average absolute number of offspring inheriting the favored genotype ($R2d2^{WSB}$) per litter. The number of resorbed embryos observed in pregnant females should distinguish the two mechanisms because it should be greater in the second than in the first scenario. In contrast, if meiotic drive is responsible for TRD then the following should be true: 1) average litter size is independent of TRD, 2) the average absolute number of offspring inheriting the favored genotype ($R2d2^{WSB}$) per litter is higher in females with TRD than in females with Mendelian segregation, and 3) the level of embryonic

lethality is independent of the presence and level of distortion. The data shown in the results require the combined action of embryonic lethality and meiotic drive. Specifically, meiotic drive is required to explain both the fact that $ALS^{observed}$ is significantly greater than predicted ($ALS^{expected}$ x TRD), and the significantly greater average absolute number of offspring with the $R2d2^{WSB}$ genotype in females with distortion.

It is interesting to speculate about the types of embryonic lethality that are consistent with our data and previous observations. Lethality is associated with distortion at *R2d2*, and thus the simplest explanation is preferential death of embryos inheriting maternal $R2d2^{NotWSB}$ alleles. However, such a scenario would require parent-of-origin-dependent death of embryos with maternal C57BL/6J, 129S1/SvIm, NOD/ShiLtJ and NZO/HILtJ *R2d2* alleles in the F1 females (Table 4.1) and CAST/EiJ, PWK/PhJ and A/J *R2d2* alleles in the CC/DO females. The lack of evidence of TRD and parent-of-origin lethality in dozens of crosses involving these alleles, combined with the lack of evidence for imprinted genes in the central region of Chr 2 [199] rules out this explanation. A more likely explanation for the combined and correlated presence of meiotic drive and lethality is that the unequal segregation of Chr 2 that leads to TRD in euploid embryos may also lead to increased aneuploidy, and thus to embryonic death (all autosomal aneuploidy is embryonic lethal in the mouse). This would also explain the slight increase in the number of resorbed embryos observed at mid-gestation (Figure 4.7 B and Table 4.2). This hypothesis makes the testable prediction that Chr 2 should be especially affected by aneuploidy in females with TRD.

4.6.2 Mapping the responder and identification of the causative allele

Known meiotic drive systems consist of one or more *responder* loci (a locus subject to preferential segregation during meiosis) and a single distorter (the effector locus required for drive at the *responder*). In this study we were able to map the *responder*, *R2d2*, to a small interval (900 kb) that is comparable in size to ones obtained in GWAS in mammals with much larger sample sizes. Such a result could not be anticipated from our initial efforts at mapping

R2d2 by determining the TR in the progeny of a modest number of females with recombinant chromosomes (Figure 4.3). In fact, we were faced with two major obstacles in our mapping efforts. First, the central region of Chr 2 contains a large (9.3 Mb) recombination-cold region in which the frequency of recombination is three-fold lower than expected in the CC [198] (2.3 vs. 6.8 cM in females and 1.4 vs 5.3 cM in males). Although this likely contributes to the overall deficit in recombinant chromosomes (we expected 23 in the 378 DO females and 4 in 61 CC lines), the complete lack of recombinants involving WSB/EiJ is striking, and, for the purposes of this study, a major impediment to the precise mapping of *R2d2*. Second, although heterozygosity is required, it is not sufficient for meiotic drive (Table 4.1 and Table 4.2). Therefore, we mapped *R2d2* by determining the minimum region of overlap for the WSB/EiJ haplotype only in females with TRD (Figure 4.3). This yielded a 9.3 Mb initial candidate interval for *R2d2*. Within that candidate interval, a single variant stands out as the most likely cause of TRD: a large copy number gain that is unique to the WSB/EiJ strain among CC founders, and apparently shared in common with SPRET/EiJ and ICR, the other inbred strains for which maternal TRD in central Chr 2 has been described. The similarities between WSB/EiJ and SPRET/EiJ extend to the presence of TRD when crossed with C57BL/6J (Table 4.1 and [192]) and Mendelian segregation when crossed with A/J (Table 4.1 and [200]).

Using whole-genome sequence alignments and variant calls from the Sanger Mouse Genomes Project [56, 196], we identified a 34-fold enrichment of sequencing reads that aligned across a 127 kb region of the reference genome. However, the reference genome is based on a single classical inbred strain, C57BL/6J; structural variation in other strains or wild mice may result in a different physical size (larger or smaller) or spatial organization of the candidate interval. Therefore, the sequencing reads that mapped to 77.70-77.82 Mb in the reference genome may in fact have derived from sequence with a different physical location in the WSB/EiJ genome. Fortunately, the presence of a third allele in CAST/EiJ (which exhibited a twofold enrichment of sequencing reads) enabled us to map the physical location of *R2d2* to a 900 kb region that (at least in the reference genome) is 6 Mb distal to where the se-

quencing reads mapped (Figure 4.5 D,E). Based on the large number of animals genotyped without observing a single recombination involving WSB/EiJ within the candidate interval, we speculate that the copy number gain is associated with the dramatic reduction in recombination observed in hybrids involving that strain. We wish to emphasize that fine mapping would have been impossible without deep sequence data for each of the strains used in these experiments [53, 56], and without combining the results of experiments completed 20 years apart [192, 193, 194, 200]. We believe that such integration of related data from old and new experiments has great potential to accelerate the progression from QTLs to causal variants.

Suppressed local recombination is common in the vicinity of meiotic drive *responders* and is often the result of structural variation. Interestingly, there were multiple small deletions (around 100-5000bp) flanking the cold region that were present in the three strains of the inter-subspecific crosses (WSB, PWK/PhJ, CAST/EiJ). There were also an enrichment for male recombinations immediately preceding and following the cold region. Those features were indicative of an the cold region being an inversion; however, the Sanger SV catalogue did not include an inversion in that region.

The 9.3 Mb candidate interval is gene-poor. Excluding the genes in the major olfactory cluster that begins $\sim 85Mb$, there are only 56 genes (Ensembl) in this region, with an average spacing of 160kb as compared to the genome wide average of 100kb. The region between 80.5Mb and 84Mb is a gene desert (7 genes). C+G content deviates significantly from the Chr 2 average of 40% only underneath the dense cluster of olfactory genes at 85Mb; C+G enrichment is normal in regions dense for protein-coding genes.

The 127 kb unit spans a single annotated gene, *Cwc22*, implicated in RNA splicing [201]. DNA copy number variation for *Cwc22* has been described previously [196]. *Cwc22* is highly expressed in mouse oocytes and fertilized eggs [202], and it is a known eQTL in mouse. Allele-specific RNA-seq of brain tissue from reciprocal crosses between WSB/EiJ, PWK/PhJ and CAST/EiJ showed extreme differential expression, with the WSB/EiJ allele more highly expressed than the other two (Crowley et al. under review). Whether expression of *Cwc22*

is causally implicated in meiotic drive and/or embryonic lethality remains to be determined. *Cwc22* has 9 paralogues (pseudogenes) in the mouse, seven on Chr 2 and two on Chr X. The Chr 2 paralogues exist in tandem at 83.95-84.00Mb, within the 900 kb region where we mapped the expansion (Figure 4.4).

In our study of subspecific origin in laboratory mice, we found that many non-*M. m. domesticus* wild-derived strains had evidence of introgression from *M. m. domesticus*, including PWK/PhJ and CAST/EiJ [40]. I used the Mouse Genome Browser [203] to determine the subspecific origin of the founder strains within the *R2d2* region. All strains except CAST/EiJ were of expected origin (i.e., classical strains and WSB/EiJ were *M. m. domesticus* and PWK/PhJ was *M. m. musculus*). Between 80-86Mb, CAST/EiJ was of *M. m. domesticus* origin. There is also a small block of *M. m. domesticus* between 78.8-79.2Mb. Importantly, CAST/EiJ did not exhibit any evidence of TRD, which meant that either the $R2d2^{WSB}$ allele is segregating in *M. m. domesticus* and that allele is not present in CAST/EiJ, or that the region between 76-78.8Mb is necessary for TRD.

Apart from its size and repetitive nature, the most important feature of the *R2d2* candidate allele is its remarkable uniformity between WSB/EiJ and SPRET/EiJ, two inbred strains that are separated by ~ 1 million years of evolution [43]. First, the number of copies of the expansion in the two strains is very similar (33 and 35). Second, compared to a genome-wide average of 1 SNP every ~ 70 bp between the two strains [56], within the 127 kb candidate allele there is only 1 SNP every 1,200 bp (a 17-fold reduction in sequence divergence). One possible explanation is that paralogous variation between different copies of the 127 kb unit are masking homologous variation between WSB/EiJ and SPRET/EiJ, i.e., since heterozygosity should not be possible in an inbred strain, any evidence of heterozygosity in whole-genome sequence alignments was treated as sequencing error by the Sanger Mouse Genomes Project, and thus SNPs are substantially undercalled in *R2d2*. Further analysis will be required to determine the age of the expansion, and whether interspecific introgression [204, 205] is required to explain an otherwise unlikely degree of sequence conservation between *M. m. domesticus*

and *M. spretus*.

Although we believe that we have correctly identified *R2d2*, we note that the causal allele may incorporate additional DNA sequences, including some that may be absent in the reference genome (similar to the origin of the sequence on maize chromosome Ab10 that causes meiotic drive in that species). If that is the case, the causal allele may be much larger than 4.3 Mb. For example, HSR alleles as large as 200 Mb have been described [206, 207].

4.6.3 Is meiotic drive at *R2d2* a polygenic trait?

Overall, we assessed TR at *R2d2* in hundreds of females carrying a single WSB/EiJ allele from dozens of distinct genetic backgrounds (Table 4.1 and Table 4.2). The presence of significantly different TR levels among F1 hybrid females demonstrates that meiotic drive is under genetic control in trans (i.e., there is at least one unlinked *distorter* locus that is genetically variable in the CC/DO). Furthermore, the presence of at least two significantly different levels of distortion indicate either that there is more than one *distorter* locus involved or alternatively that there is an allelic series at a single *distorter* locus.

In most meiotic drive systems, *responder* and *distorter* loci are tightly linked and are typically protected from decoupling by factors that inhibit recombination, such as structural variation [21, 6, 191]. Although *R2d2* resides within a recombination-cold region, the *distorter* is not closely linked to *R2d2*. If TRD was solely dependent on a linked *distorter*, then we would expect that females in which the *trans* allele is known not to give rise to TRD (A/J, PWK/PhJ, CAST/EiJ, based on the F1 data in Table 4.1) would not have TRD; however, that was not the case (Figure 4.3). Therefore, at least one unlinked modifier is required to explain the observed variability in TRD. Furthermore, it is unlikely that the unlinked modifier is located anywhere on Chr 2, since there is no place on that chromosome where at least one CC or DO female that exhibits TRD does not have an allele associated with lack of TRD.

These observations indicate that the *R2d2* meiotic drive phenotype has a complex genetic architecture, in which multiple alleles at unlinked loci interact to determine whether distortion

occurs, and to what extent. This is unique among mammalian meiotic drive systems and has significant implications for the natural history of the system and for the ease of genetic dissection. Although we had a relatively small number of phenotyped females, we attempted to map the *distorters* using the DO/QTL software (Dan Gatti and Gary Churchill, unpublished). While we found no locus that reach significance, several were highly suggestive (Figure 4.8). We are currently phenotyping additional females and will repeat the experiment using ~ 200 individuals.

4.6.4 What is the mechanism by which *R2d2* influences segregation in *cis*

Centromeres (i.e., the site of kinetochore formation) are remarkable loci that control, in *cis*, proper segregation of chromosomes during mitosis and meiosis. It is easy to envision how *responder*s at, or tightly linked to, the centromeres can influence chromosome segregation. *Responder*s located far away from centromeres are thought to influence their own segregation in *cis* by becoming "neocentromeres" and taking advantage of the inherited functional polarity of the female meiotic spindle. Based on all the available evidence, we propose that *R2d2* acts as a neocentromere through the epigenetic activation mediated by C57BL/6J, NZO/ShiLtJ, 129S1/SvImJ, and NOD/HILtJ alleles at the *distorter*(s).

The effect of activating the ectopic neocentromere at *R2d2* on the Chr 2 centromere is unknown, but the slight level of lethality suggests some coordination in the segregation process. Centromere repositioning can happen due to a relocation of the ancestral centromere (e.g. through inversion), or due to activation of a latent centromere (neocentromere) combined with deactivation of the ancestral centromere. A neocentromere may completely replace the ancestral centromere, or it may only be active during meiosis. Heritable loss of the ancestral centromere requires either a deletion of sequence that is necessary for centromere function, or some novel epigenetic mechanism. Meiosis involving chromosomes with different centromere locations (and, possibly, sequences) may lead to an increased rate of non-disjunction and a reduced rate of recombination.

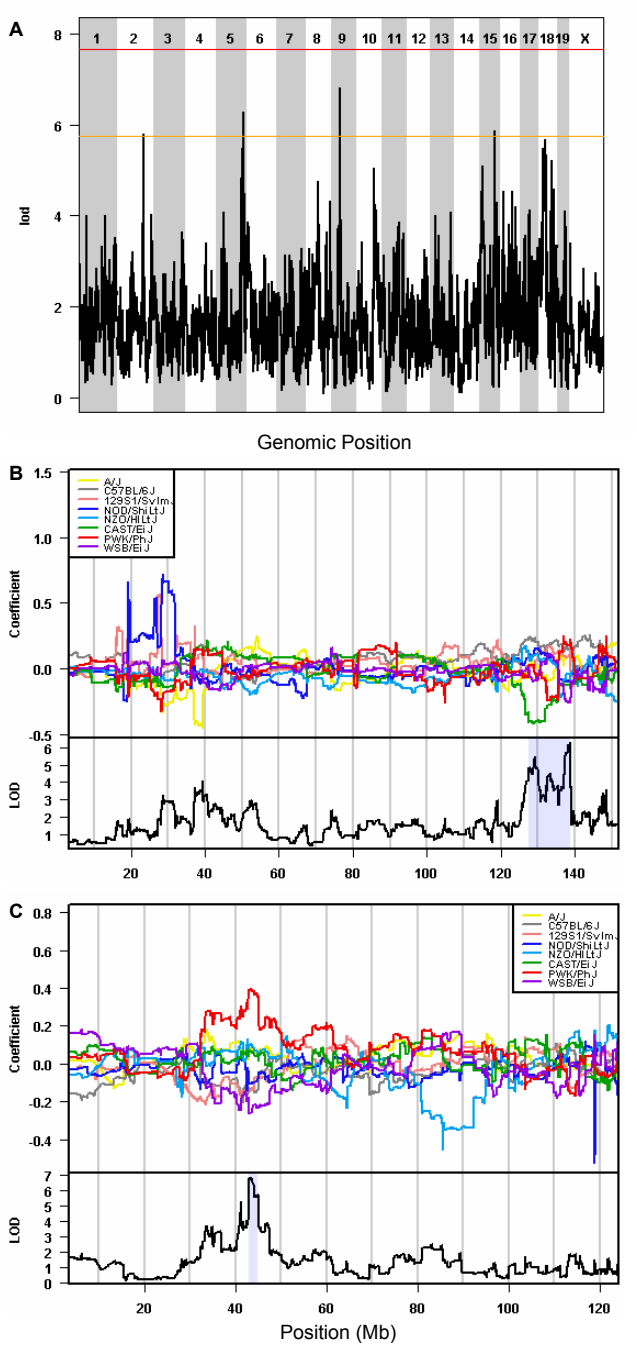


Figure 4.8: QTL mapping of modifiers identifies suggestive associations. A) Genome scan on genotype probabilities of 91 DO females using TR as a continuous phenotype. Red and orange lines indicate significance and suggestive thresholds, respectively. B-C) Effect size plots are shown for suggestive associations on B) Chr 5 and C) Chr 9. Highlighted regions in LOD-score panels show the region across which the QTL peak is above the suggestive threshold. Therefore, the largest effects are both negative and are contributed by B) CAST/EiJ and C) PWK/PhJ.

4.6.5 Revisiting TRD in the CC and DO

The conclusion that a genetically complex meiotic drive system is responsible for TRD favoring the WSB/EiJ allele at *R2d2* is fully consistent with the initial observations of TRD in the CC, with our prediction that positive selection of the WSB/EiJ allele occurred during outcrossing or in early inbreeding generations [39]. The observed levels of TRD in crosses that use DO females are consistent with presence of different alleles at the *distorter*(s) (Table 4.2 and Figure 4.6). Finally meiotic drive also explains the dramatic increase in WSB/EiJ allele frequency in eight generations of the DO.

I compared litter sizes in CC lines from the UNC breeding population that met the following criteria: 1) known to have a WSB/EiJ allele at *R2d2* in the G2:F1 generation; 2) genotyped in at least one later generation; 3) completely fixed for either the WSB/EiJ or non-WSB/EiJ *R2d2* allele in the most recently genotyped generation; and 4) more than five genotyped individuals showed fixture of the allele with no counter-evidence that the allele was still segregating. Those criteria yielded four CC lines fixed for the WSB/EiJ allele and eight lines fixed for a non-WSB/EiJ allele. There was no significant difference at any generation in litter sizes from the two sets of lines (Figure 4.9). Furthermore, I found no significant correlation (after multiple test correction) in any generation between litter size and either *R2d2* allele or CC line, and the directions of the correlations were not consistent from generation to generation.

Rescue of allelic diversity in the DO

Although the discovery and identification of TRD is an exciting development emerging from the DO pseudo-randomized mating scheme, the existence of such a locus could negatively impact the utility of the population for genetic studies. Hedrick [157] developed a mathematical model for the fixation of chromosomal variants that depended on four values: p, the wild-type allele freuqency; q, the variant allele frequency; s, the level of selection against the heterozygote; and m the segregation ratio of the variant allele in the heterozygote:

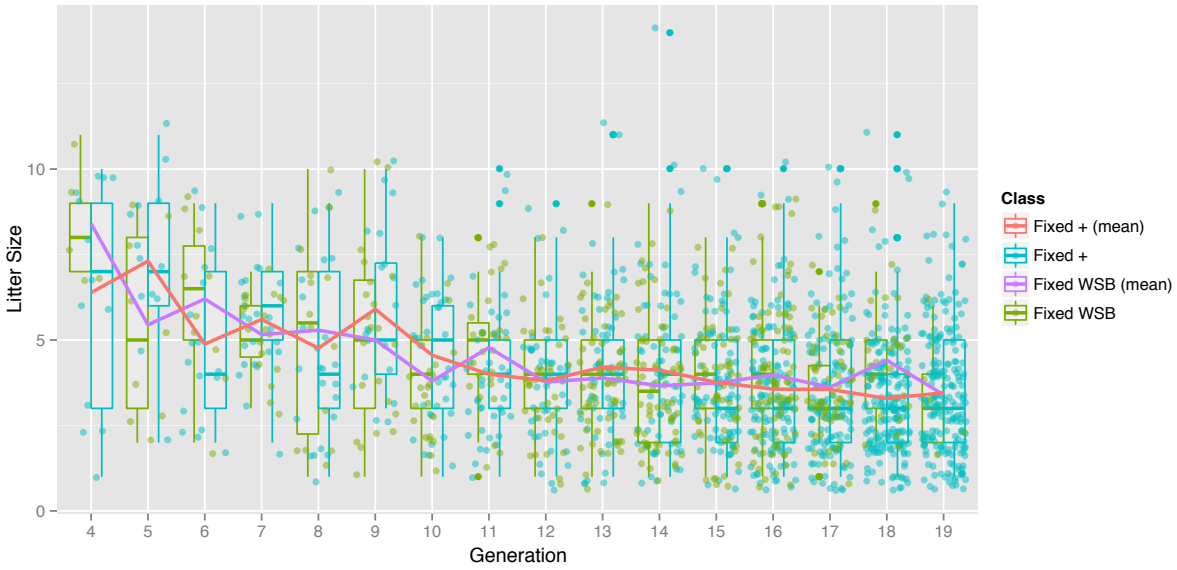


Figure 4.9: Litter sizes are not different between CC lines that fixed WSB/EiJ and non-WSB/EiJ alleles at *R2d2*. Box plots show litter sizes per generation for CC lines that ultimately fixed the WSB/EiJ allele (green) or a non-WSB/EiJ allele (blue boxes) at *R2d2*. Purple and red lines represent changes in mean litter sizes for lines that fixed WSB/EiJ and non-WSB/EiJ alleles, respectively. Initial decrease in litter sizes in all lines is due to inbreeding depression and is not correlated with any *R2d2* allele.

$$\Delta q = \frac{pq[s(4q - 2m - 1) + 2m - 1]}{2(1 - 2spq)}$$

I computed q (i.e., the frequency of the $R2d2^{WSB}$ allele) under this model given the allele frequency in the DO founder population (estimated at 0.188) and $s = 0$ for several different values of m (Figure 4.10). I found that the observed transmission ratios in the DO closely fit the model for a value of m between 0.66-0.68. Fixation is formally defined as an allele frequency of at least 0.99. Under Hedrick's model, $R2d2^{WSB}$ would become fixed after 33 generations. However, Hedrick's model considers an infinite breeding population while the DO only has a few hundred breeding pairs per generation. Therefore, the effects of genetic drift are expected to be much stronger and fixation would likely happen much earlier. In any case, fixation of $R2d2^{WSB}$ would happen much sooner than the 900 generations predicted by drift alone [180].

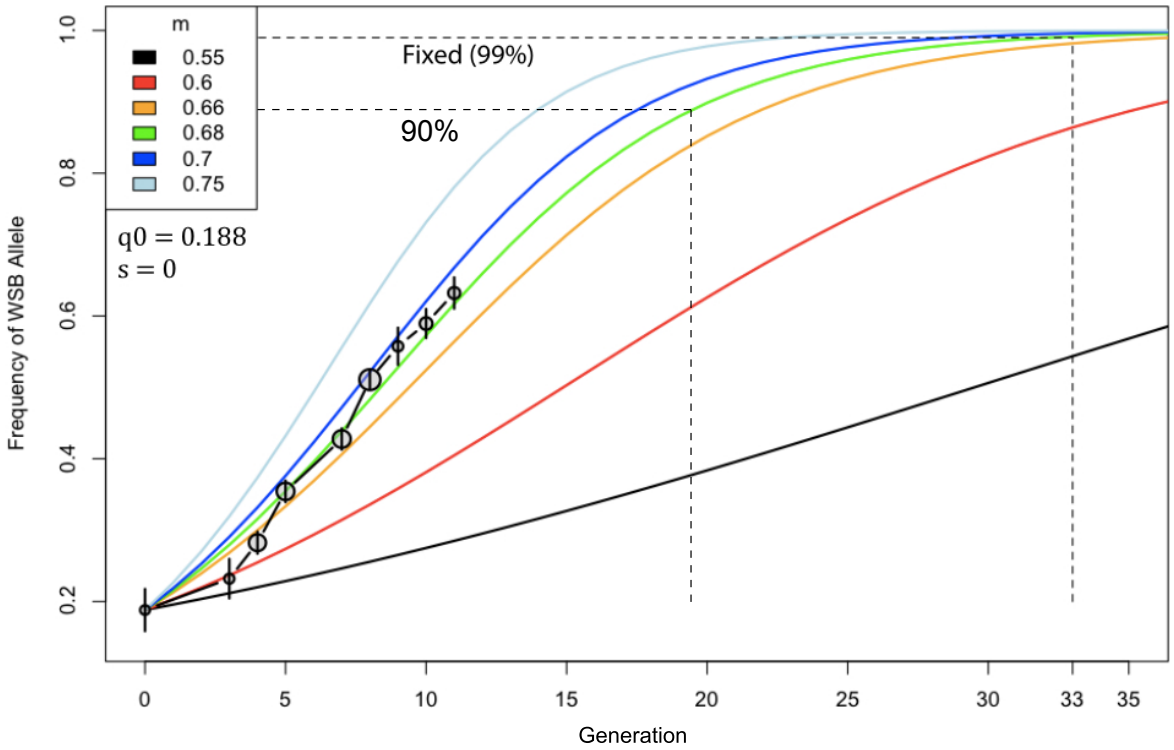


Figure 4.10: Fixation of $R2d2^{WSB}$ would occur much faster than predicted. Projections of Hedrick's equation across 35 generations of the DO for different values of m are shown as different colored lines. An initial value of q, 0.188, is estimated from the CC founder lines that comprised the DO founder population. A value of 0 is used for s due to the lack of any observable selection against heterozygotes (see next section). Lines and gray circles show the trajectory of $R2d2^{WSB}$ allele frequency in the DO over 12 generations. Dotted lines show time to fixation of $R2d2^{WSB}$, assuming m = 0.68, under A) Hedrick's model (0.99) and B) a more likely scenario that considers the greater effect of genetic drift in the limited DO breeding population (0.90).

Fortunately, the locus was discovered before complete fixation of the $R2d2^{WSB}$ allele. Although the candidate interval spans 9.3 Mb, TRD affects a much larger region in the DO because the strength of selection in favor of the WSB/EiJ allele is outpacing the rate at which recombination can degrade linkage disequilibrium in the region. Ultimately, this region would become an actual or statistical 'blind-spot' in the DO, such that the non-WSB/EiJ allele frequencies would become too small to detect allelic effects on phenotypic variation. Efforts are underway to purge the WSB/EiJ allele from the DO breeding population at this locus, rather than allow the region to become fixed. Using marker-assisted selection, progeny of heterozy-

gous WSB/EiJ carrier crosses are excluded from subsequent generations. Allele frequencies and random segregation on all other chromosomes are being preserved.

Selective sweep of *R2d2* in the DO

As discussed in the previous chapter, the ability of meiotic drive to give rise to a selective sweep in the absence of selection on organismal fitness has been hypothesized but never tested. To feasibly test the dynamics of a selfish sweep in the laboratory would require a multi-generational study in a system exhibiting a high level of TRD with minimal or absent confounding factors, such as lethality or sterility. Furthermore, such a study would require several years, and thus its importance would have to be clear in order to garner interest or funding. Arguably, the Ab10 system in maize meets the first two criteria, although it may be the lack of associated lethality or sterility that fail to make it important, at least from an agricultural standpoint. Fortunately, it was not necessary for us to prove the importance of the *R2d2* system. The DO provided an idealized, multi-generational study of the effect of a segregating, selfish gene on allele frequencies. All our data indicate that the $R2d2^{WSB}$ is undergoing positive selection in the CC and DO populations in the absence of any significant changes in fitness; in other words, a selfish selective sweep. The breeding scheme of the DO minimized the effect of ecological and sexual selection, and so the only available phenotype that may be affected to such a degree as to explain the observed TRD is fecundity. However, the almost three-fold increase in the frequency of the WSB/EiJ allele in the DO over the course of the 12 generations for which we had data occurred in the absence of any change in average litter size or variance in litter sizes (Figure 4.11).

While a selfish selective sweep has clear implications for the CC and DO, the larger evolutionary implications of selfish sweeps are less clear. On one hand, selective sweeps may be relatively rare. This appears to be the case for classic selective sweeps, at least in recent human history [208]. Furthermore, most selfish genetic elements are associated with negative phenotypic effects that, in all but exceptional cases, may negate the effects of positive selec-

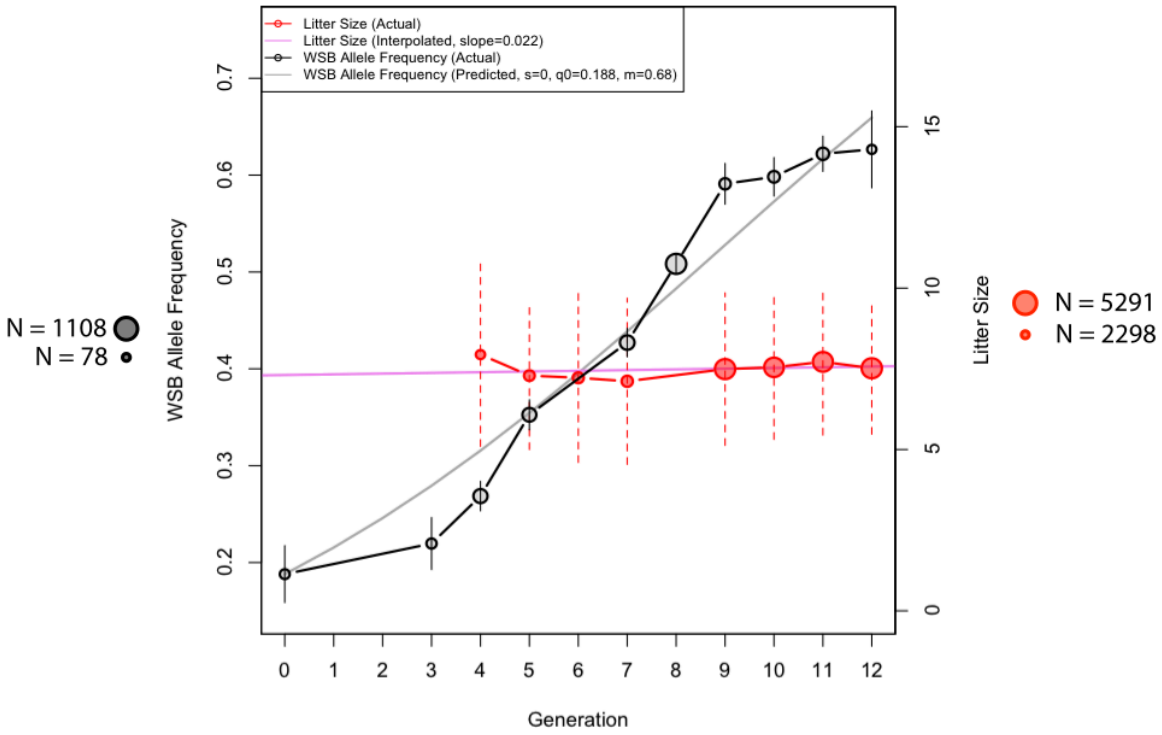


Figure 4.11: Selective sweep in the absence of changes in fitness. Per-generation WSB/EiJ allele frequencies (left y-axis and black dots) are shown compared to mean litter sizes (right y-axis and red dots). Black bars and red dotted lines indicate standard errors. Point sizes indicate sample sizes. Light gray and red lines show linear models fit to the allele frequencies and litter sizes, respectively.

tion. On the other hand, the fact that classic selective sweeps are rare does not necessarily bear on the frequency of selfish sweeps, since the latter may occur anywhere in the genome (and in fact probably occur preferentially in intergenic regions), whereas studies of classic selective sweeps only include coding and regulatory regions. The impending deluge of whole-genome sequence data will help to resolve the role of selfish sweeps in genome evolution.

4.6.6 Evolutionary origin of *R2d2*

The question of when, in evolutionary terms, *R2d2* arose is important, not only in understanding this particular meiotic drive system but also to gain a better understanding of meiotic drive in general. There are four distinct possibilities for the evolutionary origin: prior to sub-speciation in *M. musculus*, following the sub-speciation of *M. m. domesticus*, following the

colonization of the US east coast by *M. m. domesticus*, or during the development of the WSB/EiJ inbred line. This question may be answered by genetic characterization of house mouse populations on the US east coast; I consider that further under Future Directions.

We were able to genotype an inbred line, WSA, that was developed at the same time as WSB/EiJ from a closely related founder population but was never developed into a commercial line (Beverly Mock personal communication). I found that WSA had $\sim 93\%$ genome-wide genotype identity with WSB/EiJ. Within the *R2d2* region, WSA and WSB/EiJ were 100% identical. That suggested $R2d2^{WSB}$ was at least ancestral to the development of the WSB/EiJ inbred strain. The most surprising evidence we had regarding the origin of *R2d2* came from a single wild mouse from Turkey that we genotyped as part of the study in the previous chapter. That mouse had an identical genotype to WSB/EiJ across the entire *Rd2d* locus. On the other hand, the few wild mice we had from from the UK – the suspected geographic origin of mice on the US east coast – were no more similar to WSB/EiJ than any other mouse within the *R2d2* locus. That may mean that the $R2d2^{WSB}$ allele is present at low frequency throughout *M. m. domesticus*, or it may mean that the meiotic drive system originated from hybridization between two geographically distinct populations (Turkey and the UK).

R2d2 is the first observation of meiotic drive in favor of a wild-derived laboratory strain allele. This may have important implications for the study of phenotypes that arise from the interaction of divergent genetic backgrounds.

4.7 Materials and Methods

4.7.1 Ethics statement

All animal work was performed according to one of the following protocols: 1) the Guide for the Care and Use of Laboratory Animals under approved IACUC animal use protocols within the AAALAC accredited program at the University of North Carolina at Chapel Hill (Animal Welfare Assurance Number: A-3410-01); 2) the requirements of The Jackson Laboratory Animal Ethics Committees under approved protocol #JAX10001; 3) an animal protocol

approved by the North Carolina State University Institutional Animal Care and Use Committee (09-0133-B); or 4) animal study protocol approved by the NCI Animal Care and Use Committee (ASP# LCBG-013). All animals were euthanized according to the regulations of the governing protocol.

4.7.2 Published mouse crosses

The G2:F1 population has been previously reported and was genotyped on the MDA. A population of 96 (FVB/NJx(WSB/EiJxPWK/PhJ)F1)G2 mice was previously reported and was genotyped on the MegaMUGA array [198, 209]. DNAs from selected progeny from previously published (C57BL/6JxMus spretus)xC57BL/6J and (A/JxSPRET/EiJ)xA/J backcrosses [193, 200] were regenotyped on the MegaMUGA array.

4.7.3 New mouse crosses

Crosses 1-2, 7-10 and 16-17 (Table 4.1): WSB/EiJ and C57BL/6J were used in reciprocal combinations. Male F1 hybrids were backcrossed to C57BL/6J to produce the progeny of crosses 1 and 2. Female F1 hybrids were backcrossed to C57BL/6J to produce the progeny of crosses 16 and 17. The progeny of crosses 7-10 was produced in a similar way to crosses 16 and 17, except that female F1 of reciprocal matings of WSB/EiJ and CAST/EiJ were used for crosses 7 and 8, and female F1 of reciprocal matings of WSB/EiJ and PWD/EiJ were used for crosses 9 and 10. All breeding was done at the Jackson Laboratory (Bar Harbor, ME).

All other crosses: DO mice and standard mouse inbred strains (129S1/SvImJ, A/J, C57BL/6J, CAST/EiJ, FVB/NJ, NOD/ShiLtJ, NZO/H1LtJ, PWK/PhJ and WSB/EiJ) were obtained from The Jackson Laboratory (Bar Harbor, ME). CC mice were obtained from the Systems Genetics Core Facility colony at UNC Chapel Hill [83] (http://csbio.unc.edu/CCstatus/ index.py). Those mice were used to generate the following number and types of hybrid mice: nine (129S1/SvImJxWSB/EiJ)F1 females; two (A/JxWSB/EiJ)F1 females; seven (NOD/ShiLtJx WSB/EiJ)F1 females; six (NZO/HILtJxWSB/EiJ)F1 females; 10 (AU8042xCC001/Unc)F1

females; three (CC001/UncxOR15155)F1; nine (DOxCC001/Unc)F1 females and 13 (DOx CC005/Tau Unc)F1 females. F1 females were mated to FVB/NJ males and cages were surveyed three to five times per week. Litter sizes were recorded and pups were sacrificed at birth, and tissue was collected for DNA isolation. The same breeding schema was followed with 56 DO *R2d2* heterozygous females used to determine the origin of maternal TRD. All breeding was done at UNC Chapel Hill (Chapel Hill, NC).

4.7.4 DNA isolation and genotyping

Crosses 1-2, 7-10 and 16-17 (Table 4.1): DNA was prepared from spleens of 21-dpp old mice. DNA extraction and SNP genotyping were carried out as described previously [210].

All other samples: DNA for PCR-based genotyping was performed on crude whole genomic DNA extracted by heating tissue in 100ul of 25mM NaOH/0.2mM EDTA at 95°C for 60 minutes followed by the addition of 100ul of 40mM Tris-HCl. The samples were then spun at 2,000 rpm for 10 minutes and the supernatant collected for use as PCR template. All primers used in this study were designed using PRIMERQUEST software (https://www.idtdna.com/Primerquest/Home/Index). PCR reactions contained 1.5-2 mM MgCl2, 0.2-0.25 mM dNTPs, 0.2-1.8 μM of each primer and 0.5-1 units of GoTaq polymerase (Promega) in a final volume of 10-50 μL. Cycling conditions were 95°C, 2 min, 35 cycles at 95°, 55° and 72°C for 30 sec each, with a final extension at 72°C, 7 min. PCR products were loaded into a 2% agarose gel and run at 200 V for 40-120 minutes (depending on the marker). Genotypes were scored and recorded.

4.7.5 CC and DO haplotypes

An advantage of GRPs is that they make use of well-characterized founder strains, and thus phased haplotypes can be determined from genotype data everywhere in the genome. Because our samples were generated on multiple platforms, the most straight-forward method of analysis was to use haplotypes reconstructed from genotypes. Haplotypes had an additional

advantage of being more compactly represented (as a set of intervals) than genotypes, and so computations were faster and required less storage. MDA haplotype reconstructions had been done previously by Yi Liu [211]. MUGA and MegaMUGA reconstructions were done by the UNC Computational Genetics Group.

All CC lines bred at UNC were genotyped in the G2F1 generation, and most lines were genotyped in multiple additional generations. Genotyping was done on multiple platforms: MDA, MUGA and MegaMUGA. I obtained data for 3,977 arrays from 799 different CC lines. The DO has been widely adopted by the systems genetics community for the study of complex traits. Multiple researchers provided 5,022 DO samples that spanned 12 generations (3-14).

I performed QC on all samples and excluding 670 CC arrays and 423 DO arrays that had high rates of missing information and/or were outliers for heterozygosity or number of recombinations within their respective generations. For the CC, I only retained the 2,088 arrays for 499 lines bred at UNC, since those were the only lines for which we also had litter size data.

4.7.6 Estimation of embryonic lethality

DO and F1 females were euthanized by CO2 asphyxiation 12 – 18 days after delivery of the previous litter and the uterus was dissected. The number of live embryos and reabsorbed (dead) embryos was recorded. Each live embryo was dissected to isolate DNA for genotyping. Tissue from each live embryo was harvested for DNA extraction and genotyping.

4.7.7 Statistics

TR is reported as the ratio of the WSB/EiJ genotype to the total number of genotypes: WSB / (WSB + nonWSB). *P* values for aggregate data were calculated using a χ^2 goodness-of-fit test of the observed number of WSB/EiJ genotypes compared to the number of WSB/EiJ genotypes expected under the null hypothesis of equal transmission:

$$\chi^2 = \frac{(WSB - \frac{WSB+nonWSB}{2})^2}{\frac{WSB+nonWSB}{2}}$$

For small sample sizes, *P* values were instead calculated using an exact binomial test. Confidence intervals for TRs were calculated using the `BINOM R` package (Sundar Dorai-Raj unpublished). Average litter size was calculated as the mean number of offspring counted soon after birth per litter per female (± standard deviation), including the number of viable embryos counted in utero in mid-gestation DO females (unless otherwise noted).

Similarly, the average absolute number of pups inheriting each *R2d2* allele was calculated as the mean number of offspring per litter per female having each of the possible genotypes. Significance was determined using two-tailed Student *t*-test. Analysis of genotyping arrays All MDA arrays were genotyped using MouseDivGeno [41], and all MegaMUGA arrays were genotyped using Illumina® BeadStudio. We plotted number of H and N calls (as a fraction of the total number of genotypes) for each group of similar samples and excluded outliers from further analysis. For CC lines, DO animals, CCxCC F1 females and DOxCC F1 females, we inferred haplotypes using probabilistic methods [211, 198]. As an additional QC step, we grouped DO samples by generation and plotted the number of recombinations (counted as unique transitions in haplotype reconstructions) and removed outliers.

4.7.8 Linkage mapping of the *R2d2* expansion

CAST/EiJ allele in the CC G2:F1: Thirty-four MDA SNP probe sets were identified within the region corresponding to the *R2d2* expansion in the GRCm38 reference sequence. We ensured that these probes were unique using `BLAT` [212] to map them to the reference genome. In order to map the expansion allele present in the CAST/EiJ strain, the sum intensity of these probes as a phenotype and genotypes were coded as follows. First, we applied a CCS transform to the mean intensity of all probes in each probe set using MouseDivGeno [41] and summed the values for each sample. Next, the genome was divided into a set of disjoint intervals whose boundaries were defined by the 21,933 recombination events inferred in the

population [198], so that no individual would be recombinant within any of the resulting intervals. Then, using haplotype reconstructions, individuals were coded as either heterozygous (CAST/not-CAST) or homozygous (not-CAST/not-CAST) within each interval (there are no CAST homozygous individuals in this population). Of 474 individuals, 144 with a WSB/EiJ allele in the *R2d2* locus were excluded to yield a final sample size of 330. A single-locus QTL scan was then performed via Haley-Knott regression [213], treating the population as a backcross.

WSB/EiJ allele in an intercross population: Three MegaMUGA SNP probes were identified within the *R2d2* expansion region in the GRCm38 reference. Again, uniqueness was verified using `BLAT`. In order to map the expansion allele in WSB/EiJ, the sum intensity of these probes was used as a phenotype and genotypes were coded as follows. First the genome was divided into a grid of 1,000 disjoint intervals of approximately equal size, and one MegaMUGA SNP marker segregating between WSB/EiJ and PWK/PhJ was selected per interval. Individuals were coded as heterozygous (WSB/not-WSB) or homozygous (not-WSB/not-WSB) at each marker. A single-locus QTL scan was then performed using Haley-Knott regression as implemented in `R/QTL` [214], treating the population as a backcross.

4.7.9 Fine-mapping of the *R2d2* expansion

In order to refine the location of the *R2d2* expansion, we identified individual mice with recombinant chromosomes within the candidate interval defined by linkage mapping. These critical recombinants define the proximal and distal boundaries of the refined candidate interval.

CAST/EiJ allele: We partitioned the 330 G2:F1 individuals without a WSB/EiJ allele in the *R2d2* locus into two groups according to MDA sum-intensity values. From those with sum-intensity consistent with a non-CAST/EiJ expansion allele, we selected the most distal recombinants from CAST/EiJ to another haplotype. From those with sum-intensity consistent with the CAST/EiJ expansion allele, we selected the most distal recombinant from another

haplotype to CAST/EiJ. Together these recombinants define the proximal boundary of the expansion in CAST/EiJ. Similarly, in order to define the distal boundary, we selected the most proximal recombinants from CAST/EiJ to another haplotype that still had sum-intensity consistent with the CAST/EiJ expansion allele.

WSB/EiJ allele: The boundaries of the WSB/EiJ allele were mapped in the same fashion using 229 individuals spanning generations 10 through 14 of the DO, all of which have been genotyped on MegaMUGA and are recombinant for WSB/EiJ in the *R2d2* locus. We first excluded individuals homozygous for WSB/EiJ over any interval with in the locus. Then we selected the most distal recombinants from another haplotype to WSB/EiJ, which also had MegaMUGA sum-intensity values consistent with a non-WSB/EiJ expansion allele. These recombinants define the distal boundary of the expansion locus. We mapped the proximal boundary similarly.

4.7.10 Sequence variants and read depth

We retrieved BAM files of aligned reads (Oct 2012 release) and tables of all SNPs and small indels (Dec 2012 release) and structural variants (Feb 2013 release) within the initial candidate interval (76,860,362 – 86,117,205) from the Sanger Mouse Genomes Project FTP site (ftp-mouse.sanger.ac.uk). We counted a SNP as private to WSB/EiJ if that strain had a non-reference genotype that was different from the six other CC founder strains. We used the `mpileup` function of `samtools` [215] to output the read depth at each base. We defined the boundaries of the copy number expansion by identifying consecutive 100bp windows in which the average read depth was at least twice the genome-wide average read depth. We estimated the number of copies of the expansion as the modal per-base read depth.

4.8 Future directions

4.8.1 Molecular characterization of *R2d2* and modifier loci

The first step in molecular characterization of the *R2d2* locus is to determine its behavior during meiosis. Since we don't have the capability to perform such experiments in the Pardo-Manuel de Villena lab, we have established a collaboration with the Lampson lab at the University of Pennsylvania. We have provided the Lampson lab with mice from OR15155, a CC line that is homozygous for *R2d2WSB* and that exhibited drive when crossed to another CC line in our breeding experiment. The Lampson lab will cross OR15155 to B6 and obtain pre-meiotic cells from embryos. The will use fluorescent probes that hybridize to the copy number expansion to visualize in real-time the activity of the WSB/EiJ allele during meiosis.

4.8.2 Genetic characterization of *R2d2* in natural populations

The first step to determine the evolutionary origin of *R2d2* is to study the natural population of mice from which WSB/EiJ was derived. We have "crowd-sourced" the trapping of mice on the Delmarva (Delaware/Maryland/Virginia) peninsula, in radii centered on Centreville, MD. We are collaborating with Amanda Chunco-Ferris (Environmental Studies, Elon University) to establish contacts at the USDA and Maryland state agencies that are conducting wildlife studies that directly or incidentally involve the capture of house mice. We have also created a Facebook page to recruit hobbyist trappers. We prepared detailed guidelines for trapping, taking and preserving tissue samples, and recording relevant sample information. We send these instructions to trappers along with prepared sample collection tubes and a return mailing envelope to interested trappers.

We will extract DNA from tissue samples as they are returned to us and store them until we accumulate samples with sufficient geographic diversity. We will genotype these mice using either multiple PCR markers or MegaMUGA arrays, depending on funding availability. We will then compare the genotypes to WSB/EiJ to determine 1) if the wild mice have the same

genotype as WSB/EiJ within the *R2d2* and/or modifier loci, and 2) whether overall genetic diversity is lower within the *R2d2* and/or modifier loci compared to the rest of the genome. We are considering writing a small grant to fund this project.

We expect this experiment to be informative regardless of the results. There are three possible outcomes:

1. $R2d2^{WSB}$ is fixed in natural populations from Delmarva. This result may indicate a selective sweep of a novel mutation in the Delmarva populations. On the other hand, the mutation may have originated in the European population(s) from which WSB/EiJ is descended. If the later is true, it could mean that a selective sweep happened earlier (i.e., prior to the colonization of the US East Coast by *M. musculus*) or it could reflect a bottleneck event. It should be possible to differentiate between a selective sweep and a bottleneck by comparing the nucleotide diversity at *R2d2* with the rest of the genome.
2. $R2d2^{WSB}$ is segregating at intermediate frequency in natural populations. If the WSB/EiJ allele is universally present at an intermediate frequency, it must mean that there is some selective pressure against the allele in the wild, since the prediction from the mathematical model is that the allele should be fixed in a maximum of 35 generations. If the WSB/EiJ allele frequency is structured, such that it is at or near fixation in the central radii and at lower frequency more distant from the center, it could mean that a sweep is ongoing but progressing slowly due to limited gene flow, or it could mean that there is some mutation in the populations exhibiting TRD that promotes (or does not suppress) drive.
3. $R2d2^{WSB}$ is present at low frequency, or absent, in natural populations. This result has two possible interpretations. On one hand, it could mean that the WSB/EiJ allele is a mutation that arose during the development of the inbred line. On the other hand, it could mean that the WSB/EiJ allele is under neutral or weak negative selection in the wild but became fixed in the inbred line either by chance or because it was linked to a gene under artificial selection. In the later case, this result would mean that a modifier

is a necessary requirement for drive.

Experiments to further determine the origin of the $R2d2^{WSB}$ allele will involve trapping over a wider area and/or crosses between wild-caught and laboratory animals.

Chapter 5

CONCLUSIONS

I began my graduate studies with the aims of developing novel methods and technologies for genetic characterization of mouse populations, and applying those tools to the study of an important evolutionary system. Meiotic drive was a natural choice of system due to the expertise of our lab and the experience of some lab members in characterizing other meiotic drive systems [191, 216, 217]. The CRs of *M. m. domesticus* presented an exciting system because 1) meiotic drive had been suggested to play a role in the evolution of the chromosomal races but had never been tested [16]; and 2) the system had been widely studied, and many individuals had trapped large numbers of karyotypically abnormal mice along with nearby karyotypically normal controls. Early attempts to study meiotic drive in the laboratory were confounded by the fact that preferential segregation of metacentrics did not occur in crosses between wild mice and laboratory mice. There had been no previous attempt to look for genetic determinants of meiotic drive because two key factors were missing: a high-throughput and cost-effective platform for genetic study of mice, and an analysis pipeline. Therefore, I spent part of my graduate career developing those tools [40, 41, 42]. In addition, a considerable effort was required to coordinate the collection, processing and genotyping of a large number of samples from a dozen different collaborators. We became aware of the *R2d2* system late in my graduate career (December 2012), but the exciting possibility that the high level of TRD that we observed was due to meiotic drive prompted me to focus all my efforts on that project while we collected additional samples for the second phase of the GWAS of

chromosomal races. Through the contributions of many individuals, we have rapidly characterized the genetic components of a novel meiotic drive system. The advances that I and my collaborators have made in both of these systems have several implications. In this chapter I will first discuss the general implications of meiotic drive, and then discuss how our work fits into the current framework.

5.1 Implications of meiotic drive

For nearly 100 years, the synthesis of genetics and evolution has formed a theoretical basis for our understanding of how molecules, individuals and ecologies change over time. Intragenomic conflict in general, and meiotic drive in specific, fit into this system in complex ways, with each piece evolving in response to the effects of the others. Sandler and Novitski first introduced the concept of meiotic drive in 1957 [4], and since then a large body of theory has developed as to what types of effects meiotic drive could have on genes and karyotypes, both at a molecular and a population level, and what the long-term effects might be for species. Evidence is emerging to support those theories, although the pace may be slow due to the previously discussed challenges in observing and characterizing meiotic drive systems.

5.1.1 Changes in population allele frequencies

Population-level changes in allele frequencies are well described by population genetic models. Mutation and genetic drift introduce new alleles into populations, while inbreeding and adaptive selection reduce genetic diversity. Evolutionary dogma holds that the likelihood of a new mutation becoming established, increasing in frequency and even going to fixation within a population (selective sweep) is directly correlated with its effect on organismal fitness. However, the concept of intragenomic conflict raises the possibility of selfish sweeps, in which change in allele frequency and effect on organismal fitness are disconnected, or even negatively correlated.

5.1.2 Changes in centromere size and sequence

Meiotic drive has been proposed as an explanation for the centromere paradox: that, while many features of centromeres and the meiotic machinery are highly conserved, both centromeric repeat sequences and centromeric histone proteins evolve incredibly fast. A model in which centromeric drive leads to evolution of both centromeres and karyotypes was proposed independently by two different groups [16, 161]. That model is supported by several lines of evidence. First, assuming that the selective advantage of a centromere is due in part to greater attachment to the meiotic spindle, the centromeric drive theory predicts a trend toward larger centromeres. In fact, centromere size and level of TRD have been shown to be positively correlated in B chromosomes [218]. It is also expected that, since drive can only occur in an outcrossing (sexual) species, asexual species should not show the same tendency toward large centromeres. That has been shown to be the case in two closely related insect species: *Bacillus grandii* reproduces sexually and has a genome that consists of 15-20% centromeric repeats, while *B. atticus* reproduces asexually and has a genome with only 2-5% centromeric content [219]. There is also evidence that centromeres from outcrossed species are dominant over those from inbred species, such as in *Mimulus* [20]. It is important to note, however, that increases in centromere size are opposed by meiotic defects that occur when both spindle poles attach to different parts of the same centromere.

5.1.3 Changes in chromosome size and organization

DNA damage and mistakes in the machineries of DNA repair, DNA replication, recombination and chromosome segregation can lead to increases in chromosome size (repeat expansions, insertions and duplications), decreases in chromosome size (deletions) and chromosomal rearrangements (inversions, translocations, fusions and fissions). Large-scale mutations in chromosome size or structure are often deleterious because of the problems they cause during meiosis. Homologous chromosomes of different size or of substantially different sequence fail to pair, leading to recombination defects and nondisjunction. In addition, when

there are two homologous chromosomes of different length, the shorter has a transmission advantage [220, 221]. In hybrids between all-acrocentric mice and mice with homozygous Rb translocations, longer metacentric chromosomes had lower transmission rates than shorter ones [102]. In mammalian genomes, there is a well-known bias toward deletions relative to insertions [222]. Insertions and deletions tend to be small since larger changes have a greater likelihood of disrupting functional sequence. These biases against changes in chromosome morphology present another paradox: karyotypes are known to change relatively frequently over evolutionary time, yet theory holds that it should be very difficult, if not impossible, for a new chromosomal variant to become established. Intragenomic conflict resolves this paradox. Meiotic drive can overcome deleterious effects of new chromosomal variants if the level of selection in favor of the variant when in heterozygosity during female meiosis is more than twice as great as the selection against the variant due to its negative effect on fitness [157] (Figure 5.1). Differential selection of chromosomal variants during meiosis is well established [16].

Duplications and large insertions may occur by several mechanisms, most notably through the transposition and reverse transcription activity of parasitic DNA elements such as LINEs and SINEs. Transposable elements are also known to induce chromosomal rearrangement [223, 224]. Inversions serve to disrupt recombination and therefore may "lock in" large insertions, such as on Ab10 in maize and *t*-haplotypes* in mouse [225, 226]. Inversions that move the centromere closer to the middle of the chromosome were found to be selected for during female meiosis in flies [227, 228]. There is also mounting evidence that centromeres can change location in the absence of rearrangement, perhaps due to competition between established and *de novo* (neo)centromeres [229, 26].

5.1.4 Karyotype evolution and speciation

Over the past 100 years, karyotypes have been reported for thousands of eukaryotic species. In eutherian mammals alone, karyotypes vary between diploid numbers of 6 (India muntjac, *Muntiacus muntjac*) and 102 (*Tympanoctomys barrerae*, a South American rodent). The most

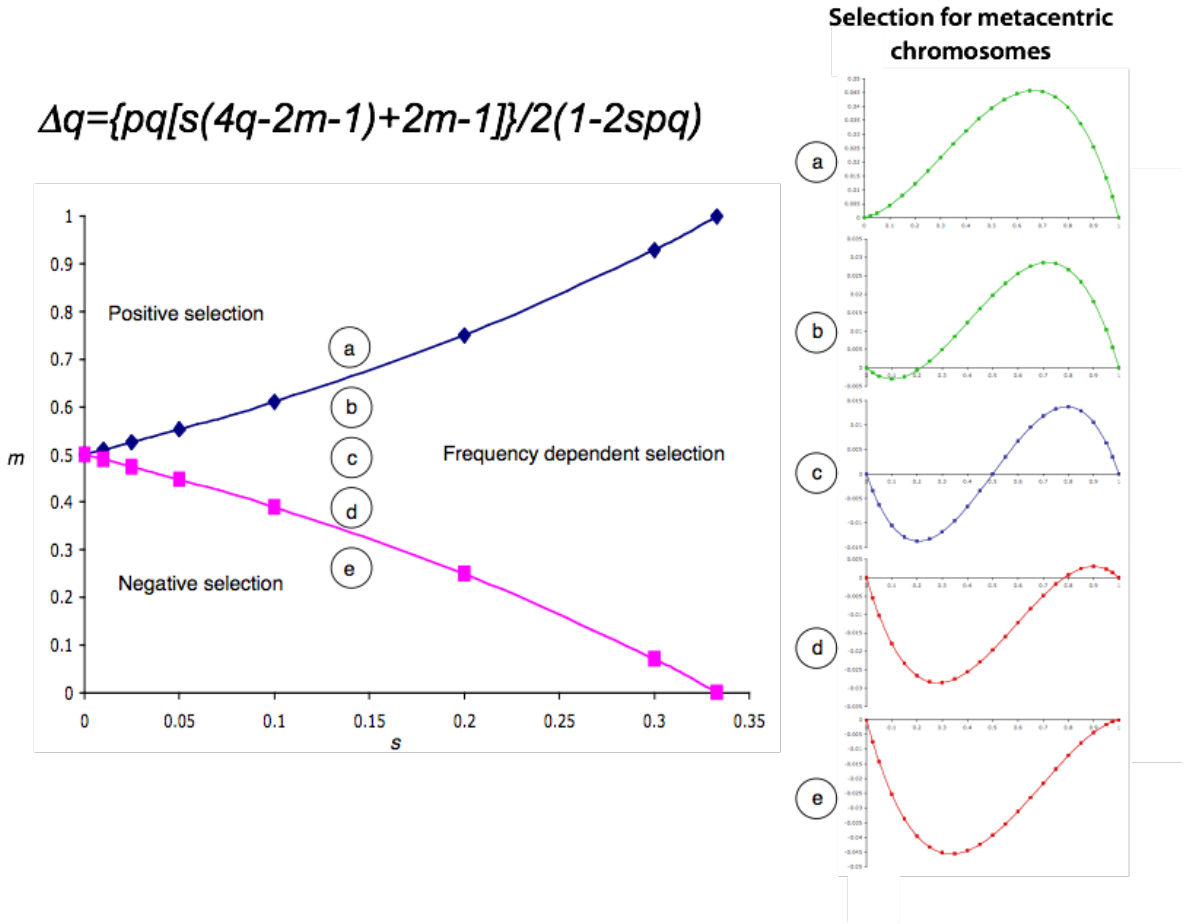


Figure 5.1: Fixation probability depends on opposing forces of selection. The change in allele frequency of a chromosomal variant is predicted by Hedrick's equation (see Chapter 4). The left panel shows how the selection coefficients (m, in favor of the new chromosomal variant, and s, against the heterozygote) interact to determine the change in allele frequencies. The panels at right show δq over time, where the allele frequencies are balanced at *t0* and a single allele becomes fixed at *t1*. Meiotic drive in favor of the chromosomal variant that is strong with respect to s leads to the fixation of the variant (a, b); weak, neutral or negative selection leads to the maintenance of the variant at an intermediate level (c) or elimination of the variant (d, e), depending on s. Adapted from [157], courtesy of Fernando Pardo-Manuel de Villena.

recent reconstructions of the ancestral eutherian karyotype agree on a diploid number of 46 [230], which is nearly identical to the mean 2N of eutherian mammals [16]. It is obvious from both the diversity of karyotypes and the wide deviations of extant species from the ancestral state that karyotypes have experienced rapid evolution. Even more striking is the fact that karyotype composition is not randomly distributed. Instead, karyotypes are overwhelmingly biased toward having a single type of chromosome – either all-acrocentric or all-metacentric – exactly the inverse of what is expected under a model of random karyotype evolution [16] (Figure 5.2). Furthermore, karyotypes show the same pattern within taxa; very few taxa have more species with intermediate karyotype compositions than with extreme compositions. When considered within a phylogenetic context, it is apparent that the direction of bias shifted multiple times over evolutionary history.

Both the bias toward extreme karyotype compositions and the frequent switching of bias can be explained by meiotic drive. Homologous chromosomes that differ in centromere number (fissions and fusions, including Rb translocations), position (inversions and centromere repositioning) or size (centromeric repeat expansion and mutation) represent functional heterozygosity at a locus that mediates attachment of a chromosome to the meiotic spindle (the third requirement for meiotic drive, see Chapter 1). White (1978) observed that each species appears to allow (and select for) certain types of chromosomal changes over others (karyotypic orthoselection), and it has been proposed that mammals exhibit karyotypic orthoselection for Robertsonian translocations [154]. When coupled with asymmetry of the meiotic spindle during female meiosis, chromosomal variants may experience strong and consistent preferential segregation leading to rapid fixation. It seems that karyotypes often experience punctuated equilibrium; that is, they are stable for long periods of time and then undergo rapid evolution toward a new equilibrium [231]. Just how fast these changes may occur is uncertain, but studies of knobs in maize and Rb translocations in the house mouse suggest that it is on the order of only thousands of generations in the former, and possibly decades in the latter [110].

Generally, meiotic drive of chromosomal rearrangements is observed to favor the dominant

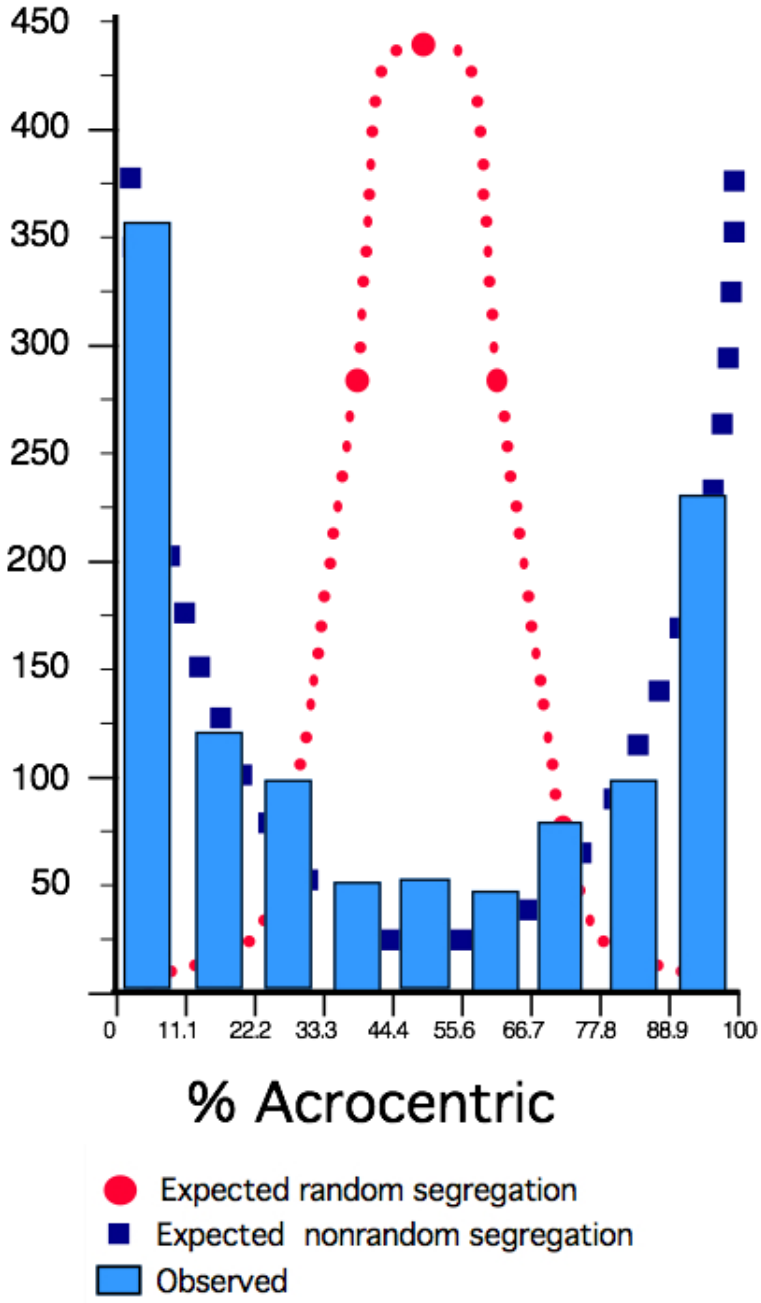


Figure 5.2: Distribution of acrocentric chromosomes in karyotypes of 1,170 mammalian species. The observed distribution (blue bars) closely matches the expectation from non-random segregation (blue dotted line), and is inverse of the expectation from random segregation (red dotted line). Adapted from [16], courtesy of Fernando Pardo-Manuel de Villena.

chromosome form in a species. The best example of this is the aforementioned difference in drive between mouse (all-acrocentric karyotype, drive in favor of acrocentrics) and humans (primarily metacentric karyotype, drive in favor of metacentrics). It is important to consider, however, that heterogeneity may occur not only at the level of homologous chromosomes but also in the polarity of the meiotic spindle (i.e., the direction of bias). When that is the case, a chromosomal variant that is selected against in a species may nonetheless be selected for within a sub-population of that species that is divergent for spindle polarity. The effect is amplified in species with low levels of drive and/or high rates of inbreeding. A mechanism determining spindle polarity is currently unknown, as is how such a mechanism might undergo frequent inversion. Sandler and Novitski (1957) hypothesized that drag (i.e., decreased fitness) resulting from meiotic drive itself may give abnormally high selective value to mutations that modify the nature of meiosis. Regardless of how they occur, these drastic and rapid shifts in karyotype composition are thought to be a key (and perhaps even the primary) catalyst of speciation.

5.1.5 Human population genetics and health

Rb translocations cause well known problems during meiosis, including increased rates of nondisjunction [232]. Several studies have found evidence of meiotic checkpoints that induce meiotic arrest and apoptosis in cells with synapsis and segregation errors [233], but those pathways are imperfect and may degrade with maternal age [234]. Rb translocations are the most common structural chromosome aberration in humans (0.12% of live births [235]). Rb translocations carried by females are subject to centromeric drive, leading to an increase in the rate of spontaneous abortion [236] and of offspring with trisomy disorders (Down, Edward and Patau syndromes). Studies in mice and humans have shown a high rate of non-disjunction during male meiosis that is associated with reduced fertility [237]. Cells with disregulation of *c-Myc*, an oncoprotein, had a higher incidence of Rb translocation formation, indicating a link between genome instability and chromosomal rearrangement [238].

A third of all human trisomies detected at miscarriage result from nondisjunction of chromosome 16 at maternal meiosis I [239]. Moreover, trisomy 16 does not show the dramatic increase with maternal age of other trisomies, suggesting that nondisjunction of this chromosome has a different etiology from that of other chromosomes. A large block of pericentromeric heterochromatin is variable in size within human populations [240], and there is strong evidence of recent selection spanning this centromere [241].

Sandler and Novitski made the interesting observation that non-random segregation could play an important role in human population genetic changes, even in the presence of high rates of embryonic lethality [4]. Fertility is not typically a limiting factor in human fecundity; in fact humans often take steps to limit the number of their offspring. Therefore, early embryonic lethality even in a large percentage of fertilizations will not ultimately effect the number of children a couple has. A successful strategy for a selfish human gene would be to sacrifice fertility to increase its likelihood of transmission.

5.2 Contributions of my studies

5.2.1 The Wild Mouse Genetic Survey (WMGS)

We have developed a diverse archive of DNAs from natural populations of *M. m. domesticus* that will facilitate studies of meiotic drive and of natural mouse populations. Our collection includes 44 CRs and multiple ST populations. We have genotyped a substantial number of those mice, along with controls from the other *M. musculus* subspecies and closely related species, creating a rich database of genetic diversity in the house mouse, the WMGS. Several collaborators are already exploiting this data set for diverse studies, and we expect that resource to lead to many important findings. We anticipate that the WMGS will grow as more studies take advantage of affordable mouse genotyping arrays (and eventually whole-genome sequencing) to study natural mouse populations.

Our collection of DNAs from a large number of DO females, which have been genotyped on MegaMUGA, and their offspring also constitute an important resource, both for further

characterization of *R2d2* and for other potential projects.

5.2.2 Characterization of meiotic drive using genotyping arrays

My studies demonstrate the continued utility and versatility of genotyping arrays despite the availability of higher resolution technologies, such as whole-genome sequencing. I and my collaborators have developed and/or adapted methods for using genotyping arrays to locate regions that exhibit TRD, identify QTL for multiple phenotypes related to non-Mendelian chromosome segregation and fine-map significant associations using relatively small sample sizes. In addition, I have demonstrated utility of genotype and intensity data for phylogenetic and population-genetic characterization of the populations being studied. I expect that others will build on these studies and further characterize both the CRs and the *R2d2* meiotic drive system.

5.2.3 Genetic control of meiotic drive

In most meiotic drive systems that have been described previously, there has only been evidence for the involvement of a single locus (the *responder*). It has generally been assumed that meiotic drive works like a toggle – on or off – determined by the *responder* allele. The Ab10 system in maize is an exception – four modifier genes have been identified that work in concert to promote the drive of the large knob. However, those modifiers are all tightly linked to the *responder*. The few reports of unlinked modifiers of meiotic drive have either been inconclusive [33] or have been disproven [242, 34].

The *R2d2* system provides strong evidence that both the presence and level of meiotic drive are subject to influence by unliked genetic factors. In addition, we have identified QTL associated with the accumulation of metacentric chromosomes in CRs. The fact that all 19 autosomes are involved in Rb translocations and yet there are only a small number of significant or suggestive QTL indicates that there are factors that act in *trans* to promote the segregation of metacentrics to the oocyte. I suggest that meiotic drive should be viewed as at least having

the potential to be a complex trait. Where possible, I encourage revisiting known meiotic drive systems to look for evidence of genetic control.

Meiotic drive as a complex trait

If meiotic drive is a complex trait, then what benefit to the organism is there from promoting the transmission of one type of chromosomal variant over another? And what is the possible underlying mechanism? One answer was proposed by Qumsiyeh [19], who suggested that karyotypic orthoselection was a result of selection for high vs. low recombination. In general, a high rate of recombination promotes genetic variation through the creation of new combinations of allelic variants, while low recombination tends to decrease variation. High genetic variability can be useful for a species that is expanding and adapting to new and different environments (such as has happened in the mouse over the past 250,000 years, and most dramatically within the past 10,000 years), whereas low genetic variability is useful for adapting to specialized environments, such as is the case for at least some of the CRs, which tend to be located in isolated areas such as islands and mountain valleys. Selection for a particular level of recombination also selects for particular types of chromosomal variants: high recombination tends to increase diploid number and low recombination tends to decrease diploid number. Therefore, recombination rate may be a fitness characteristic in a species or subpopulation, and thus subject to selection, with karyotypic orthoselection as a byproduct. Variants that promote or suppress recombination may be selected for in certain environments, and selected against in others, which explains why multiple modifiers might be segregating in a population. As to the mechanism, modifiers of recombination are known in multiple organisms, such as the recombination enhancer of the Ab10 system and *Prdm9* [243] in the mouse. These and other known QTLs [244] are the first candidates to examine for an involvement in meiotic drive.

In contrast to the model of Qumsiyeh, the causal arrow may point in the other direction: modifiers may use chromosomal variants as proxies to promote or suppress recombination.

For example, Rb translocations have a negative effect on recombination [245]. It has also been shown that the level of recombination is positively correlated with fundamental number (the number of chromosome arms in a karyotype) [246]. While Rb translocations do not affect the number of chromosome arms, other types of rearrangements do, such as centromere repositioning. Modifiers may also promote/suppress other features of chromosome morphology, such as chromosome size or number of centromeres [11]. In this model, the underlying mechanism would have to be a protein involved in either the attachment of the chromosome to the meiotic spindle or the control of spindle orientation and attachment to the cellular cortex. Further characterization of the molecular components of female meiosis will be required to uncover a candidate protein.

A third model to consider is centromeric drive [161], in which centromeric repeat sequences and centromere-associated proteins co-evolve. In that model, centromeric proteins may be considered modifiers acting on specific sequences, or on epigenetic features that arise from the content, orientation or organization of sequences. Genes that code for centromeric proteins could enhance their own transmission by influencing the segregation of the linked centromere to the oocyte, while unlinked genes would benefit by suppressing drive. Alternately, a modifier may be a protein that promotes or suppresses instability of the centromere. More frequent "breakage" in or near acrocentric centromeres would encourage a greater rate of Rb translocation formation or other types of rearrangements. One class of elements found near centromeres that are often associated with genomic instability are transposable elements [247].

5.2.4 The time-scales of meiotic drive systems

A final observation from the two systems that I studied is that meiotic drive appears to occur over a broad range of time scales. There is general consensus that the karyotypic diversification of *M. m. domesticus* has occurred within the last 10,000 years. While the relative ages of the various CRs have not been precisely determined, there is evidence that new races

can arise over extremely short time periods [110]. In addition, some systems, such as the races near Barcelona, appear to be in flux and lack clearly defined CRs.

In the DO, I have demonstrated that meiotic drive can give rise to a selective sweep with the ability to eliminate genetic variation in a large genomic region over a very short time period (33 generations or less). On the other hand, *R2d2* allele associated with meiotic drive appears to be segregating within wild *M. m. domesticus* mice, and to be independent of karyotype or geography, which suggests that it is a very old mutation. This may be supported by the presence of the allele in SPRET/EiJ, although it seems more likely that the allele was transmitted to that strain by recent introgression from *M. m. domesticus*.

In summary, our results suggest that meiotic drive has the potential to be a powerful force for reshaping genomes and karyotypes across evolutionarily short time-scales.

5.3 Future applications of our work

This chapter and the Future Directions outlined in the previous two chapters indicate the important implications of this work. We have provided the first mammalian model of meiotic drive that may be characterized at a molecular level. First, the existence of modifier loci means that the system is perturbable, enabling a systems genetics approach to characterization. Second, the model exists within an inbred resource of essentially unlimited size. Techniques for mouse reproductive biology are well developed; it is possible to super-ovulate female mice to obtain large quantities of oocytes. Second, lethality does not play a substantial role in the observation of drive, and so normal-sized litters of offspring with both allelic types are available. Third, the strength of TRD means that relatively small sample sizes are required to determine the presence of drive in test crosses.

I expect that this work will contribute to several fields beyond the study of the genetics of meiotic drive. First, this work lends support to the role of meiotic drive in the evolution of karyotypes, especially centromeric drive. If molecular characterization reveals that *R2d2* functions as a centromere during meiosis, then our model will be important for study-

ing the formation of neocentromeres and the interaction of neocentromeres with wild-type centromeres, and it may help to identify new proteins involved in meiotic spindle formation, regulation and interaction with chromosomes. If *R2d2* is not functioning as a centromere during meiosis, then we have discovered a novel mechanisms by which a non-centromeric locus can influence the meiotic segregation of centromeres. Second, an understanding of how and why metacentrics are preferentially segregated to the oocyte during human female meiosis may suggest preventative measures against improper chromosome segregation. Finally, there are several exciting potential biotechnological applications of this work. In combination with genome engineering tools, a strong mammalian meiotic drive system could be exploited to quickly introduce new genes into an existing inbred or outbred population. For example, new congenic panels could created in 1-2 generations rather than 10-20 by random incorporation of a locus under strong drive. Furthermore, the requirement for modifiers means that meiotic drive can be made inducible. Additionally, a greater understanding of neocentromeres may enable the creation of stable, *de novo* mammalian artificial chromosomes (MACs). Engineered MACs that are also capable of being driven to fixation would constitute an ideal, mitotically stable vector for model organism genetic engineering or, potentially, human gene therapy. Below I discuss one final ongoing collaboration that could have important future applications.

5.3.1 Genetic control of invasive mouse populations

Islands are our greatest reservoir of biological diversity. Islands represent only 3% of the earth's land area, but they harbor 20% of all known species. Unfortunately, island species are also under the greatest threat of extinction. Fifty percent of all endangered species are on islands, and 80% of all known extinctions have occurred on islands. Island bird populations are under particular threat: 95% of all known bird extinctions have occurred on islands (http://www.islandconservation.org).

Invasive species are one of the greatest threats to endangered island species, especially

rats and house mice. It has been estimated that 60% of all island animal extinctions have been caused at least in part by invasive species. Invasive species may pose direct threats, as predators of endangered species, or indirect threats, for example by outcompeting native species for food or other resources.

Control of invasive species is often difficult or impossible using currently available methods. The most common method of eradicating invasive rodent populations is the direct application of rodenticides such as Warfarin. This approach is only feasible, financially and logistically, on relatively small islands. Often, native species must be removed prior to the application of rodenticide to prevent collateral damage. This approach has a high rate of failure due to the ability of only a few survivors to repopulate the island in a relatively short time. A second approach whose negative side-effects are well known is the introduction of a second non-native species to eradicate an existing invasive species.

The Farallon Islands are a timely and important case study in the control of invasive mouse populations. The Farallons are a small (85.4 ha) group of islands off the coast of San Francisco (CA, USA). The Farallon Islands were discovered and named in 1539 by Spanish explorer Juan Rodriguez Cabrillo. They have since been recognized as a breeding ground for 14 species of seabirds, several of them endangered. There are no native land animals; at one time seals frequented the island but were hunted to extinction, though they have begun to recolonize the islands in the past 20 years. There are no human settlements save for a research station, and visitation is highly restricted. At some point, house mice were introduced to the Farallon Islands and quickly expanded to become the densest rodent populations in the world (http://www.restorethefarallones.org/learn/) due to the absence of predators (all of the birds on the island exclusively eat fish and other sea life). Burrowing Owls, native to mainland California, has discovered the abundant food on the island and has begun to migrate there. The owls also feed on the eggs and chicks of the native bird species. The Ashy Storm Petrel has been particularly sensitive to the vagrant owls and is considered highly endangered. The state has proposed eradicating the house mouse population by saturating the island with rodenti-

cide pellets. That approach is strongly opposed by environmental groups, and the two sides are currently at a stand-still.

In 2012, a summit was held at NC State University in which experts in island conservation, invasive species control, and mouse genetics came together to discuss genetic solutions to the problem on the Farallon Islands. Our lab attended the meeting, and we proposed to characterize the genetic background of the Farallon mice by array genotyping. We determined that the mice are almost completely *M. m. domesticus*, and of a mixed European origin, perhaps representing multiple colonization events (Didion, Threadgill, FPMV unpublished). We were particularly concerned that the Farallon mice may harbor a warfarin-resistant allele of *Vkorc1*, which is known to be segregating in European mice and was suggested to have originated from inter-specific mating with *M. spretus* [205]. However, sequencing of the *Vkorc1* gene showed that the Farallon mice do not have the resistant allele.

Following the discovery of the *R2d2* meiotic drive system, we turned our attention to the possibility of exploiting the system to introduce a sterility-associated gene, such as a foreign allele of *Prdm9* [243]. Such a model would involve linking the deleterious gene to *R2d2 in vitro* in a cell line derived from a driving CC strain. The Threadgill lab has already begun to create ES cell lines from CC founder strains and CC lines. The ES cells would then be microinjected into blastocysts in a pregnant female of the same CC line, or into an outbred (DO) background; outbred animals may be better suited to reproductive competition in a natural environment. The engineered animals would be released onto the island, with the expectation that the *R2d2*-linked allele would spread rapidly through females and eventually lead to a dearth of fertile males.

A similar release study (although not done with transgenic mice) was conducted by Berry and colleagues in the early 1980s [158]. They released 77 mice from Eday, one of the Orkney Islands in northern Scotland, onto the Isle of May, off the east cost of Scotland near Edinburgh. Both populations were *M. m. domesticus*, but were genetically and karyotypically distinct. Most notably, mice from the Isle of May had the standard 40 chromosome karyotype, and

mice from Eday were fixed for three Rb translocations (2n = 34) [158]. Follow-up studies showed that the Eday mice interbred with the Ornkey mice and that the Rb translocations had become established at high frequency.

Chapter 6

APPENDIX A: GENOTYPING ARRAY METHODS

A-1 DNA isolation and array processing

DNA for array processing was isolated as described previously [40, 39, 198, 209]. All sample preparation was performed by our laboratory technicians. MDA processing was done either at the UNC Genomics Core or at the Jackson Laboratories. MUGA and MegaMUGA processing was performed by Neogen/GeneSeek (Lincoln, NE).

A-2 Normalization

Intensity values are generated from a mixture of true hybridization signal and noise. Noise has multiple sources: 1) variability in DNA quantity in unsynchronized cells; 2) variability in DNA quantity or sequence due to mosaicism; 3) contaminants in the sample; 4) biases or other errors during DNA extraction, fragmentation and amplification; 5) low or non-uniform molecular weight of degraded DNA; 6) non-specific hybridization (i.e., hybridization of DNA to a non-complementary probe); 7) probe-specific biases due to characteristics of the DNA sequence; 8) inherent biases in the hybridization or fluorescence chemistry; 9) non-uniform DNA concentration or quality across samples in a batch; 10) within-batch array processing errors due to equipment malfunction or improper calibration; 11) variability in sample handling or array processing across batches; and 12) non-uniqueness of a probe sequence (i.e., a probe matches multiple places in the genome).

Given that the purpose of using a SNP array is to determine the correct genotypes of a sample, it is not necessary to determine which noise factors are contributing and to what degree, but only to subtract as much noise as possible so that the intensity value from each probe is highly correlated with the copy number of that probe sequence in the sample DNA. This correction process is called normalization, since it typically takes the form of fitting each probe and each sample to a normal distribution. By default, the array processing software provided by manufacturers of the microarrays and scanning hardware (Affymetrix, Illumina) performs several platform-specific normalizations to correct for array-specific noise (sources 1-5) and probe-specific noise (sources 6-8). In addition, a set of reference arrays can be used to correct for batch-specific noise (sources 9-11), although platforms differ as to whether this is a default or optional step. Despite these steps, all sources may generate random noise for which it is difficult or impossible to correct. Genotyping methods (discussed next) are therefore designed to be noise-tolerant.

Third-party normalization steps may be necessary for certain applications such as CNV calling. For example, the `PennCNV` software [248] requires correcting for "genomic waves," a poorly understood effect related to C+G sequence composition. The `tQN` software [249] specifically addresses "dye bias" in Illumina arrays, a systematic effect caused by inherent differences in the two fluorophores that is not addressed by Illumina's `BeadStudio` software.

A-3 Clustering and genotyping

Genotyping is the process of converting continuous intensity values into discrete genotype values. Traditional genotyping methods are concerned with classifying samples into one of three categories for each probe: homozygous A (AA), homozygous B (BB) or heterozygous (AB), where A and B represent the two alleles of the SNP. Samples that do not fit into any of the three categories are uncalled (no-call, N), and are typically treated as missing information in downstream analyses. The computer science field of machine learning has produced

several clustering methods that have been applied to genotyping. Most of these methods are fully or partially supervised, meaning they require some guidance from the user, such as a set of reference arrays [63] or sample-specific parameters that must be guessed by the user. `Alchemy` is an example of a model-based algorithm that performs well on small batches and in the absence of heterozygosity, which is particularly important for genotyping inbred model organisms such as the mouse [64].

A-4 Quality control

I have developed quality control (QC) standards for evaluating the genotype accuracy of each array. First I count the numbers of heterozygous (H) and N calls. Second, I compare the distribution of sum intensities to a reference distribution using a Kolmogorov-Smirnov test. Sum intensity is the total intensity for a SNP, which is calculated by summing the intensities for the two alleles (typically abbreviated as A+B for MDA and X+Y for MUGA and MegaMUGA). The reference distribution for MegaMUGA is based on a batch of 384 samples representing many genetic backgrounds. Although I have empirically identified thresholds for these measures, I find that the best results are achieved by analyzing similar samples jointly and identifying outliers. I achieve the best results by first excluding SNPs that perform poorly, either universally or within groups of similar samples.

I use several additional QC measures when available or applicable. If I have replicate samples, I test the genotype concordance between them. Every sample of an inbred line can be considered a replicate of the same genome. When I work with outbred samples, I also test SNPs for significant deviations from Hardy-Weinberg equilibrium (HWE); inbred strains violate the prerequisites of random mating, large effective population size and no selection. SNPs that are not in HWE are important for further investigation because they signal population structure. SNPs that are diagnostic for a population (i.e., present at higher rate of homozygosity than predicted by HWE) may be associated with loci that are under selective pressure, and thus may be relevant to the phenotype being studied.

Other measures are detailed in the following sections. I have developed an `R` package, `megamugaQC`, that automates all of these QC measures for MegaMUGA arrays (unpublished). I have also developed a separate package, `CLASP`, that extends the methods of genotype reproducibility testing to the validation of mouse cell line identity (Didion et al. in prep).

A-5 Copy number analysis

For any given probe, hybridization intensity is strongly correlated with the number of copies of the target sequence that are present in the sample. Individual probes are generally not a reliable measure of copy number for a larger region due to the possibility of non-specific hybridization or non-uniqueness of the probe; however, joint analysis of multiple contiguous probes can provide a reliable estimate of copy number. In fact, this is how CNV and comparative genome hybridization (CGH) arrays work. The primary difference between SNP and CVN/CGH arrays are that the later employ probes that are invariant. Sum intensities for SNP array markers are reliable substitutes for invariant probes. Furthermore, SNP arrays provide allele-specific information that can be used to determine the source of unbalanced gain and loss events [250]. The MDA has been used for CNV analysis with `PennCNV` [248]. I have found that the Genome Alteration Print (`GAP`) algorithm [88] produced the best results for MegaMUGA. `GAP` also provides for analysis of inbred samples, whereas most methods rely on a high rate of heterozygosity since they have been designed to work with human data.

A-5.1 Sex determination

Genetic sex is determined by Chr Y presence (male) or absence (female). Chr X copy number is a less reliable determinant of genetic sex because females may possess a single X-chromosome (XO) and males may possess two X-chromosomes (XXY). Non-standard sex chromosome copy numbers are also associated with phenotypic differences (e.g. Turner syndrome, Klinefelter syndrome) that may be important to account for. In principle, sex chromosome copy number determination is just a special case copy number analysis, however in

practice it is more complex. Two baselines must be determined for both Chr X and Chr Y – male and female. Hybridization intensity and copy number are not related linearly; therefore, the single-copy baseline cannot be determined by linear interpolation from the two-copy baseline, and the zero-copy baseline has a non-zero intensity value due to background noise. Generally, larger (and sex-balanced) sample sizes are needed to determine the sex chromosome baselines than the autosomal baselines. Special consideration is also needed for Chr Y because markers in the pseudo-autosomal region (PAR, a region of homology between the Chrs X and Y where recombination occurs in males) will have elevated intensity value since they hybridize to both Chrs X and Y. I developed an algorithm that accurately predicts sex chromosome karyotype, and therefore genetic sex (unpublished). The algorithm uses a mixed model based on sum intensity in a moving window, and also heterozygosity information when present.

A-6 Phasing and imputation

Phasing is the process of inferring the two parental haplotypes given only genotype information (I refer to this interchangeably as haplotype reconstruction). Phasing is trivial for inbred samples because the haplotypes are identical to each other and to the genotypes. Phasing is more complex but still highly accurate in individuals for whom the founder haplotypes are known, such as parent-child trios, inbred strain intercrosses, RILs and even outbred populations such as the DO. In samples where the founders are not known, such as traditional outbred stocks and wild-caught mice, phasing is a probabilistic process that depends on a catalogue of reference haplotypes or population-specific allele frequencies for the alleles of the marker set being phased. Phasing is generally limited to haplotype blocks, within which there is no evidence of a recombination event.

For the CC and DO populations, two haplotype reconstruction methods are available (Gatti and Churchill, unpublished and [211]). Both cluster sample intensities in order to assign probabilities to the 36 possible founder states at each marker. The methods differ slightly

in their details: DO/QTL performs clustering based only on the samples in the batch being genotyped, and it uses transition probabilities based on the expected number of recombination breakpoints (which can be estimated from the sample type and generation); the method of Liu et al. uses clusters pre-computed from a large number of reference samples, and does not use sample-specific transition probabilities. For wild-caught mouse samples, I use two different probabilistic methods, fastPHASE [132] and IMPUTE2 [133]. The accuracy of phasing in the current study is limited because I have used the simplest but least accurate method, which conditions each sample on all other samples. Both softwares provide similar results. We currently have a collaboration with Elodie Gazave at Cornell University to identify reference haplotypes from homozygous regions of wild-caught and wild-derived samples that should improve results in the future.

Imputation is a probabilistic process coupled to phasing that attempts to infer missing genotypes. For example, if the two parental haplotypes are inferred across ten markers given child genotypes for nine of those markers, the child genotype at the tenth marker can be inferred from the parental haplotypes. Whole-genome imputation was performed on 100 commonly used inbred strains [251, 42] that were genotyped on MDA [40] based on 17 strains sequenced by the Sanger Institute [56]. I use imputation as a pre-processing step for association studies in wild mice (discussed later).

A-7 Relatedness

For any given marker, a pair of samples may share zero, one or two alleles in common. Allele sharing may be due to common ancestry (identiy by descent, IBD) or to chance (identity by state, IBS). For an outbreeding population, the probability of allele sharing due to IBD can be estimated for any degree of relationship. For example, the probability of monozygotic twins sharing two alleles IBD is always 1.0, whereas the probability of two unrelated individuals being IBD for any alleles is 0.0. The standard measures of relatedness are 1) kinship (k or θ), which is the probability that two alleles – one chosen at random from each individual –

are IBD, and 2) coefficient of relatedness (*r*), which is the expected fraction of shared alleles that are IBD. In an unstructured population, these values are directly related ($r = 2 * k$) and the expected values for different degrees of relatedness are well known. For structured populations, the expected values are either obvious, such as for inbred lines ($k = r = 1$), or must be estimated. I determined the expected values for the DO from a simulation using the `ValBreed` software (Will Valdar unpublished).

Actual *k* and *r* values may be derived from a pedigree. However, pedigrees are generally not available for outbred mice. In principle, a pedigree could be determined for each DO individual; in practice, this task is not feasible for the staff at the Jackson Laboratory. Alternatively, the degree of allele sharing between a pair of individuals can be estimated from genotypes at a set of SNP markers. In the CC and DO, this calculation is made simpler and more accurate by comparing haplotype blocks rather than individual genotypes. The relationship between the two individuals can be determined by comparing the actual and estimated fraction of shared alleles for different degrees of relatedness.

For samples from unstructured populations (wild-caught individuals), I use the method of Stevens et al. (2011), which is based on calculations by SNPDuo [252]. This method is based on the IBS2* ratio [253]: $\frac{IBS2*}{IBS2*+IBS0}$, where IBS2* is the fraction of heterozygous-concordant SNPs (i.e., both samples are heterozygous) and IBS0 is the fraction of homozygous-discordant SNPs (i.e., each sample is homozygous for a different allele). The expected value of IBS2* for unrelated individuals is 2/3, and a two-sided Z-test can be used to determine the significance of deviation from the expectation. Pairwise comparisons that are significant after multiple test correction are flagged as possibly related.

A-8 Tree reconstruction

Phylogenetic tree reconstruction methods are well established [84]. Tree reconstruction can have two purposes: establishing the relationships between individuals (topology) and establishing the distance between individuals (divergence time estimation). Most of my anal-

ysis only require topological information. Therefore I typically use the computationally simple neighbor-joining algorithm. Neighbor-joining trees are constructed by first computing a pairwise distance matrix and then iteratively finding the pairs of taxa with the shortest joint pairwise distance. Thus, the tree is constructed in a bottom-up fashion. The robustness of each branch can be estimated by bootstrapping, which means running a large number of iterations of the tree construction algorithm (typically 100 or 1000), collapsing the results into a consensus tree, and then counting how many times each branch in the consensus tree appears in the set of permuted trees.

REFERENCES

[1] T. H. Morgan, C. B. Bridges, and A. H. Sturtevant, *The genetics of* Drosophila melanogaster. Garland Publishing, Inc, Aug. 1925. 2

[2] S. Gershenson, "A New Sex-Ratio Abnormality in *Drosopholia obscura*," *Genetics*, vol. 13, pp. 488–507, Nov. 1928. 2

[3] F. Pardo-Manuel de Villena, E. de la Casa-Esperon, T. L. Briscoe, and C. Sapienza, "A genetic test to determine the origin of maternal transmission ratio distortion. Meiotic drive at the mouse *Om* locus," *Genetics*, vol. 154, pp. 333–342, Jan. 2000. 2

[4] L. Sandler and E. Novitski, "Meiotic drive as an evolutionary force," *American Naturalist*, pp. 105–110, 1957. 2, 134, 141

[5] B. C. Turner and D. D. Perkins, "Spore killer, a chromosomal factor in *Neurospora* that kills meiotic products not containing it," *Genetics*, vol. 93, no. 3, pp. 587–606, 1979. 2

[6] M. F. Lyon, "Transmission ratio distortion in mouse *t*-haplotypes is due to multiple distorter genes acting on a responder locus," *Cell*, vol. 37, pp. 621–628, June 1984. 2, 115

[7] A. Burt and R. Trivers, *Genes in Conflict*. Harvard University Press, Mar. 2006. 2, 11

[8] R. B. Nicklas, "How cells get the right chromosomes," *Science*, vol. 275, pp. 632–637, Jan. 1997. 2

[9] M.-H. Verlhac and A. Villeneuve, *Oogenesis: The Universal Process*. John Wiley & Sons, Mar. 2010. 2, 4

[10] H. V. Crouse, "The Controlling Element in Sex Chromosome Behavior in *Sciara*," *Genetics*, vol. 45, pp. 1429–1443, Oct. 1960. 3

[11] F. Pardo-Manuel de Villena and C. Sapienza, "Nonrandom segregation during meiosis: the unfairness of females," *Mammalian genome*, vol. 12, pp. 331–339, May 2001. 4, 5, 144

[12] G. M. Hewitt, "Meiotic drive for B-chromosomes in the primary oocytes of *Myrmeleotettix maculatus* (Orthopera: Acrididae)," *Chromosoma*, vol. 56, pp. 381–391, July 1976. 5

[13] R. LeMaire-Adkins and P. A. Hunt, "Nonrandom segregation of the mouse univalent X chromosome: evidence of spindle-mediated meiotic drive," *Genetics*, vol. 156, pp. 775–783, Oct. 2000. 6

[14] H. S. Malik and J. J. Bayes, "Genetic conflicts during meiosis and the evolutionary origins of centromere complexity," *Biochemical Society transactions*, vol. 34, pp. 569–573, Aug. 2006. 7

[15] F. Pardo-Manuel de Villena and C. Sapienza, "Transmission ratio distortion in offspring of heterozygous female carriers of Robertsonian translocations," *Human Genetics*, vol. 108, pp. 31–36, Jan. 2001. 8

[16] F. Pardo-Manuel de Villena and C. Sapienza, "Female meiosis drives karyotypic evolution in mammals," *Genetics*, vol. 159, pp. 1179–1189, Nov. 2001. 8, 51, 52, 81, 96, 133, 135, 136, 138, 139

[17] B. J. Dinkel, E. A. O'Laughlin-Phillips, N. S. Fechheimer, and R. G. Jaap, "Gametic products transmitted by chickens heterozygous for chromosomal rearrangements," *Cytogenetics and cell genetics*, vol. 23, no. 1-2, pp. 124–136, 1979. 8

[18] J. B. Searle, "Factors responsible for a karyotypic polymorphism in the common shrew, *Sorex araneus*," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 229, pp. 277–298, Dec. 1986. 8

[19] M. B. Qumsiyeh, "Evolution of Number and Morphology of Mammalian Chromosomes," *Journal of heredity*, 1994. 8, 143

[20] L. Fishman and J. H. Willis, "A novel meiotic drive locus almost completely distorts segregation in *Mimulus* (monkeyflower) hybrids," *Genetics*, vol. 169, pp. 347–353, Jan. 2005. 8, 94, 96, 135

[21] E. S. Buckler, T. L. Phelps-Durr, C. S. Buckler, R. K. Dawe, J. F. Doebley, and T. P. Holtsford, "Meiotic drive of chromosomal knobs reshaped the maize genome," *Genetics*, vol. 153, pp. 415–426, Sept. 1999. 9, 10, 11, 94, 115

[22] L. Fishman and A. Saunders, "Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers," *Science*, vol. 322, pp. 1559–1562, Dec. 2008. 9

[23] M. M. Rhoades, "Preferential Segregation in Maize," *Genetics*, vol. 27, pp. 395–407, July 1942. 9

[24] S. Agulnik, S. Adolph, H. Winking, and W. Traut, "Zoogeography of the chromosome 1 HSR in natural populations of the house mouse (*Mus musculus*)," *Hereditas*, vol. 119, no. 1, pp. 39–46, 1993. 9

[25] K. H. Choo, "Centromere DNA dynamics: latent centromeres and neocentromere formation," *The American Journal of Human Genetics*, vol. 61, pp. 1225–1233, Dec. 1997. 9

[26] M. Rocchi, N. Archidiacono, W. Schempp, O. Capozzi, and R. Stanyon, "Centromere repositioning in mammals," *Heredity*, vol. 108, no. 1, pp. 59–67, 2011. 9, 136

[27] R. K. Dawe and W. Z. Cande, "Induction of centromeric activity in maize by suppressor of meiotic drive 1," *PNAS*, vol. 93, pp. 8512–8517, Aug. 1996. 9, 11

[28] W. J. Peacock, E. S. Dennis, M. M. Rhoades, and A. J. Pryor, "Highly repeated DNA sequence limited to knob heterochromatin in maize," *PNAS*, vol. 78, pp. 4490–4494, July 1981. 10

[29] L. B. Kanizay, P. S. Albert, J. A. Birchler, and R. K. Dawe, "Intragenomic conflict between the two major knob repeats of maize," *Genetics*, vol. 194, pp. 81–89, May 2013. 10

[30] H. G. Yu, E. N. Hiatt, A. Chan, M. Sweeney, and R. K. Dawe, "Neocentromere-mediated chromosome movement in maize," *The Journal of cell biology*, vol. 139, pp. 831–840, Nov. 1997. 10, 12

[31] Y. Brandvain and G. Coop, "Scrambling eggs: Meiotic Drive and the Evolution of Female Recombination Rates," *Genetics*, vol. 190, no. February, pp. 709–723, 2011. 11, 74

[32] A. E. Longley, "Abnormal Segregation during Megasporogenesis in Maize," *Genetics*, vol. 30, p. 100, Jan. 1945. 11

[33] U. Nur and B. L. Brett, "Control of Meiotic Drive of B Chromosomes in the Mealybug, *Pseudococcus affinis* (obscurus)," *Genetics*, vol. 115, pp. 499–510, Mar. 1987. 12, 142

[34] M. J. Puertas, M. González-Sánchez, S. Manzanero, F. Romera, and M. M. Jiménez, "Genetic control of the rate of transmission of rye B chromosomes. IV. Localization of the genes controlling B transmission rate," *Heredity*, vol. 80, no. 2, pp. 209–213, 1998. 12, 142

[35] G. Wu, L. Hao, Z. Han, S. Gao, K. E. Latham, F. Pardo-Manuel de Villena, and C. Sapienza, "Maternal transmission ratio distortion at the mouse *Om* locus results from meiotic drive at the second meiotic division," *Genetics*, vol. 170, pp. 327–334, May 2005. 12

[36] E. Jenczewski, M. Gherardi, I. Bonnin, J. M. Prosperi, I. Olivieri, and T. Huguet, "Insight on segregation distortions in two intraspecific crosses between annual species of *Medicago* (Leguminosae)," *Theoretical and Applied Genetics*, vol. 94, pp. 682–691, Apr. 1997. 13, 94

[37] Y. Harushima, M. Nakagahra, M. Yano, T. Sasaki, and N. Kurata, "A genome-wide survey of reproductive barriers in an intraspecific hybrid," *Genetics*, vol. 159, pp. 883–892, Oct. 2001. 13

[38] D. L. Aylor, W. Valdar, W. Foulds-Mathes, R. J. Buus, R. A. Verdugo, R. S. Baric, M. T. Ferris, J. A. Frelinger, M. Heise, M. B. Frieman, L. E. Gralinski, T. A. Bell, J. P. Didion, K. Hua, D. L. Nehrenberg, C. L. Powell, J. Steigerwalt, Y. Xie, S. N. P. Kelada, F. S. Collins, I. V. Yang, D. A. Schwartz, L. Branstetter, E. J. Chesler, D. R. Miller,

J. Spence, E. Y. Liu, L. McMillan, A. Sarkar, J. R. Wang, W. Wang, Q. Zhang, K. W. Broman, R. Korstanje, C. Durrant, R. Mott, F. A. Iraqi, D. Pomp, D. W. Threadgill, F. Pardo-Manuel de Villena, and G. A. Churchill, "Genetic analysis of complex traits in the emerging Collaborative Cross," *Genome research*, vol. 21, pp. 1213–1222, Aug. 2011. 13, 14, 91, 94, 95, 98

[39] Collaborative Cross Consortium, "The genome architecture of the Collaborative Cross mouse genetic reference population," *Genetics*, vol. 190, pp. 389–401, Feb. 2012. 13, 17, 37, 77, 91, 92, 93, 94, 95, 96, 98, 118, 150

[40] H. Yang, J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell, C. E. Welsh, F. Bonhomme, A. H.-T. Yu, M. W. Nachman, J. Piálek, P. K. Tucker, P. Boursot, L. McMillan, and F. Pardo-Manuel de Villena, "Subspecific origin and haplotype diversity in the laboratory mouse," *Nature genetics*, vol. 43, pp. 648–655, July 2011. 14, 15, 17, 20, 22, 29, 32, 36, 53, 55, 56, 60, 84, 88, 92, 95, 114, 133, 150, 155

[41] J. P. Didion, H. Yang, K. Sheppard, C.-P. Fu, L. McMillan, F. Pardo-Manuel de Villena, and G. A. Churchill, "Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias," *BMC genomics*, vol. 13, p. 34, 2012. 14, 20, 22, 23, 58, 84, 127, 133

[42] J. P. Didion and F. Pardo-Manuel de Villena, "Deconstructing *Mus gemischus*: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse," *Mammalian genome*, vol. 24, pp. 1–20, Feb. 2013. 14, 15, 16, 17, 31, 133, 155

[43] H. Suzuki, T. Shimada, M. Terashima, K. Tsuchiya, and K. Aplin, "Temporal, spatial, and ecological modes of evolution of Eurasian *Mus* based on mitochondrial and nuclear gene sequences," *Molecular phylogenetics and evolution*, vol. 33, pp. 626–646, Dec. 2004. 14, 114

[44] A. Geraldes, P. Basset, B. Gibson, K. L. Smith, B. Harr, A. H.-T. Yu, N. Bulatova, Y. Ziv, and M. W. Nachman, "Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes," *Molecular Ecology*, vol. 17, pp. 5349–5363, Dec. 2008. 14

[45] F. Bonhomme and J. B. Searle, "Chapter 11," in *House mouse phylogeography*, Cambridge University Press, July 2012. 14

[46] H. Yonekawa, K. Moriwaki, O. Gotoh, N. Miyashita, Y. Matsushima, L. M. Shi, W. S. Cho, X. L. Zhen, and Y. Tagashira, "Hybrid origin of Japanese mice "*Mus musculus molossinus*": evidence from restriction analysis of mitochondrial DNA," *Molecular Biology and Evolution*, vol. 5, pp. 63–78, Jan. 1988. 15, 32

[47] E. M. Prager, C. Orrego, and R. D. Sage, "Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen," *Genetics*, vol. 150, pp. 835–861, Oct. 1998. 15

[48] M. Terashima, S. Furusawa, N. Hanzawa, K. Tsuchiya, A. Suyanto, K. Moriwaki, H. Yonekawa, and H. Suzuki, "Phylogeographic origin of Hokkaido house mice (*Mus musculus*) as indicated by genetic markers with maternal, paternal and biparental inheritance," *Heredity*, vol. 96, pp. 128–138, Jan. 2006. 15

[49] A. Stamatakis, T. Ludwig, and H. Meier, "RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees," *Bioinformatics*, vol. 21, pp. 456–463, Feb. 2005. 15

[50] F. Bonhomme, J.-L. Guénet, B. Dod, K. Moriwaki, and G. Bulfield, "The polyphyletic origin of laboratory inbred mice and their rate of evolution," *Biological Journal of the Linnean Society*, vol. 30, no. 1, pp. 51–58, 1987. 17, 31, 88

[51] J.-L. Guénet and F. Bonhomme, "Wild mice: an ever-increasing contribution to a popular mammalian model," *Trends in genetics*, vol. 19, pp. 24–31, Jan. 2003. 17, 76

[52] R. A. Gibbs, J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, C. A. Gill, R. D. Green, D. Hamernik, S. M. Kappes, S. Lien, and The Bovine HapMap Consortium, "Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds," *Science*, vol. 324, no. 5926, pp. 528–532, 2009. 17, 20

[53] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. J. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigó, M. S. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. J. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. V. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. A. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos,

V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. W. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S.-P. Yang, S.-P. Yang, E. M. Zdobnov, M. C. Zody, and E. S. Lander, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, pp. 520–562, Dec. 2002. 17, 18, 113

[54] R. J. Mural, M. D. Adams, E. W. Myers, H. O. Smith, G. L. G. Miklos, R. Wides, A. Halpern, P. W. Li, G. G. Sutton, J. Nadeau, S. L. Salzberg, R. A. Holt, C. D. Kodira, F. Lu, L. Chen, Z. Deng, C. C. Evangelista, W. Gan, T. J. Heiman, J. Li, Z. Li, G. V. Merkulov, N. V. Milshina, A. K. Naik, R. Qi, B. C. Shue, A. Wang, J. Wang, X. Wang, X. Yan, J. Ye, S. Yooseph, Q. Zhao, L. Zheng, S. C. Zhu, K. Biddick, R. Bolanos, A. L. Delcher, I. M. Dew, D. Fasulo, M. J. Flanigan, D. H. Huson, S. A. Kravitz, J. R. Miller, C. M. Mobarry, K. Reinert, K. A. Remington, Q. Zhang, X. H. Zheng, D. R. Nusskern, Z. Lai, Y. Lei, W. Zhong, A. Yao, P. Guan, R.-R. Ji, Z. Gu, Z.-Y. Wang, F. Zhong, C. Xiao, C.-C. Chiang, M. Yandell, J. R. Wortman, P. G. Amanatides, S. L. Hladun, E. C. Pratts, J. E. Johnson, K. L. Dodson, K. J. Woodford, C. A. Evans, B. Gropman, D. B. Rusch, E. Venter, M. Wang, T. J. Smith, J. T. Houck, D. E. Tompkins, C. Haynes, D. Jacob, S. H. Chin, D. R. Allen, C. E. Dahlke, R. Sanders, K. Li, X. Liu, A. A. Levitsky, W. H. Majoros, Q. Chen, A. C. Xia, J. R. Lopez, M. T. Donnelly, M. H. Newman, A. Glodek, C. L. Kraft, M. Nodell, F. Ali, H.-J. An, D. Baldwin-Pitts, K. Y. Beeson, S. Cai, M. Carnes, A. Carver, P. M. Caulk, A. Center, Y.-H. Chen, M.-L. Cheng, M. D. Coyne, M. Crowder, S. Danaher, L. B. Davenport, R. Desilets, S. M. Dietz, L. Doup, P. Dullaghan, S. Ferriera, C. R. Fosler, H. C. Gire, A. Gluecksmann, J. D. Gocayne, J. Gray, B. Hart, J. Haynes, J. Hoover, T. Howland, C. Ibegwam, M. Jalali, D. Johns, L. Kline, D. S. Ma, S. MacCawley, A. Magoon, F. Mann, D. May, T. C. McIntosh, S. Mehta, L. Moy, M. C. Moy, B. J. Murphy, S. D. Murphy, K. A. Nelson, Z. Nuri, K. A. Parker, A. C. Prudhomme, V. N. Puri, H. Qureshi, J. C. Raley, M. S. Reardon, M. A. Regier, Y.-H. C. Rogers, D. L. Romblad, J. Schutz, J. L. Scott, R. Scott, C. D. Sitter, M. Smallwood, A. C. Sprague, E. Stewart, R. V. Strong, E. Suh, K. Sylvester, R. Thomas, N. N. Tint, C. Tsonis, G. Wang, G. Wang, M. S. Williams, S. M. Williams, S. M. Windsor, K. Wolfe, M. M. Wu, J. Zaveri, K. Chaturvedi, A. E. Gabrielian, Z. Ke, J. Sun, G. Subramanian, J. C. Venter, C. M. Pfannkoch, M. Barnstead, and L. D. Stephenson, "A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome," *Science*, vol. 296, pp. 1661–1671, May 2002. 17

[55] K. A. Frazer, E. Eskin, H. M. Kang, M. A. Bogue, D. A. Hinds, E. J. Beilharz, R. V. Gupta, J. Montgomery, M. M. Morenzoni, G. B. Nilsen, C. L. Pethiyagoda, L. L. Stuve, F. M. Johnson, M. J. Daly, C. M. Wade, and D. R. Cox, "A sequence-based variation

map of 8.27 million SNPs in inbred mouse strains," *Nature*, vol. 448, pp. 1050–1053, Aug. 2007. 17, 19, 32, 88

[56] T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellåker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assunção, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, and D. J. Adams, "Mouse genomic variation and its effect on phenotypes and gene regulation," *Nature*, vol. 477, pp. 289–294, Sept. 2011. 17, 18, 23, 25, 32, 36, 55, 95, 103, 112, 113, 114, 155

[57] J.-Y. Xu, G.-B. Xu, and S.-L. Chen, "A new method for SNP discovery," *BioTechniques*, vol. 46, pp. 201–208, Mar. 2009. 17

[58] C. M. Wade, E. J. Kulbokas, A. W. Kirby, M. C. Zody, J. C. Mullikin, E. S. Lander, K. Lindblad-Toh, and M. J. Daly, "The mosaic structure of variation in the laboratory mouse genome," *Nature*, vol. 420, pp. 574–578, Dec. 2002. 18

[59] K. Wong, S. Bumpstead, L. van der Weyden, L. G. Reinholdt, L. G. Wilming, D. J. Adams, and T. M. Keane, "Sequencing and characterization of the FVB/NJ mouse genome," *Genome biology*, vol. 13, p. R72, Aug. 2012. 18, 23

[60] H. Yang, Y. Ding, L. N. Hutchins, J. Szatkiewicz, T. A. Bell, B. J. Paigen, J. H. Graber, F. Pardo-Manuel de Villena, and G. A. Churchill, "A customized and versatile high-density genotyping array for the mouse," *Nature Methods*, vol. 6, pp. 663–666, Sept. 2009. 19, 30, 55

[61] G. C. Kennedy, H. Matsuzaki, S. Dong, W.-m. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M. S. Phillips, M. T. Boyce-Jacino, S. P. A. Fodor, and K. W. Jones, "Large-scale genotyping of complex DNA," *Nature biotechnology*, vol. 21, pp. 1233–1237, Oct. 2003. 19

[62] A. R. Boyko, P. Quignon, L. Li, J. J. Schoenebeck, J. D. Degenhardt, K. E. Lohmueller, K. Zhao, A. Brisbin, H. G. Parker, B. M. vonHoldt, M. Cargill, A. Auton, A. Reynolds, A. G. Elkahloun, M. Castelhano, D. S. Mosher, N. B. Sutter, G. S. Johnson, J. Novembre, M. J. Hubisz, A. Siepel, R. K. Wayne, C. D. Bustamante, and E. a. Ostrander, "A Simple Genetic Architecture Underlies Morphological Variation in Dogs," *PLoS biology*, vol. 8, p. e1000451, Aug. 2010. 20

[63] Affymetrix Inc, "BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array." 20, 23, 152

[64] M. H. Wright, C. W. Tung, K. Zhao, A. Reynolds, S. R. McCouch, and C. D. Bustamante, "ALCHEMY: a reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations," *Bioinformatics*, vol. 26, pp. 2952–2960, Dec. 2010. 23, 152

[65] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, pp. 1754–1760, July 2009. 23

[66] D. M. Church, L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein, X. She, C. J. Bult, R. Agarwala, J. L. Cherry, M. DiCuccio, W. Hlavina, Y. Kapustin, P. Meric, D. Maglott, Z. Birtle, A. C. Marques, T. Graves, S. Zhou, B. Teague, K. Potamousis, C. Churas, M. Place, J. Herschleb, R. Runnheim, D. Forrest, J. Amos-Landgraf, D. C. Schwartz, Z. Cheng, K. Lindblad-Toh, E. E. Eichler, C. P. Ponting, and Mouse Genome Sequencing Consortium, "Lineage-specific biology revealed by a finished genome assembly of the mouse," *PLoS biology*, vol. 7, p. e1000112, May 2009. 25

[67] The International HapMap 3 Consortium, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, pp. 52–58, Sept. 2010. 29

[68] C. E. Bishop, P. Boursot, B. Baron, F. Bonhomme, and D. Hatat, "Most classical *Mus musculus domesticus* laboratory mouse strains carry a *Mus musculus musculus* Y chromosome," *Nature*, vol. 315, pp. 70–72, May 1985. 31

[69] K. Moriwaki, T. Shiroishi, H. Yonekawa, N. Miyashita, and Y. Sagai, "Genetic status of Japanese wild mice and immunological characters of their H-2 antigens," in *Teratocarcinoma and Embryonic Cell Interactions* (T. Muramatsu, G. Gachelin, A. A. Monscona, and Y. Ikawa, eds.), pp. 41–56, Tokyo: Japan Scientific Soc. Press and Academic Press, Aug. 1982. 31

[70] K. Paigen and J. T. Eppig, "A mouse phenome project," *Mammalian genome*, vol. 11, pp. 715–717, Sept. 2000. 31, 92

[71] H. Yang, T. A. Bell, G. A. Churchill, and F. Pardo-Manuel de Villena, "On the subspecific origin of the laboratory mouse," *Nature genetics*, vol. 39, pp. 1100–1107, Sept. 2007. 32, 33, 88

[72] F. Staubach, A. Lorenc, P. W. Messer, K. Tang, D. A. Petrov, and D. Tautz, "Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*)," *PLoS Genetics*, vol. 8, p. e1002891, Aug. 2012. 33, 74

[73] D. J. Lawson, G. Hellenthal, S. R. Myers, and D. Falush, "Inference of population structure using dense haplotype data," *PLoS Genetics*, vol. 8, no. 1, p. e1002453, 2012. 34, 60

[74] J. Wang, K. J. Moore, Q. Zhang, F. Pardo-Manuel de Villena, W. Wang, and L. McMillan, "Genome-wide compatible SNP intervals and their properties," in *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pp. 43–52, ACM, 2010. 36

[75] R. R. Hudson and N. L. Kaplan, "Statistical properties of the number of recombination events in the history of a sample of DNA sequences," *Genetics*, vol. 111, pp. 147–164, Sept. 1985. 36

[76] J. Piálek, M. Vyskocilová, B. Bímová, D. Havelková, J. Piálková, P. Dufková, V. Bencová, L. Dureje, T. Albrecht, H. C. Hauffe, M. Macholán, P. Munclinger, R. Storchova, A. Zajícová, V. Holán, S. Gregorová, and J. Forejt, "Development of unique house mouse resources suitable for evolutionary studies of speciation," *Journal of heredity*, vol. 99, pp. 34–44, Jan. 2008. 36

[77] T. Salcedo, A. Geraldes, and M. W. Nachman, "Nucleotide variation in wild and inbred mice," *Genetics*, vol. 177, pp. 2277–2291, Dec. 2007. 36

[78] M. Nei and W. H. Li, "Mathematical model for studying genetic variation in terms of restriction endonucleases," *PNAS*, vol. 76, pp. 5269–5273, Oct. 1979. 36

[79] P. Boursot and K. Belkhir, "Mouse SNPs for evolutionary biology: beware of ascertainment biases," *Genome research*, vol. 16, pp. 1191–1192, Oct. 2006. 37

[80] B. Harr, "Genomic islands of differentiation between house mouse subspecies," *Genome research*, vol. 16, pp. 730–737, June 2006. 37

[81] S. L. Burgess-Herbert, S.-W. Tsaih, I. M. Stylianou, K. Walsh, A. J. Cox, and B. Paigen, "An experimental assessment of in silico haplotype association mapping in laboratory mice," *BMC genetics*, vol. 10, p. 81, 2009. 37

[82] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, "Efficient control of population structure in model organism association mapping," *Genetics*, vol. 178, pp. 1709–1723, Mar. 2008. 37

[83] C. E. Welsh and L. McMillan, "Accelerating the inbreeding of multi-parental recombinant inbred lines generated by sibling matings," *G3*, vol. 2, pp. 191–198, Feb. 2012. 37, 93, 94, 124

[84] J. Felsenstein, "Notices," *Cladistics*, vol. 5, pp. 163–166, June 1989. 40, 156

[85] T. Jombart, F. Balloux, and S. Dray, "adephylo: new tools for investigating the phylogenetic signal in biological traits," *Bioinformatics*, vol. 26, pp. 1907–1909, Aug. 2010. 40

[86] D. C. Wallace, "Mitochondrial DNA sequence variation in human evolution and disease," *PNAS*, vol. 91, pp. 8739–8746, Sept. 1994. 40

[87] H. Rajabi-Maham, A. Orth, R. Siahsarvie, P. Boursot, J. Darvish, and F. Bonhomme, "The southeastern house mouse *Mus musculus castaneus* (Rodentia: Muridae) is a polytypic subspecies," *Biological Journal of the Linnean Society*, vol. 107, no. 2, pp. 295–306, 2012. 41

[88] T. Popova, E. Manié, D. Stoppa-Lyonnet, G. Rigaill, E. Barillot, and M. Stern, "Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays," *Genome biology*, vol. 10, no. 11, p. R128, 2009. 41, 153

[89] J. D. Calaway, A. B. Lenarcic, J. P. Didion, J. R. Wang, J. B. Searle, L. McMillan, W. Valdar, and F. Pardo-Manuel de Villena, "Genetic architecture of skewed X inactivation in the laboratory mouse," *PLoS Genetics*, vol. 9, no. 10, p. e1003853, 2013. 44

[90] A. Gropp, U. Tettenborn, and E. Von Lehmann, "Chromosomenuntersuchungen bei der Tabakmaus (M. poschiavinus) und bei Tabakmaus-Hybriden," *Experientia*, vol. 25, no. 8, pp. 875–876, 1969. 44

[91] J. Piálek, H. C. Hauffe, and J. B. Searle, "Chromosomal variation in the house mouse," *Biological Journal of the Linnean Society*, vol. 84, no. 3, pp. 535–563, 2005. 44, 45, 48, 50, 79, 80

[92] E. R. Hauser, M. Boehnke, and S. W. Guo, "Affected sib-pair interval mapping and exclusion for complex genetic traits: Inferring identity by descent status from relatives," *The American Journal of Human Genetics*, vol. 55, Sept. 1994. 45

[93] E. Solano, R. Castiglia, and M. Corti, "A new chromosomal race of the house mouse, *Mus musculus domesticus*, in the Vulcano Island-Aeolian Archipelago, Italy," *Hereditas*, vol. 144, no. 3, pp. 75–77, 2007. 45

[94] J. C. Auffray, P. Fontanillas, J. Catalan, and J. Britton-Davidian, "Developmental stability in house mice heterozygous for single Robertsonian fusions," *Journal of heredity*, vol. 92, no. 1, pp. 23–29, 2001. 45

[95] A. Gropp and H. Winking, "Robertsonian Translocations: Cytology, Meiosis, Segregation Patterns and Biological Consequences of Heterozygosity," *Symp zool Soc Lond*, pp. 141–181, Jan. 1981. 45, 51, 81

[96] E. Capanna, "Robertsonian numerical variation in animal speciation: *Mus musculus*, an emblematic model," in *Mechanisms of speciation* (B. C, ed.), pp. 155–177, New York: Alan R. Liss, Inc, Jan. 1982. 45

[97] R. J. Baker and J. W. Bickham, "Speciation by monobrachial centric fusions," *PNAS*, vol. 83, pp. 8245–8248, Nov. 1986. 45

[98] H. C. Hauffe and J. B. Searle, "Chromosomal heterozygosity and fertility in house mice (*Mus musculus domesticus*) from Northern Italy," *Genetics*, vol. 150, no. 3, pp. 1143–1154, 1998. 45

[99] N. Chatti, J. Britton-Davidian, J. Catalan, J. C. Auffray, and K. Saïd, "Reproductive trait divergence and hybrid fertility patterns between chromosomal races of the house mouse in Tunisia: analysis of wild and laboratory-bred males and females," *Biological Journal of the Linnean Society*, vol. 84, no. 3, pp. 407–416, 2005. 45

[100] J. Britton-Davidian, H. Sonjaya, J. Catalan, and G. Cattaneo-Berrebi, "Robertsonian heterozygosity in wild mice: fertility and transmission rates in Rb(16.17) translocation heterozygotes," *Genetica*, vol. 80, no. 3, pp. 171–174, 1990. 45

[101] B. M. Wallace, J. B. Searle, and C. A. Everett, "The effect of multiple simple Robertsonian heterozygosity on chromosome pairing and fertility of wild-stock house mice (*Mus musculus domesticus*)," *Cytogenetic and genome research*, vol. 96, no. 1-4, pp. 276–286, 2002. 45, 47

[102] R. Castiglia and E. Capanna, "Contact zone between chromosomal races of *Mus musculus domesticus*. 2. Fertility and segregation in laboratory-reared and wild mice heterozygous for multiple robertsonian rearrangements," *Heredity*, vol. 85 (Pt 2), no. March, pp. 147–156, 2000. 45, 81, 136

[103] J. B. Searle, Y. N. Navarro, and G. Ganem, "Further studies of a staggered hybrid zone in *Mus musculus domesticus* (the house mouse)," *Heredity*, vol. 71, pp. 523–531, 1993. 45, 50, 79

[104] R. Johannisson and H. Winking, "Synaptonemal complexes of chains and rings in mice heterozygous for multiple Robertsonian translocations," *Chromosome Research*, vol. 2, pp. 137–145, Mar. 1994. 45

[105] J. B. Searle, "A Hybrid Zone Comprising Staggered Chromosomal Clines in the House Mouse (*Mus musculus domesticus*)," *Proceedings of the Royal Society B: Biological Sciences*, vol. 246, pp. 47–52, Oct. 1991. 45

[106] S. Adolph and J. Klein, "Genetic variation of wild mouse populations in south Germany. I. Cytogenetic Study," *Genetical Research*, vol. 2, pp. 117–134, Jan. 1983. 47

[107] J. Catalan, J. C. Auffray, F. Pellestor, and J. Britton-Davidian, "Spontaneous occurrence of a Robertsonian fusion involving chromosome 19 by single whole-arm reciprocal translocation (WART) in wild-derived house mice," *Chromosome Research*, vol. 8, no. 7, pp. 593–601, 2000. 47

[108] E. Capanna and C. A. Redi, "Whole-arm reciprocal translocation (WART) between Robertsonian chromosomes: finding of a Robertsonian heterozygous mouse with karyotype derived through WARTs," *Chromosome Research*, vol. 3, pp. 135–137, Mar. 1995. 47, 51

[109] M. Corti, "Chromosomal speciation and reticulate evolution: Testing phylogenetic hypotheses with compatibility and parsimony," *Bolletino di zoologia*, vol. 58, pp. 307–319, Jan. 1991. 47

[110] S. Garagna, M. Zuccotti, C. A. Redi, and E. Capanna, "Trapping speciation," *Nature*, vol. 390, pp. 241–242, Nov. 1997. 47, 79, 80, 138, 144

[111] H. C. Hauffe, "Evolution of the chromosomal races of *Mus musculus domesticus* in the Rhaetian Alps: the roles of wholearm reciprocal translocation and zonal raciation," *Biological Journal of the Linnean Society*, vol. 62, no. 2, pp. 255–278, 1997. 47, 79

[112] T. A. White, M. Bordewich, and J. B. Searle, "A network approach to study karyotypic evolution: the chromosomal races of the common shrew (*Sorex araneus*) and house

mouse (*Mus musculus*) as model systems," *Systematic biology*, vol. 59, no. 3, pp. 262–276, 2010. 47

[113] P. Franchini, R. Castiglia, and E. Capanna, "Reproductive isolation between chromosomal races of the house mouse *Mus musculus domesticus* in a parapatric contact area revealed by an analysis of multiple unlinked loci," *Journal of Evolutionary Biology*, vol. 21, pp. 502–513, Mar. 2008. 47

[114] P. Franchini, P. Colangelo, E. Solano, E. Capanna, E. Verheyen, and R. Castiglia, "Reduced gene flow at pericentromeric loci in a hybrid zone involving chromosomal races of the house mouse *Mus musculus domesticus*," *Evolution*, vol. 64, pp. 2020–2032, July 2010. 47, 71, 83

[115] M. T. Davisson and E. C. Akeson, "Recombination suppression by heterozygous Robertsonian chromosomes in the mouse," *Genetics*, vol. 133, pp. 649–667, Mar. 1993. 47, 71

[116] N. Chatti, K. Saïd, J. Catalan, J. Britton-Davidian, and J. C. Auffray, "Developmental Instability in Wild Chromosomal Hybrids of the House Mouse," *Evolution*, vol. 53, p. 1268, Aug. 1999. 47

[117] A. Ruiz-Herrera, M. Farré, M. Ponsà, and T. J. Robinson, "Selection against Robertsonian fusions involving housekeeping genes in the house mouse: integrating data from gene expression arrays and chromosome evolution," *Chromosome Research*, vol. 18, pp. 801–808, Sept. 2010. 48, 81

[118] İ. Gündüz, M. J. López-Fuster, J. Ventura, and J. B. Searle, "Clinal analysis of a chromosomal hybrid zone in the house mouse," *Genetical Research*, vol. 77, pp. 41–51, Feb. 2001. 50

[119] S. Chakrabarti and A. Chakrabarti, "Spontaneous Robertsonian fusion leading to karyotype variation in the house mouse—first report from Asia," *Experientia*, vol. 33, no. 2, pp. 175–177, 1977. 50

[120] K. Moriwaki, H. Yonekawa, O. Gotoh, M. Minezawa, H. Winking, and A. Gropp, "Implications of the genetic divergence between European wild mice with Robertsonian translocations from the viewpoint of mitochondrial DNA," *Genetical Research*, vol. 43, pp. 277–287, June 1984. 50

[121] M. W. Nachman and J. B. Searle, "Why is the house mouse karyotype so variable?," *Trends in Ecology and Evolution*, vol. 10, pp. 397–402, Oct. 1995. 50

[122] M. W. Nachman, S. N. Boyer, J. B. Searle, and C. F. Aquadro, "Mitochondrial DNA variation and the evolution of Robertsonian chromosomal races of house mice, *Mus domesticus*," *Genetics*, vol. 136, pp. 1105–1120, Mar. 1994. 51

[123] E. Capanna and R. Castiglia, "Chromosomes and speciation in *Mus musculus domesticus*," *Cytogenetic and genome research*, vol. 105, no. 2-4, pp. 375–384, 2004. 51

[124] N. N. Vorontsov and E. A. Lyapunova, "Explosive chromosomal speciation in seismic active regions," in *Chromosomes today*, pp. 279–294, Dordrecht: Springer Netherlands, 1984. 51, 79

[125] C. I. Amos, "Successful design and conduct of genome-wide association studies," *Human Molecular Genetics*, vol. 16 Spec No. 2, pp. R220–5, Oct. 2007. 52

[126] J. M. Satagopan, D. A. Verbel, E. S. Venkatraman, K. E. Offit, and C. B. Begg, "Two-stage designs for gene-disease association studies," *Biometrics*, vol. 58, pp. 163–170, Mar. 2002. 52

[127] A. Gropp and H. Winking, "Robertsonian translocations: cytology, meiosis, segregation pattern and biological consequences of heterozygosity," in *Biology of the House Mouse* (R. J. Berry, ed.), pp. 141–181, New York/London: Academic Press, 1981. 53

[128] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, pp. 904–909, July 2006. 56

[129] G. H. Hardy, "Mendelian proportions in a mixed population," *Science*, vol. 28, pp. 49–50, July 1908. 58

[130] S. W. Guo and E. A. Thompson, "Performing the exact test of Hardy-Weinberg proportion for multiple alleles," *Biometrics*, vol. 48, pp. 361–372, June 1992. 58

[131] S. Ihle, I. Ravaoarimanana, M. Thomas, and D. Tautz, "An analysis of signatures of selective sweeps in natural populations of the house mouse," *Molecular Biology and Evolution*, vol. 23, pp. 790–797, Apr. 2006. 58, 76

[132] P. Scheet, "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*, vol. 78, pp. 629–644, Apr. 2006. 60, 155

[133] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing," *Nature genetics*, July 2012. 60, 155

[134] B. Yalcin, J. Nicod, A. Bhomra, S. Davidson, J. Cleak, L. Farinelli, M. Østerås, A. Whitley, W. Yuan, X. Gan, M. Goodson, P. Klenerman, A. Satpathy, D. Mathis, C. Benoist, D. J. Adams, R. Mott, and J. Flint, "Commercially available outbred mice for genome-wide association studies," *PLoS Genetics*, vol. 6, Sept. 2010. 61, 77

[135] S. Wright, "Coefficients of inbreeding and relationship," *The American naturalist*, vol. 56, pp. 330–338, 1922. 61

[136] C. C. Laurie, D. A. Nickerson, A. D. Anderson, B. S. Weir, R. J. Livingston, M. D. Dean, K. L. Smith, E. E. Schadt, and M. W. Nachman, "Linkage disequilibrium in wild mice," *PLoS Genetics*, vol. 3, no. 8, p. e144, 2007. 62, 66, 76, 77

[137] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, pp. 559–575, Sept. 2007. 64

[138] S. R. Browning, "Missing data imputation and haplotype phase inference for genome-wide association studies," *Human Genetics*, vol. 124, pp. 439–450, Oct. 2008. 66

[139] B. Servin, "Imputation-based analysis of association studies: candidate regions and quantitative traits," *PLoS Genetics*, vol. 3, p. e114, July 2007. 66

[140] Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, June 2007. 68

[141] M. Stephens and D. J. Balding, "Bayesian statistical methods for genetic association studies," *Nature Reviews Genetics*, vol. 10, pp. 681–690, Oct. 2009. 68

[142] R. C. Johnson, G. W. Nelson, J. L. Troyer, J. A. Lautenberger, B. D. Kessing, C. A. Winkler, and S. J. O'brien, "Accounting for multiple comparisons in a genome-wide association study (GWAS)," *BMC genomics*, vol. 11, p. 724, 2010. 68

[143] R. Faria and A. Navarro, "Chromosomal speciation revisited: rearranging theory with pieces of evidence," *Trends in Ecology and Evolution*, vol. 25, pp. 660–669, Nov. 2010. 71

[144] D. W. Förster, M. L. Mathias, J. Britton-Davidian, and J. B. Searle, "Origin of the chromosomal radiation of Madeiran house mice: a microsatellite analysis of metacentric chromosomes," *Heredity*, vol. 110, pp. 380–388, Apr. 2013. 71, 77

[145] J. M. Smith and J. Haigh, "The hitch-hiking effect of a favourable gene," *Genet Res*, vol. 23, no. 1, pp. 23–25, 1974. 74

[146] N. L. Kaplan, R. R. Hudson, and C. H. Langley, "The "hitchhiking effect" revisited," *Genetics*, vol. 123, pp. 887–899, Dec. 1989. 74

[147] D. C. Presgraves, P. R. Gérard, A. Cherukuri, and T. W. Lyttle, "Large-scale selective sweep among segregation distorter chromosomes in African populations of *Drosophila melanogaster*," *PLoS Genetics*, vol. 5, no. 5, p. e1000463, 2009. 74

[148] E. Axelsson, A. Albrechtsen, A. P. van, L. Li, H. J. Megens, A. L. Vereijken, R. P. Crooijmans, M. A. Groenen, H. Ellegren, E. Willerslev, and R. Nielsen, "Segregation distortion in chicken and the evolutionary consequences of female meiotic drive in birds," *Heredity*, vol. 105, pp. 290–298, Sept. 2010. 74

[149] R. Nielsen, S. H. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. D. Bustamante, "Genomic scans for selective sweeps using SNP data," *Genome research*, vol. 15, pp. 1566–1575, Nov. 2005. 74

[150] R. J. Berry and F. H. Bronson, "Life history and bioeconomy of the house mouse," *Biological reviews of the Cambridge Philosophical Society*, vol. 67, pp. 519–550, Nov. 1992. 77

[151] I. O. Brahim, N. Chatti, J. Britton-Davidian, and K. Saïd, "Origin and evolution of the Robertsonian populations of the house mouse (Rodentia, Muridae) in Tunisia based on allozyme studies," *Biological Journal of the Linnean Society*, vol. 84, no. 3, pp. 515–521, 2005. 77

[152] M. N. Y. Navarro and J. Britton-Davidian, "Genetic structure of insular Mediterranean populations of the house mouse," *Biological Journal of the Linnean Society*, vol. 36, pp. 377–390, Jan. 2008. 77

[153] R. Bandyopadhyay, A. Heller, C. Knox-DuBois, C. McCaskill, S. A. Berend, S. L. Page, and L. G. Shaffer, "Parental origin and timing of de novo Robertsonian translocation formation," *The American Journal of Human Genetics*, vol. 71, pp. 1456–1462, Dec. 2002. 79

[154] F. Pardo-Manuel de Villena, "Evolution of the mammalian karyotype," in *Mammalian Genomics* (A. Ruvinsky and J. A. M. Graves, eds.), pp. 317–348, CABI, Dec. 2004. 79, 80, 138

[155] S. Garagna, M. Zuccotti, and E. Capanna, "High-resolution organization of mouse telomeric and pericentromeric DNA," *Cytogenetic and genome research*, 2002. 79

[156] S. Garagna, D. Broccoli, C. A. Redi, J. B. Searle, H. J. Cooke, and E. Capanna, "Robertsonian metacentrics of the house mouse lose telomeric sequences but retain some minor satellite DNA in the pericentromeric area," *Chromosoma*, vol. 103, pp. 685–692, July 1995. 79

[157] P. W. Hedrick, "The establishment of chromosomal variants," *Evolution*, vol. 35, no. 2, pp. 322–332, 1981. 80, 96, 118, 136, 137

[158] P. Scriven, "Robertsonian translocations introduced into an island population of house mice," *Journal of Zoology*, vol. 227, pp. 493–502, 1992. 80, 148

[159] J. F. Dallas, F. Bonhomme, P. Boursot, J. Britton-Davidian, and V. Bauchau, "Population genetic structure in a Robertsonian race of house mice: evidence from microsatellite polymorphism," *Heredity*, vol. 80 (Pt 1), pp. 70–77, Jan. 1998. 80

[160] R. Schulz, L. A. Underkoffler, J. N. Collins, and R. J. Oakey, "Nondisjunction and transmission ratio distortion of Chromosome 2 in a (2.8) Robertsonian translocation mouse strain," *Mammalian genome*, vol. 17, no. 3, pp. 239–247, 2006. 81

[161] S. Henikoff, K. Ahmad, and H. S. Malik, "The centromere paradox: stable inheritance with rapidly evolving DNA," *Science*, vol. 293, pp. 1098–1102, Aug. 2001. 81, 135, 144

[162] A. Geraldes, P. Basset, K. L. Smith, and M. W. Nachman, "Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination," *Molecular Ecology*, vol. 20, pp. 4722–4736, Nov. 2011. 85

[163] G. Wang, Y. Yang, and J. Ott, "Genome-wide conditional search for epistatic disease-predisposing variants in human association studies," *Human heredity*, vol. 70, no. 1, pp. 34–41, 2010. 86

[164] G. Zheng, B. Freidlin, Z. Li, and J. L. Gastwirth, "Genomic control for association studies under various genetic models," *Biometrics*, vol. 61, pp. 186–192, Mar. 2005. 86

[165] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at UCSC," *Genome research*, vol. 12, pp. 996–1006, June 2002. 86

[166] P. Havlak, R. Chen, K. J. Durbin, A. Egan, Y. Ren, X.-Z. Song, G. M. Weinstock, and R. A. Gibbs, "The Atlas genome assembly system," *Genome research*, vol. 14, pp. 721–732, Apr. 2004. 86

[167] B. B. Orcheski and T. M. Davis, "An enhanced method for sequence walking and paralog mining: TOPO® Vector-Ligation PCR," *BMC research notes*, vol. 3, p. 61, 2010. 86

[168] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature genetics*, vol. 43, pp. 491–498, Apr. 2011. 89

[169] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. A. T. McVean, "*De novo* assembly and genotyping of variants using colored de Bruijn graphs," *Nature genetics*, vol. 44, pp. 226–232, Feb. 2012. 89

[170] J. T. Simpson and R. Durbin, "Efficient *de novo* assembly of large genomes using compressed data structures," *Genome research*, vol. 22, pp. 549–556, Mar. 2012. 89

[171] A. Helmrich, K. Stout-Weider, K. Hermann, E. Schrock, and T. Heiden, "Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes," *Genome research*, vol. 16, pp. 1222–1230, Oct. 2006. 89

[172] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang, "*De novo* assembly of human genomes with massively parallel short read sequencing," *Genome research*, vol. 20, pp. 265–272, Feb. 2010. 89

[173] A. Fujimoto, H. Nakagawa, N. Hosono, K. Nakano, T. Abe, K. A. Boroevich, M. Nagasaki, R. Yamaguchi, T. Shibuya, M. Kubo, S. Miyano, Y. Nakamura, and T. Tsunoda,

"Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing," *Nature genetics*, vol. 42, pp. 931–936, Oct. 2010. 89

[174] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, "Repbase Update, a database of eukaryotic repetitive elements," *Cytogenetic and genome research*, vol. 110, no. 1-4, pp. 462–467, 2005. 90

[175] J. L. Peirce, L. Lu, J. Gu, L. M. Silver, and R. W. Williams, "A new set of BXD recombinant inbred lines from advanced intercross populations in mice," *BMC genetics*, vol. 5, p. 7, Apr. 2004. 91

[176] T. Hrbek, R. A. de Brito, B. Wang, L. S. Pletscher, and J. M. Cheverud, "Genetic characterization of a new set of recombinant inbred lines (LGXSM) formed from the intercross of SM/J and LG/J inbred mouse strains," *Mammalian genome*, vol. 17, no. 5, pp. 417–429, 2006. 91

[177] B. J. Bennett, C. R. Farber, L. Orozco, H. M. Kang, A. Ghazalpour, N. Siemers, M. Neubauer, I. Neuhaus, R. Yordanova, B. Guan, A. Truong, W.-p. Yang, A. He, P. Kayne, P. Gargalovic, T. Kirchgessner, C. Pan, L. W. Castellani, E. Kostem, N. Furlotte, T. A. Drake, E. Eskin, and A. J. Lusis, "A high-resolution association mapping panel for the dissection of complex traits in mice," *Genome Res*, 2010. 92

[178] J. M. Cheverud, T. T. Vaughn, L. S. Pletscher, A. C. Peripato, E. S. Adams, C. F. Erikson, and K. J. King-Ellison, "Genetic architecture of adiposity in the cross of LG/J and SM/J inbred mice," *Mammalian genome*, vol. 12, no. 1, pp. 3–12, 2001. 92

[179] R. Mott, C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint, "A method for fine mapping quantitative trait loci in outbred animal stocks," *PNAS*, vol. 97, pp. 12649–12654, Nov. 2000. 92

[180] K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill, "High-resolution genetic mapping using the Mouse Diversity outbred population," *Genetics*, vol. 190, pp. 437–447, Feb. 2012. 92, 94, 119

[181] D. W. Threadgill, K. W. Hunter, and R. W. Williams, "Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort," *Mammalian genome*, vol. 13, pp. 175–178, Apr. 2002. 92

[182] G. A. Churchill, D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie, J. Beatty, W. D. Beavis, J. K. Belknap, B. Bennett, W. Berrettini, A. Bleich, M. Bogue, K. W. Broman, K. J. Buck, E. Buckler, M. Burmeister, E. J. Chesler, J. M. Cheverud, S. Clapcote, M. N. Cook, R. D. Cox, J. C. Crabbe, W. E. Crusio, A. Darvasi, C. F. Deschepper, R. W. Doerge, C. R. Farber, J. Forejt, D. Gaile, S. J. Garlow, H. Geiger, H. Gershenfeld, T. Gordon, J. Gu, W. Gu, G. de Haan, N. L. Hayes, C. Heller, H. Himmelbauer, R. Hitzemann, K. Hunter, H.-C. Hsu, F. A. Iraqi, B. Ivandic, H. J. Jacob, R. C. Jansen, K. J. Jepsen, D. K. Johnson, T. E. Johnson, G. Kempermann, C. Kendziorski, M. Kotb,

R. F. Kooy, B. Llamas, F. Lammert, J.-M. Lassalle, P. R. Lowenstein, L. Lu, A. Lusis, K. F. Manly, R. Marcucio, D. Matthews, J. F. Medrano, D. R. Miller, G. Mittleman, B. A. Mock, J. S. Mogil, X. Montagutelli, G. Morahan, D. G. Morris, R. Mott, J. H. Nadeau, H. Nagase, R. S. Nowakowski, B. F. O'Hara, A. V. Osadchuk, G. P. Page, B. Paigen, K. Paigen, A. A. Palmer, H.-J. Pan, L. Peltonen-Palotie, J. Peirce, D. Pomp, M. Pravenec, D. R. Prows, Z. Qi, R. H. Reeves, J. Roder, G. D. Rosen, E. E. Schadt, L. C. Schalkwyk, Z. Seltzer, K. Shimomura, S. Shou, M. J. Sillanpää, L. D. Siracusa, H.-W. Snoeck, J. L. Spearow, K. Svenson, L. M. Tarantino, D. W. Threadgill, L. A. Toth, W. Valdar, F. Pardo-Manuel de Villena, C. Warden, S. Whatley, R. W. Williams, T. Wiltshire, N. Yi, D. Zhang, M. Zhang, and F. Zou, "The Collaborative Cross, a community resource for the genetic analysis of complex traits," *Nature genetics*, vol. 36, pp. 1133–1137, Nov. 2004. 92

[183] T. Bhattacharyya, S. Gregorová, O. Mihola, M. Anger, J. Sebestova, P. Denny, P. Simecek, and J. Forejt, "Mechanistic basis of infertility of mouse intersubspecific hybrids," *PNAS*, vol. 110, pp. E468–77, Feb. 2013. 93

[184] K. W. Broman, "Genotype probabilities at intermediate generations in the construction of recombinant inbred lines," *Genetics*, vol. 190, pp. 403–412, Feb. 2012. 93

[185] F. Teuscher and K. W. Broman, "Haplotype Probabilities for Multiple-Strain Recombinant Inbred Lines," *Genetics*, vol. 175, pp. 1267–1274, Dec. 2006. 93

[186] A. B. Lenarcic, K. L. Svenson, G. A. Churchill, and W. Valdar, "A general Bayesian approach to analyzing diallel crosses of inbred strains," *Genetics*, vol. 190, pp. 413–435, Feb. 2012. 93

[187] Y. Gong and F. Zou, "Varying coefficient models for mapping quantitative trait loci using recombinant inbred intercrosses," *Genetics*, vol. 190, pp. 475–486, Feb. 2012. 93

[188] W. Zhang, R. Korstanje, J. Thaisz, F. Staedtler, N. Harttman, L. Xu, M. Feng, L. Yanas, W. Valdar, and K. DiPetrillo, "Genome-wide association mapping of quantitative traits in outbred mice," *G3*, vol. 2, pp. 167–174, Feb. 2012. 93

[189] C. Durrant, H. Tayem, B. Yalcin, J. Cleak, L. Goodstadt, F. Pardo-Manuel de Villena, R. Mott, and F. A. Iraqi, "Collaborative Cross mice and their power to map host susceptibility to *Aspergillus fumigatus* infection," *Genome research*, vol. 21, pp. 1239–1248, Aug. 2011. 95, 98

[190] M. A. White, B. Steffy, T. Wiltshire, and B. A. Payseur, "Genetic dissection of a key reproductive barrier between nascent species of house mice," *Genetics*, vol. 189, pp. 289–304, Sept. 2011. 95

[191] K. Kim, S. Thomas, I. B. Howard, H. E. Doherty, F. Y. Ideraabdullah, D. A. Detwiler, and F. Pardo-Manuel de Villena, "Meiotic drive at the *Om* locus in wildderived inbred mouse strains," *Biological Journal of the Linnean Society*, vol. 84, no. 3, pp. 487–492, 2005. 96, 115, 133

[192] L. D. Siracusa, W. G. Alvord, W. A. Bickmore, N. A. Jenkins, and N. G. Copeland, "Interspecific backcross mice show sex-specific differences in allelic inheritance," *Genetics*, vol. 128, pp. 813–821, Aug. 1991. 96, 112, 113

[193] L. B. Rowe, J. H. Nadeau, R. Turner, W. N. Frankel, V. A. Letts, J. T. Eppig, M. S. H. Ko, S. J. Thurston, and E. H. Birkenmeier, "Maps from two interspecific backcross DNA panels available as a community genetic mapping resource," *Mammalian genome*, vol. 5, pp. 253–274, May 1994. 96, 113, 124

[194] X. Montagutelli, R. Turner, and J. H. Nadeau, "Epistatic control of non-Mendelian inheritance in mouse interspecific crosses," *Genetics*, vol. 143, pp. 1739–1752, Aug. 1996. 96, 113

[195] J. L. Rocha, E. J. Eisen, F. Siewerdt, L. D. Vleck, and D. Pomp, "A large-sample QTL study in mice: III. Reproduction," *Mammalian genome*, vol. 15, pp. 878–886, Nov. 2004. 96

[196] B. Yalcin, K. Wong, A. Agam, M. Goodson, T. M. Keane, X. Gan, C. Nellåker, L. Goodstadt, J. Nicod, A. Bhomra, P. Hernandez-Pliego, H. Whitley, J. Cleak, R. Dutton, D. Janowitz, R. Mott, D. J. Adams, and J. Flint, "Sequence-based characterization of structural variation in the mouse genome," *Nature*, vol. 477, pp. 326–329, Sept. 2011. 103, 112, 113

[197] S. A. Kelly, T. A. Bell, S. R. Selitsky, R. J. Buus, K. Hua, G. M. Weinstock, T. Garland, F. Pardo-Manuel de Villena, and D. Pomp, "A novel intronic single nucleotide polymorphism in the myosin heavy polypeptide 4 gene is responsible for the mini-muscle phenotype characterized by major reduction in hind-limb muscle mass in mice," *Genetics*, vol. 195, pp. 1385–1395, Dec. 2013. 103

[198] E. Y. Liu, A. P. Morgan, E. J. Chesler, W. Wang, G. A. Churchill, and F. Pardo-Manuel de Villena, "High-Resolution Sex-Specific Linkage Maps of the Mouse Reveal Polarized Distribution of Crossovers in Male Germline," *Genetics*, Feb. 2014. 103, 112, 124, 127, 128, 150

[199] C. Gregg, J. Zhang, J. E. Butler, D. Haig, and C. Dulac, "Sex-specific parent-of-origin allelic expression in the mouse brain," *Science*, vol. 329, pp. 682–685, Aug. 2010. 111

[200] C. D. Eversley, T. Clark, Y. Xie, J. Steigerwalt, T. A. Bell, F. P. de Villena, and D. W. Threadgill, "Genetic mapping and developmental timing of transmission ratio distortion in a mouse interspecific backcross," *BMC genetics*, vol. 11, no. 1, p. 98, 2010. 112, 113, 124

[201] A.-L. Steckelberg, V. Boehm, A. M. Gromadzka, and N. H. Gehring, "CWC22 connects pre-mRNA splicing and exon junction complex assembly," *Cell reports*, vol. 2, pp. 454–461, Sept. 2012. 113

[202] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch, "A

gene atlas of the mouse and human protein-encoding transcriptomes," *PNAS*, vol. 101, pp. 6062–6067, Apr. 2004. 113

[203] J. R. Wang, F. Pardo-Manuel de Villena, and L. McMillan, "Comparative analysis and visualization of multiple collinear genomes," *BMC Bioinformatics*, vol. 13 Suppl 3, p. S13, 2012. 114

[204] R. Greene-Till, Y. Zhao, and S. C. Hardies, "Gene flow of unique sequences between *Mus musculus domesticus* and *Mus spretus*," *Mammalian genome*, vol. 11, pp. 225–230, Mar. 2000. 114

[205] Y. Song, S. Endepols, N. Klemann, D. Richter, F.-R. Matuschka, C.-H. Shih, M. W. Nachman, and M. H. Kohn, "Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice," *Current biology*, vol. 21, pp. 1296–1301, Aug. 2011. 114, 148

[206] D. Weichenhan, B. Kunze, H. Winking, M. van Geel, K. Osoegawa, P. J. de Jong, and W. Traut, "Source and component genes of a 6-200 Mb gene cluster in the house mouse," *Mammalian genome*, vol. 12, pp. 590–594, Aug. 2001. 115

[207] W. Traut, I. M. Rahn, H. Winking, B. Kunze, and D. Weichehan, "Evolution of a 6-200 Mb long-range repeat cluster in the genus *Mus*," *Chromosoma*, vol. 110, pp. 247–252, Aug. 2001. 115

[208] R. D. Hernandez, J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. A. T. McVean, 1000 Genomes Project Consortium, G. Sella, and M. Przeworski, "Classic selective sweeps were rare in recent human evolution," *Science*, vol. 331, pp. 920–924, Feb. 2011. 121

[209] A. R. Rogala, A. P. Morgan, A. M. Christensen, T. J. Gooch, T. A. Bell, D. R. Miller, V. L. Godfrey, and F. Pardo-Manuel de Villena, "The Collaborative Cross as a Resource for Modeling Human Disease: CC011/Unc, a New Mouse Model for Spontaneous Colitis," *Mammalian genome*, pp. 1–14, 2014. 124, 150

[210] P. M. Petkov, Y. Ding, M. A. Cassell, W. Zhang, G. Wagner, E. E. Sargent, S. Asquith, V. Crew, K. A. Johnson, P. Robinson, V. E. Scott, and M. V. Wiles, "An efficient SNP system for mouse genome scanning and elucidating strain relationships," *Genome research*, vol. 14, pp. 1806–1811, Sept. 2004. 125

[211] E. Y. Liu, Q. Zhang, L. McMillan, F. Pardo-Manuel de Villena, and W. Wang, "Efficient genome ancestry inference in complex pedigrees with inbreeding," *Bioinformatics*, vol. 26, pp. i199–207, June 2010. 126, 127, 154

[212] W. J. Kent, "BLAT–the BLAST-like alignment tool," *Genome research*, vol. 12, pp. 656–664, Apr. 2002. 127

[213] C. S. Haley and S. A. Knott, "A simple regression method for mapping quantitative trait loci in line crosses using flanking markers," *Heredity*, vol. 69, no. 4, pp. 315–324, 1992. 128

[214] K. W. Broman, H. Wu, Ś. Sen, and G. A. Churchill, "R/qtl: QTL mapping in experimental crosses," *Bioinformatics*, vol. 19, no. 7, pp. 889–890, 2003. 128

[215] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078–2079, Aug. 2009. 129

[216] T. A. Bell, E. de la Casa-Esperón, H. E. Doherty, F. Y. Ideraabdullah, K. Kim, Y. Wang, L. A. Lange, K. Wilhemsen, E. M. Lange, C. Sapienza, and F. Pardo-Manuel de Villena, "The paternal gene of the DDK syndrome maps to the Schlafen gene cluster on mouse chromosome 11," *Genetics*, vol. 172, pp. 411–423, Jan. 2006. 133

[217] F. Y. Ideraabdullah, K. Kim, J. L. Moran, D. Beier, and F. Pardo-Manuel de Villena, "Rescue of the mouse DDK syndrome by parent-of-origin-dependent modifiers," *Biology of reproduction*, vol. 76, pp. 286–293, Jan. 2007. 133

[218] E. Kaszás and J. A. Birchler, "Meiotic transmission rates correlate with physical features of rearranged centromeres in maize," *Genetics*, vol. 150, pp. 1683–1692, Dec. 1998. 135

[219] B. Mantovani, F. Tinti, L. Bachmann, and V. Scali, "The *Bag320* satellite DNA family in *Bacillus* stick insects (Phasmatodea): different rates of molecular evolution of highly repetitive DNA in bisexual and parthenogenic taxa," *Molecular Biology and Evolution*, vol. 14, pp. 1197–1205, Dec. 1997. 135

[220] E. Novitski, "Non-Random Disjunction in *Drosophila*," *Genetics*, vol. 36, p. 267, May 1951. 136

[221] E. Novitski, "Nonrandom Disjunction in *Drosophila*," *Annual review of genetics*, vol. 1, pp. 71–86, Dec. 1967. 136

[222] T. R. Gregory, "Insertion-deletion biases and the evolution of genome size," *Gene*, vol. 324, pp. 15–34, Jan. 2004. 136

[223] E. L. Walker, T. P. Robbins, T. E. Bureau, J. Kermicle, and S. L. Dellaporta, "Transposon-mediated chromosomal rearrangements and gene duplications in the formation of the maize R-r complex," *The EMBO journal*, vol. 14, pp. 2350–2363, May 1995. 136

[224] M. Cáceres, J. M. Ranz, A. Barbadilla, M. Long, and A. Ruiz, "Generation of a widespread *Drosophila* inversion by a transposable element," *Science*, vol. 285, pp. 415–418, July 1999. 136

[225] R. J. Mroczek, J. R. Melo, A. C. Luce, E. N. Hiatt, and R. K. Dawe, "The maize Ab10 meiotic drive system maps to supernumerary sequences in a large complex haplotype," *Genetics*, vol. 174, pp. 145–154, Sept. 2006. 136

[226] M. J. Justice and V. C. Bode, "Genetic analysis of mouse *t* haplotypes using mutations induced by ethylnitrosourea mutagenesis: the order of *T* and *qk* is inverted in *t* mutants," *Genetics*, vol. 120, pp. 533–543, Oct. 1988. 136

[227] C. Heemert, "Somatic pairing and meiotic nonrandom disjunction in a pericentric inversion of *Hylemya antiqua* (Meigen)," *Chromosoma*, vol. 59, no. 3, pp. 193–206, 1977. 136

[228] G. G. Foster and M. J. Whitten, "Meiotic drive in Lucilia cuprina and chromosomal evolution," *American Naturalist*, pp. 403–415, 1991. 136

[229] B. Dutrillaux, "Chromosomal evolution in primates: tentative phylogeny from *Microcebus murinus* (Prosimian) to man," *Human Genetics*, vol. 48, pp. 251–314, May 1979. 136

[230] T. J. Robinson, "Mammalian Chromosomal Evolution: From Ancestral States to Evolutionary Regions," *Evolutionary Biology-Concepts*, 2010. 138

[231] J. Britton-Davidian, J. Catalan, G. Ganem, J. C. Auffray, R. Capela, M. Biscoito, J. B. Searle, and M. de Luz Mathias, "Rapid chromosomal evolution in island mice," *Nature*, vol. 403, no. 6766, p. 158, 2000. 138

[232] L. A. Underkoffler, L. E. Mitchell, A. R. Localio, S. M. Marchegiani, J. Morabito, J. N. Collins, and R. J. Oakey, "Molecular analysis of nondisjunction in mice heterozygous for a Robertsonian translocation," *Genetics*, vol. 161, no. 3, pp. 1219–1224, 2002. 140

[233] S. Eaker, A. Pyle, J. Cobb, and M. A. Handel, "Evidence for meiotic spindle checkpoint from analysis of spermatocytes from Robertsonian-chromosome heterozygous mice," *Journal of cell science*, vol. 114, pp. 2953–2965, Aug. 2001. 140

[234] T. Chiang, R. M. Schultz, and M. A. Lampson, "Meiotic origins of maternal age-related aneuploidy," *Biology of reproduction*, vol. 86, pp. 1–7, Jan. 2012. 140

[235] J. Nielsen and M. Wohlert, "Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark," *Human Genetics*, vol. 87, pp. 81–83, May 1991. 140

[236] S. Kolgeci, J. Kolgeci, M. Azemi, R. Shala, A. Daka, and M. Sopjani, "Reproductive Risk of the Silent Carrier of Robertsonian Translocation," *Medical Archives*, vol. 67, no. 1, pp. 56–59, 2013. 140

[237] A. Daniel, "Distortion of female meiotic segregation and reduced male fertility in human Robertsonian translocations: consistent with the centromere model of co-evolving centromere DNA/centromeric histone (CENP-A)," *American Journal of Medical Genetics*, vol. 111, pp. 450–452, Sept. 2002. 140

[238] A. Guffei, Z. Lichtensztejn, A. Gonçalves dos Santos Silva, S. F. Louis, A. Caporali, and S. Mai, "*c-Myc*–Dependent Formation of Robertsonian Translocation Chromosomes in Mouse Cells," *Neoplasia*, vol. 9, no. 7, pp. 578–588, 2007. 140

[239] R. Garcia-Cruz, A. Casanovas, M. Brieño-Enríquez, P. Robles, I. Roig, A. Pujol, L. Cabero, M. Durban, and M. Garcia Caldés, "Cytogenetic analyses of human oocytes provide new data on non-disjunction mechanisms and the origin of trisomy 16," *Human reproduction (Oxford, England)*, vol. 25, pp. 179–191, Jan. 2010. 141

[240] R. S. Verma, H. Dosik, H. A. Lubs, and U. Francke, "Size and pericentric inversion heteromorphisms of secondary constriction regions (h) of chromosomes 1, 9, and 16 as detected by CBG technique in Caucasians: Classification, frequencies, and incidence," *American Journal of Medical Genetics*, vol. 2, no. 4, pp. 331–339, 1978. 141

[241] S. H. Williamson, M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante, and R. Nielsen, "Localizing Recent Adaptive Evolution in the Human Genome," *PLoS Genetics*, vol. 3, p. e90, June 2007. 141

[242] F. Romera, M. M. Jiménez, and M. J. Puertas, "Genetic control of the rate of transmission of rye B chromosomes. I. Effects in 2B × 0B crosses," *Heredity*, vol. 66, pp. 61–65, Feb. 1991. 142

[243] O. Mihola, Z. Trachtulec, C. Vlcek, J. C. Schimenti, and J. Forejt, "A mouse speciation gene encodes a meiotic histone H3 methyltransferase," *Science*, vol. 323, pp. 373–375, Jan. 2009. 143, 148

[244] B. L. Dumont and B. A. Payseur, "Genetic analysis of genome-scale recombination rate evolution in house mice," *PLoS Genetics*, vol. 7, p. e1002116, June 2011. 143

[245] P. M. Borodin, I. P. Gorlov, A. I. Agulnik, S. I. Agulnik, and A. O. Ruvinsky, "Chromosome pairing and recombination in mice heterozygous for different translocations in chromosomes 16 and 17," *Chromosoma*, vol. 101, no. 4, pp. 252–258, 1991. 143

[246] F. Pardo-Manuel de Villena and C. Sapienza, "Recombination is proportional to the number of chromosome arms in mammals," *Mammalian genome*, vol. 12, no. 4, pp. 318–322, 2001. 144

[247] L. H. Wong and K. H. A. Choo, "Evolutionary dynamics of transposable elements at the centromere," *Trends in genetics*, vol. 20, pp. 611–616, Dec. 2004. 144

[248] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. A. Grant, H. Hakonarson, and M. Bucan, "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data," *Genome research*, vol. 17, pp. 1665–1674, Nov. 2007. 151, 153

[249] J. Staaf, J. Vallon-Christersson, D. Lindgren, G. Juliusson, R. Rosenquist, M. Höglund, Å. Borg, and M. Ringnér, "Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios," *BMC Bioinformatics*, vol. 9, no. 1, p. 409, 2008. 151

[250] S. Vattathil and P. Scheet, "Haplotype-based profiling of subtle allelic imbalance with SNP arrays," *Genome research*, vol. 23, pp. 152–158, Jan. 2013. 153

[251] J. R. Wang, F. Pardo-Manuel de Villena, H. A. Lawson, J. M. Cheverud, G. A. Churchill, and L. McMillan, "Imputation of single-nucleotide polymorphisms in inbred mice using local phylogeny," *Genetics*, vol. 190, pp. 449–458, Feb. 2012. 155

[252] E. D. O. Roberson and J. Pevsner, "Visualization of shared genomic regions and meiotic recombination in high-density SNP data," *PLoS One*, vol. 4, no. 8, p. e6711, 2009. 156

[253] W.-C. Lee, "Testing the genetic relation between two individuals using a panel of frequency-unknown single nucleotide polymorphisms," *Annals of human genetics*, vol. 67, pp. 618–619, Nov. 2003. 156