Stained-glass windows on the evolutionary process: Patterns of genomic ancestry in hybrid zones

By

Megan E. Frayer

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Genetics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral examination: 8/1/2022

The dissertation is approved by the following members of the Final Oral Committee:

Bret A. Payseur, Professor, Genetics and Medical Genetics
David A. Baum, Professor, Botany
Colin N. Dewey, Professor, Biostatistics and Medical Informatics
Nicole T. Perna, Professor, Genetics
John E. Pool, Professor, Genetics

Table of Contents

Table of Tables

Table of Figures

Acknowledgements

There are many people that I need to thank for getting me through this Ph.D. journey. First and foremost, I need to thank my advisor, Bret. I cannot understate the impact that his mentorship has had on me. Bret, thank you for all of your guidance, encouragement and conversation. Thank you for not giving up on me even when I had given up on myself. I could not have made it through the pandemic without your understanding and support. I also want to thank all of the Payseur lab members that I have learned from over the years. Peicheng, Mark, Amy, Alex, John, Richard, Michelle, Lauren, April, Jered, Mikey, Kevin and Emma -- it has been a pleasure to work with you. Thank you for your help, feedback, support, and camaraderie. I will sincerely miss our lab meetings, book clubs, and ice cream trips.

Madison has been an amazing place to be a graduate student. I have learned so much from the Genetics and Evolution communities, and I have been honored to be a part of them. Thank you for sharing your science with me, teaching me, mentoring me, listening to my talks, and even going along with some of my crazy ideas. I would especially like to thank my committee for their feedback, support, encouragement, and for smiling throughout my many rambling committee meetings. I also want to thank the CHTC, especially Lauren and Christina, for making my thesis research possible. Taking on a computational project was a steep learning curve for me, but your help allowed me to take on huge projects and gain skills I didn't think were possible. Thank you for the time you have spent working with me, and for the over 8 million computational hours that went into this dissertation.

Finally, I want to thank my family and friends. Khailee, thank you for keeping me sane. Jake, thank you for supporting me and always being ready to go along for unpredictable ride that is a research career. Mom, Dad, and Katie, thank you for making me who I am and for making everything in my life possible. Thank you so much to my aunts, uncles, cousins, grandparents, and in-laws for supporting and encouraging me every step of the way.

Abstract

The outcome of speciation is put to the test in regions where incipient species' ranges overlap and hybrid offspring are produced, known as hybrid zones. Delineating the parental species that contributed each part of a hybrid genome, the ancestry, can provide insights into the history of hybrid individuals and the progression of speciation between these lineages. The genomic locations where regions of different ancestry meet are called junctions. Junctions are historical recombination events that occurred in hybrids, and thus are records of the history of the zone in which they are found. In this dissertation, I explore the effect of demography on junctions and then investigate junctions in a classic hybrid zone. I find that junctions are very responsive to demographic history, particularly the age of the hybrid zone, the population size, and the rate of migration within a stepping-stone model. I also find that population substructure uniquely impacts ancestry in a way that is not captured by single population models. I also introduce a novel metric to describe the sharing of junctions between individuals—the junction frequency spectrum. I then review my attempt to integrate the junction frequency spectrum and other metrics of ancestry into an inference method using approximate Bayesian computation. My approach is unable to accurately estimate parameters, but I outline suggestions that might allow this inference to be successful in the future. Finally, I identify hundreds of thousands of junctions in mice from two populations within the European house mouse hybrid zone between *Mus musculus domesticus* and *M. m. musculus*. I show that these junctions are variable between chromosomes and across the genome. I present the first estimates of the junction frequency spectrum in a hybrid population. The populations that I survey show unique patterns of ancestry and junction sharing, indicating that they have distinct histories. I use junctions along with the site frequency spectrum to infer the age of the hybrid zone as a few thousand years. Finally, I identify several regions of the genome that are outliers for ancestry metrics as candidate regions for incompatibilities between *M. m. musculus* and *M. m. domesticus*.

Chapter 1

Introduction

Speciation is the evolutionary process which codifies varieties into distinct evolutionary branches—generating new species, and, in turn, generating the immense biodiversity of our planet. The question of how new species arise has long puzzled biologists. It is the title and unaddressed question of Darwin's seminal "abstract"—*On the Origin of Species* (Darwin, 1859). Nearly 80 years later, the architects of the Modern Synthesis recognized the importance of speciation as they grappled with the integration of genetics and natural history. Dobzhansky wrote that "The origin and functioning of the isolating mechanisms constitute one of the most important problems of the genetics of populations" (pg. 14, Dobzhansky 1937). Despite years of investigation and progress, there are still many gaps in our understanding of this fundamental process. It has been observed that, even as our molecular and genomic tools advance, the questions we ask remain the same (Harrison and Larson, 2014).

Part of what makes speciation difficult to study is that it is difficult to define. What is a species? When is it fully distinguished from another closely related species? There are many competing species concepts, defining species based on diverse criteria such as breeding system, ecology and phylogeny (Coyne and Orr, 2004). Reclassification between species and subspecies is common. For example, the house mouse, one of the most ubiquitous small mammals, has been classified as anywhere from 1 to 133 distinct species, with varieties fluctuating between species and subspecies (Berry and Bronson, 1992; Boursot et al., 1993). While there is no one-size-fits-all definition, mammologists most often think about species barriers in terms of reproductive isolation—barriers that prevent the interbreeding of lineages and prevent their collapse into a single panmictic gene pool (largely in line with the Biological Species Concept; Mayr 1942). When the barriers to reproduction are incomplete, distinct lineages may produce offspring—known as hybrids. Hybrids exist at the boundaries of our biological definitions, making them a powerful but messy source of information about species and speciation (Harrison, 1993).

Reproductive isolation can have many components that act at different stages of reproduction (Coyne and Orr, 2004). Prezygotic barriers prevent some or all interbreeding between lineages,

preventing hybrids. Postzygotic barriers reduce the fitness of hybrid offspring, either by reducing their viability or fertility. Barriers can be extrinsic (driven by the environmental context) or intrinsic (innate to the biology of the organism). Different forms of reproductive isolation may act at different stages along the speciation continuum, although support for this idea is mixed. For example, intrinsic post-zygotic barriers are often thought of as late-stage speciation barriers (Coyne and Orr, 2004), but they can evolve rapidly or early in speciation (Coughlan and Matute, 2020). Furthermore, different types of reproductive isolation may vary in their ability to affect the outcome of speciation. For example, the strength required for pre-zygotic barriers to prevent species collapse is larger than for reductions in hybrid fitness (Irwin, 2020).

All forms of reproductive isolation can influence the genomes of the lineages involved. In cases where hybrids cannot be produced, genomes of the divergent lineages will reflect that lack of gene flow. Individual mutations that produce maladaptive or low-fitness hybrid phenotypes may be the target of selection and show distinct patterns such as low introgression across a geographic region (Barton and Hewitt, 1985). A major focus of speciation research has been identifying genes that may contribute to reproductive isolation between species, sometimes called “barrier loci” (Maheshwari and Barbash, 2011).

The Dobzhansky-Muller model is one of the most widely studied models of barrier loci. This model posits that mutations in interacting genes fix in independent lineages, creating defects when the mutations come together in the hybrid offspring of those lineages and imposing a reproductive barrier between the parent lineages (Dobzhansky, 1936; Muller, 1942). These incompatibilities can accumulate faster-than-linearly over divergence time, exhibiting a “snowball effect” (Orr, 1995). The Dobzhansky-Muller model is useful and appealing in its simplicity, but there are several questions outstanding—What is the role of polymorphism in incompatibility loci (Cutter, 2012)? It is generally believed that the number of incompatibility loci must be large to reduce gene flow across the entire genome (Barton and

Bengtsson, 1986), and the efficacy of such barriers in the face of gene flow have been questioned (Bank et al., 2012; Lindtke and Buerkle, 2015). Furthermore, the genetic signature of these incompatibilities can be complex. The expectation that incompatibilities will lead to linkage disequilibrium in ancestry can be difficult to measure, even with large sample sizes (Schumer and Brandvain, 2016).

Speciation is also expected to have a unique impact on the X chromosome. The X chromosome is distinct from autosomes from a population genetic standpoint in that it undergoes recombination less frequently and has a reduced effective population size. The observation that the X chromosome often has a disproportionate effect on hybrid incompatibility is dubbed the large X effect (Coyne and Orr, 1989, 2004). Hypotheses for the cause of this effect have included: impact of hemizygosity on effect sizes and selection (Muirhead and Presgraves, 2016; Turelli and Orr, 1995); higher density of incompatibilities on the X (Coyne and Orr, 1989; Masly and Presgraves, 2007); increased non-synonymous evolution (Charlesworth et al., 1987), meiotic drive (Tao and Hartl, 2003) or changes in gene position (Moyle et al., 2010) on the X relative to the autosomes; disruptions in the X-specific processes of dosage compensation (Orr, 1989) or X inactivation (Larson et al., 2017; Masly and Presgraves, 2007). The analogous effect has been seen on the Z chromosome in ZW systems (Irwin, 2018). Any or all of these factors could contribute to a disproportionate role for the X chromosome, and could lead to reduced introgression of the X chromosome, which is often observed (Presgraves, 2018). Along with the unique effect on the X chromosome itself, a general asymmetrical effect on the hybrid sexes has also been observed—the hemizygous sex is more likely to experience hybrid sterility or inviability (Haldane, 1922). Haldane's rule and the large X effect are considered to be two rules of speciation (Coyne and Orr, 1989, 2004).

A common approach to studying speciation is the use of "natural laboratories"—hybrid zones (Hewitt, 1988). Hybrid populations are pools of recombinant individuals that allow new genetic combinations between the parental strains to form and be investigated (Barton and Gale, 1993;

Harrison and Larson, 2014). Some hybrid zones are tension zones, maintained by a balance between selection on hybrids and migration into the hybrid zone from non-hybrids (Barton and Hewitt, 1985). Others are more transient or patchy, sometimes called mosaic hybrid zones (Rand and Harrison, 1989). Introgression across hybrid zones has been classically studied with clines (Endler, 1977). Geographic clines describe introgression over the physical space of a hybrid zone (Barton, 1983; Barton and Gale, 1993; Szymura and Barton, 1986). More recently, a complementary genomic cline approach has been used to describe variation in introgression within the genomes of hybrid individuals (Gompert and Buerkle, 2011). Their ability to apply such approaches and gain insights into the ongoing evolutionary forces have lead hybrid zones them to be called “windows” on the evolutionary process (Harrison, 1990).

Ancestry can be used to summarize the unique effects of hybridization on the genome. Patterns of ancestry are inferred by categorizing parts of the genome based on the parental source population from which they originated. As two species diverge, the ability to differentiate their genomic identity increases. Recombination in hybrids mixes these ancestries along the genome. If sufficient divergence between the parental taxa has occurred, it is possible to describe this pattern. Numerous methods exist for inferring ancestry at individual sites along the genome (Baran et al., 2012; Brisbin et al., 2012; Corbett-Detig and Nielsen, 2017; Guan, 2014; Price et al., 2009; Wegmann et al., 2011). Ancestry-based approaches can be used to reconstruct selection and demography (Gompert and Buerkle, 2013), and have been used to explore evolutionary history in humans (Bycroft et al., 2019; Hellenthal et al., 2014; Henn et al., 2012; Sankararaman et al., 2014) and many other species (Chiou et al., 2021; Duranton et al., 2019; Galaverni et al., 2017; Kenney and Sweigart, 2016; Lavretsky et al., 2019; Leitwein et al., 2018; Suarez-Gonzalez et al., 2016; Wang et al., 2017). Much of this work focused on ancestry proportions.

First explored by Fisher (Fisher, 1949, 1954) in the context of inbreeding, junctions are break points between segments of different ancestry along chromosomes. They are inherently informative

about hybridization because they are the direct results of recombination events in ancestral hybrid individuals. Once formed, junctions are inherited like point mutations (Fisher, 1954), making them useful records of the hybridization history that will be impacted by selection and demography.

Junctions and tract lengths (the distances between junctions) are responsive to demography and selection. Many types of selection, including underdominance and epistatic incompatibilities, can lead to regions of reduced junction density around a selected locus within a hybrid genome (Hvala et al., 2018) or increased tract length (Sedghifar et al., 2016). Tract lengths decrease as junction density increases over time (Baird, 1995; Baird et al., 2003; Liang and Nielsen, 2014; Pool and Nielsen, 2009). Junction density also responds to migration—migration slows the breakdown of tracts (Gravel, 2012; Pool and Nielsen, 2009), whereas an isolated population will quickly approach an equilibrium junction number (Chapman and Thompson, 2002). Population size can also affect the rate of junction formation and the time required to reach an equilibrium (Janzen et al., 2018).

Junctions may be particularly well-suited to the investigation of genomes in hybrid zones. First, junctions are direct markers of hybridization—unlike sequence variations, they cannot arise outside the context of hybridization. Second, junction patterns may record incompatibility alleles even after they have been removed from the population. This arises in situations under which incompatibility loci lead to the fixation of a compatible variant and thus the removal of the incompatibility from the population (Lindtke and Buerkle, 2015). Simulations in complex population structures with migration have suggested that even after the incompatible allele is gone, the reduction in heterozygous ancestry (called heterogenicity) and junction number at the locus can persist (Hvala et al., 2018).

Empirical investigation of ancestry patterns is difficult, which has limited its utilization in some taxa. Ancestry must be inferred, which requires sampling of reference populations. Approaches that rely on track lengths require phasing—determining which variants belong together on which homologous chromosome. However, in many cases, statistical phasing relies on the same patterns that we would like

to investigate and thus confounds our inference of ancestry. This limits the application of these methods to species with familial knowledge (Chiou et al., 2021), populations with little heterozygosity (Pool, 2015; Pool and Nielsen, 2009), or other scenarios where knowledge of the population makes statistical phasing possible. Junction identity, on the other hand, can be determined without phase.

Exploring the natural context of junctions is aided by a general understanding of both hybridization history and the genome of the organisms of interest. House mice make a strong subject for this type of study due to their status as a genetic model organism. Mice are one of the oldest and best characterized models for biomedical research (Phifer-Rixey and Nachman, 2015). As a result, we have extensive resources not available in many other natural systems, including a high quality reference genome (Church et al., 2009), detailed recombination maps (Cox et al., 2009), and extensive genome annotation resources (Blake et al., 2011).

House mice originated in Southwestern Asia, and spread out over most of the globe, while diversifying into several subspecies (Boursot et al., 1993; Duvaux et al., 2011; Phifer-Rixey and Nachman, 2015; Suzuki et al., 2013). In Europe, there are two subspecies: *Mus musculus musculus* (henceforth referred to as *musculus*) came through Russia into Eastern Europe, while *M. m. domesticus* (henceforth *domesticus*) came into Western Europe via North Africa and the Mediterranean. Their modern ranges meet at a hybrid zone that runs from Norway to the Black Sea (Boursot et al., 1993; Jones et al., 2010). House mice first colonized Central Europe ~6,000 years ago, setting a likely ceiling for the age of the hybrid zone (Cucchi et al., 2005, 2011). The hybrid zone was first described morphologically in Central Europe (Zimmerman, 1949) and Denmark (Ursin, 1952); it was first characterized genetically by Hunt and Selander (1973). There is differential movement of alleles across the hybrid zone (Janoušek et al., 2012; Macholán et al., 2007; Payseur et al., 2004; Teeter et al., 2008, 2010), including low movement of the X chromosome relative to the autosomes (Macholán et al., 2007; Tucker et al., 1992). The zone does

not align with any known ecological gradient (Baird and Macholan, 2012), and may be moving westward into *domesticus* territory (Wang et al., 2011).

Several forms of reproductive isolation contribute to the semipermeable barrier between *musculus* and *domesticus*. The most studied barrier is reduced hybrid fertility. Hybrid mice show reduced fertility in the wild (Albrechtová et al., 2012; Turner et al., 2012), but the mechanisms are best understood in F_1 hybrids produced in the lab. Males with a C57BL/6J (a mainly *domesticus* classical inbred strain) father and PWD (a wild-derived *musculus* inbred strain) mother are completely sterile due to an incompatibility involving *Prdm9*—the only known mammalian speciation gene (reviewed by Forejt, Jansa, and Parvanov 2021)—as well as elements on the PWD X chromosome, and a sufficient level of heterozygosity on the autosomes (Gregorova et al., 2018). The alleles involved in this incompatibility are variable in house mice (Britton-Davidian et al., 2005; Larson et al., 2018; Mukaj et al., 2020), sometimes highly variable over relatively small distances (Vara et al., 2019), and several *Prdm9* variants are capable of causing sterility (Mukaj et al., 2020).

While *Prdm9* is the best studied gene that contributes to reproductive isolation between *musculus* and *domesticus*, it is likely one of many incompatibilities between these subspecies. Male hybrids display a wide range of sterility traits such as reduced testis weight (Dzur-Gejdosova et al., 2012; Turner and Harr, 2014; Turner et al., 2012) and sperm defects—including sperm morphology abnormalities, reduced sperm count, and reduced sperm motility (Dzur-Gejdosova et al., 2012; Good et al., 2008; Turner et al., 2012; White et al., 2011)—which have been characterized at multiple stages of spermatogenesis (Schwahn et al., 2018). QTL have been mapped in several crosses for these traits. While most of this reduced fertility is found in males as predicted by Haldane's Rule (Haldane, 1922), there is also reduced fertility found in females (Bhattacharyya et al., 2014; Suzuki and Nachman, 2015). Hybrid mice may also experience developmental instability (Auffray et al. 1996; but see Mikula and Macholán 2008), or increased parasite load (Sage et al. 1986; but see Balard and Heitlinger 2022).

Isolation between the parental taxa may be further driven by differences in sperm competition (Dean and Nachman, 2009) or by assortative mate choice (Laukaitis et al., 1997; Smadja and Ganem, 2002, 2005). Reinforcement against hybridization may be developing in some parts of the zone (Hurst et al., 2017; Latour et al., 2014; Loire et al., 2017). Despite all these avenues of reproductive isolation, no F_1 hybrid has ever been found in the hybrid zone, and the hybrids that do occur are highly recombinant. In context of the steep clines across the hybrid zone, it is clear that we do not have a complete picture of the strength of isolation between these species.

This dissertation aims to expand our understanding of ancestry and junctions in natural hybrid populations. Chapter 2 expands our theoretical understanding of junctions. I explore the behavior of junctions within a complex population structure with varying demographic histories. Chapter 2 also explores applicable summary statistics of ancestry patterns and introduces a novel descriptor: the junction frequency spectrum. Chapter 3 attempts to integrate these summary statistics into a simulation-based inference method. This method is not successful, but I offer insights about how to better develop such a method in the future. Finally, Chapter 4 applies the ancestry framework to the European house mouse hybrid zone to better understand the complex history of admixture that has occurred. The junction frequency spectrum is estimated for the first time in a hybrid population. I present candidate incompatibility regions based on our metrics of ancestry.

Chapter 2

Demographic history shapes genomic ancestry in hybrid zones

This chapter was published as a peer reviewed article in *Ecology and Evolution*.

Frayer, M.E., and Payseur, B.A. (2021). Demographic history shapes genomic ancestry in hybrid zones. Ecology and Evolution *11*, 10290–10302. https://doi.org/10.1002/ece3.7833.

Contributions were as follows: M.E.F. and B.A.P. designed the study. M.E.F. conducted simulations and analyses. M.E.F. and B.A.P. wrote the paper.

Abstract

Demographic factors such as migration rate and population size can impede or facilitate speciation. In hybrid zones, reproductive boundaries between species are tested and demography mediates the opportunity for admixture between lineages that are partially isolated. Genomic ancestry is a powerful tool for revealing the history of admixed populations, but models and methods based on local ancestry are rarely applied to structured hybrid zones. To understand the effects of demography on ancestry in hybrids zones, we performed individual-based simulations under a stepping-stone model, treating migration rate, deme size, and hybrid zone age as parameters. We find that the number of ancestry junctions (the transition points between genomic regions with different ancestries), as well as heterogenicity (the genomic proportion heterozygous for ancestry), are often closely connected to demographic history. Reducing deme size reduces junction number and heterogenicity. Elevating migration rate increases heterogenicity, but migration affects junction number in more complex ways. We highlight the junction frequency spectrum as a novel and informative summary of ancestry that responds to demographic history. A substantial proportion of junctions are expected to fix when migration is limited or deme size is small, changing the shape of the spectrum. Our findings suggest that genomic patterns of ancestry could be used to infer demographic history in hybrid zones.

Introduction

During speciation, lineages independently accumulate genetic variation. When these lineages come back into contact and mate, certain combinations of mutations may reduce the fitness of hybrids and confer barriers to reproduction between the lineages. If a balance is reached between migration and selection against hybrids in the region of contact, a stable population structure can form, with demes containing individuals of mixed ancestry bridging the gap between the original lineages.

Migration across this metapopulation, known as a hybrid zone, shapes its dynamics by controlling the flow of alleles. Migration tends to homogenize allele frequencies between demes (Wright, 1931), slowing genomic divergence. Restricting migration in a subdivided metapopulation can facilitate local adaptation, whereas demes begin to behave as a single panmictic population when migration is high (Barton and Whitlock, 1997; Maruyama and Kimura, 1980; Whitlock and Barton, 1997). Hybrid zones may additionally feature migration from source populations, which replenishes chromosomes and combinations of alleles found outside of the region of contact (Barton, 1979a; Feldman and Christiansen, 1974; Harrison, 1990). Stable hybrid zones are often described using a tension zone model, under which the balance between migration across the zone and selection against hybrids maintains sigmoidal clines in allele frequencies (Barton, 1979a; Barton and Hewitt, 1985). In this scenario, migration works in direct opposition to selection to establish hybrid zone structure.

The dynamics of a hybrid zone are also governed by genetic drift. Drift increases the variance in allele frequency among demes, leading to steeper clines, even at neutral loci (Polechová and Barton, 2011). In addition, drift reduces the efficacy of selection in small populations (Kimura et al., 1963). Drift may be a strong force in hybrid zones because there are reasons to suspect that hybrid populations will often be small. If selection against hybrids is strong, low hybrid fitness could limit population growth. Hybrid zones generally occur at the edges of species ranges, which can have low population densities

(Bridle and Vines, 2007). Furthermore, range edges can be highly fragmented (Bridle and Vines, 2007), which may reduce migration and strengthen drift. Finally, drift is a potential explanation for instances of variable hybridization outcomes across unique meetings of the same species (Mandeville et al., 2017).

The genomic pattern underlying the observed genetic variation in a hybrid zone is ancestry. Over time, the configuration of ancestry along chromosomes changes as meiotic recombination breaks down segments inherited from each lineage. This expectation has spurred the creation of methods that use inferred ancestry to estimate admixture time (Corbett-Detig and Nielsen, 2017; Liang and Nielsen, 2014; Medina et al., 2018; Moorjani et al., 2011; Pool and Nielsen, 2009). Migration generally acts on ancestry in opposition to recombination and time, by replacing some of the chromosomes that were previously shuffled by recombination. Drift affects the rate at which ancestry patterns are fixed or lost from a hybrid population. Incorporating drift into analytical models better captures the behavior of ancestry in admixed populations (Gravel, 2012).

Patterns of local ancestry in individuals can be described by inferring the genomic locations of junctions. Junctions are transition points between tracts of alternative ancestries along a chromosome, first used by Fisher (1949, 1954) to model the effects of inbreeding. Junctions are formed by recombination events between chromosomes with different ancestries, and are inherited like point mutations (Fisher, 1954). Junctions can be counted, and the distances between junctions ("tract lengths") can be measured. The distribution of tract lengths is the inverse of the distribution of junction density.

Junction density and tract length respond to demographic history and selection (Gravel, 2012; Hvala et al., 2018; Janzen et al., 2018; Pool and Nielsen, 2009). Junction density increases as tracts shorten over time (Liang and Nielsen, 2014; Pool and Nielsen, 2009). In an isolated population, junction density approaches an equilibrium value, and the population converges on a single ancestry pattern (Chapman and Thompson, 2002). Analytical models show that migration leads to longer tract lengths

(Gravel, 2012; Pool and Nielsen, 2009). In a hybrid swarm model, increasing population size raises the rate of junction formation and the time required to reach the maximum number of junctions (Janzen et al., 2018).

Although ancestry-based approaches are increasingly used to reconstruct demography in humans (Bryc et al., 2010; Bycroft et al., 2019; Hellenthal et al., 2014) and other species (Lavretsky et al., 2019; Leitwein et al., 2018), their application to hybrid zones between divergent lineages remains limited compared to analyses of allele frequency clines (Gompert et al., 2017; Payseur and Rieseberg, 2016). Junction-based methods have been applied to understand hybrid speciation (Buerkle and Rieseberg, 2008; Ungerer et al., 1998), but not in structured hybrid zones. Part of the explanation for this deficit is a lack of specific theoretical predictions for ancestry in hybrid zones with realistic population structure. Existing analytical models of ancestry necessarily make important simplifying assumptions. Chapman and Thompson (2002) assumed a single pulse of admixture, whereas Pool and Nielsen (2009) considered an island model of migration. In a spatially explicit model of admixture, Sedghifar et al. (2015) derived useful expressions for tract length as a function of individual dispersal distance and time since initial contact, but ignored genetic drift. Hvala et al. (2018) simulated ancestry in a stepping-stone framework that considered drift, but explored a limited part of the demographic parameter space.

Given the growing success of ancestry-based frameworks for interpreting genomic data and the insights that hybrid zones provide about speciation, we examined how demographic history affects ancestry in a structured hybrid zone. We performed neutral simulations under a stepping-stone model, emphasizing the effects of time, gene flow and drift on the dynamic behavior of ancestry junctions and heterogenicity (the heterozygosity of ancestry). Our results highlight the sensitivity of ancestry to demography and motivate the application of ancestry-based methods to infer history in real hybrid zones.

Methods

Individual-based, forward-in-time simulations were run using *forqs* (Kessner and Novembre, 2014). This program records recombination events and tracks the founding haplotypes of a population through time. By starting with two haplotypes that represent two source populations, the resulting haplotype blocks can be used to follow ancestry and to determine which recombination events are also ancestry junctions. We analyzed these haplotypes directly, bypassing the generation of nucleotide sequences and assuming perfect knowledge of ancestry.

We simulated a hybrid zone based on the stepping-stone model (Feldman and Christiansen, 1974; Kimura and Weiss, 1964). The stepping-stone model is often used to describe hybrid zones (Barton, 1979b; De La Torre et al., 2015; Dudek et al., 2019; Gavrilets, 1997) and it better captures spatial dynamics than the Wright-Fisher model of admixture that has been employed to examine ancestry in admixed populations (Gravel, 2012; Liang and Nielsen, 2014). We used the `LinearSteppingStone` configuration in *forqs*, which generates a string of subpopulations ("demes") connected by migration. We modeled five hybrid demes connecting two source populations that remained unadmixed (Figure 1). At generation 0, the two source populations were established. Then, the hybrid zone was sequentially filled by individuals from the source populations over the next few generations, and the central hybrid population was formed in generation 4 as a 50/50 mix. Consequently, the first phase of each simulation consisted of initial admixture and eventual loss of parental individuals from the hybrid demes.

At the establishment of each deme, the number of individuals was fixed at a given deme size. Because we used an individual-based simulator, this deme size was the actual number of individuals generated in that deme. All individuals were equally likely to contribute to the next generation.

Individuals reproduced as diploid hermaphrodites. We followed the fate of a single pair of 1 Megabase (Mb) chromosomes. In each non-overlapping generation, chromosomes completed meiosis. A Poisson-distributed number of crossovers was generated, and these crossovers were placed along the chromosomes at random positions drawn from a uniform distribution. Generations were treated as the unit of time, and our results could be applied to organisms with any generation time. For simplicity, we assumed a recombination rate of 0.51 cM/Mb, equal to the genomic average for house mice (Cox et al., 2009), a classic genetic and genomic model for studying reproductive isolation in hybrid zones (Boursot et al., 1993; Sage et al., 1993; Teeter et al., 2010; Tucker et al., 1992; Turner et al., 2014). Because we focused on the effects of demography, natural selection was absent from all simulations. Migration occurred with the same rate across demes, with only neighboring demes exchanging migrants (Figure 1). We assumed no migration back into the source populations; half of the migrants chosen from the outermost hybrid demes were discarded, rather than being transferred to the source populations.

From each simulation, 18 individuals were sampled from the central population. We chose this number as a practical sample size for studies of real hybrid zones, where determination of fine-scale ancestry will typically require whole genome sequencing. One hundred replicates were run for each combination of parameters.

For comparison, simulations were also run under a different model with only one hybrid population receiving migrants from two source populations, hereafter referred to as the “hybrid swarm” model. These simulations were run using the same framework in *forqs*, but with only one hybrid “deme” filled in the first generation of the simulation.

Demography

The stepping-stone model features several parameters that could affect the dynamics of hybridization. We focused on deme size, migration rate, and time since hybrid zone formation. We

conducted simulations for several values of each parameter (Table 1) to determine how these parameters shape genomic ancestry patterns. We chose an initial set of values to cover a broad parameter space, and subsequently added values to further clarify observed patterns. Ranges of parameter values were chosen with actual hybrid zones in mind.

For a subset of parameter combinations, we explored a case without migration. These simulations were run in the stepping-stone framework described above, but once all demes were established, migration between them was eliminated. This approach enabled direct comparison to simulations of the stepping-stone model. These simulations were used in conjunction with simulations of the hybrid swarm model to better understand behavior under the stepping-stone model, as well as to make direct comparison to existing analytical predictions.

Summary Statistics

The output of simulations contained ancestry information about each of the two chromosomes in an individual. Nevertheless, recognizing the many challenges associated with reconstructing haplotype phase in hybrid zones, we focused on summary statistics that could be obtained from unphased data.

Summary statistics were chosen to reflect basic patterns of ancestry. First, junction number was counted in an individual as the number of switch points between ancestries (Figure 2A) (Fisher, 1954). Second, heterogenicity was computed as the proportion of an individual genome harboring ancestry from both source populations (*i.e.* different ancestries on the two chromosomes; Figure 2B) (Fisher, 1954). Third, we tabulated the frequency of each junction across the sample, thereby generating a "junction frequency spectrum" analogous to the site frequency spectrum commonly used to describe single nucleotide polymorphisms (Braverman et al., 1995; Gutenkunst et al., 2009; Tajima, 1989). Because junctions only form when hybridization has occurred, each junction can be considered derived

relative to the ancestral state of having no junction. Thus, we treated the junction frequency spectrum as unfolded. Because junctions can be inferred even when they are invariant in the sample (and the population), we included fixed junctions in the frequency spectrum.

For junction number and heterogenicity, we computed the mean, median, variance and skew across the 18 sampled individuals. For junction number, we also computed the total count in the sample. We examined the junction frequency spectrum graphically, representing each frequency category as the proportion of junctions found a given number of times in the sample. We further characterized the junction frequency spectrum by computing the number of singleton junctions (those occurring on only one chromosome in the sample), the proportion of singleton junctions, the number of unique junctions (the number of independently occurring junctions, regardless of their frequency), and the skew of the spectrum.

All simulation scripts, input files and results from this study have been deposited on Dryad (https://doi.org/10.5061/dryad.3tx95x6gk).

Results

Summary Statistics

Results for all summary statistics of ancestry we computed are available on Dryad (https://doi.org/10.5061/dryad.3tx95x6gk). Here, we focus on results for mean junction number, mean heterogenicity, and the junction frequency spectrum. These summary statistics were the most sensitive to demography across the parameter space we surveyed.

Ancestry Over Time

The first demographic parameter we examined was time since hybrid zone formation, measured in generations. Because our simulations began with two separate populations, the hybrid zone was established in the initial phase of the simulations. For most parameter combinations (all but the highest migration rate and the smallest deme size), the central population samples are nearly all hybrids by the 500^{th} generation, so we focus on timepoints after 500 generations. Junctions are expected to accumulate over time, approaching an equilibrium that reflects a balance between migration, recombination and drift (Chapman and Thompson, 2002; Hvala et al., 2018), and our findings match that expectation (Figure 3A). For simulations without migration, we found that the mean junction number at equilibrium matches predictions from analytical theory (Janzen et al., 2018).

Heterogenicity decreases over time, also approaching an equilibrium value (Figure 3B). In most cases, it appears that heterogenicity reaches equilibrium before junction number. For example, when the migration rate is 0.001 and the deme size is 3,000, junction number settles between 12,000 and 14,000 generations, whereas heterogenicity barely changes between 6,000 and 8,000 generations.

The shape of the junction frequency spectrum (hereafter denoted as JFS) changes over time. The proportion of singletons decreases and the tail of the distribution lengthens, indicating that

junctions are rising to higher frequencies (Figure 3C-F). The JFS approaches equilibrium near the same time as junction number.

Effects of Migration on Ancestry

The second demographic parameter we considered was migration rate. The migration rate modifies both the time to equilibrium and the equilibrium values for junction number and heterogenicity (Figure 3A,B).

While we expected more migration to decrease equilibrium junction number due to the addition of unadmixed chromosomes (from source populations), we find that the relationship is more complicated (Figure 4A). Changes in the level of migration can increase or decrease the number of junctions. Very high migration reduces junction numbers compared to the case of no migration, whereas very low migration and no migration result in similar junction numbers. However, in between these extremes, there is a zone where increasing migration raises the number of junctions (relative to no migration). Small changes in migration rate can generate substantial effects. For example, samples simulated with an intermediate migration rate of 0.0001 (and a deme size of 3,000) have an average of 57.9 junctions at generation 60,000 and are still adding junctions, whereas samples simulated with the similar migration rate of 0.0004 stop accumulating junctions at an average of 33.5 near generation 34,000.

The effect of migration rate on heterogenicity is more straightforward than the effect on junction number (Figure 4B). Without migration, heterogenicity goes to zero over time as observed for an admixed Wright-Fisher model (Chapman and Thompson, 2002), but with migration heterogenicity decreases to a non-zero equilibrium value. With migration, heterogenicity is lost more slowly than the exponential decay predicted for a population with no migration (Chapman and Thompson, 2002).

Decreasing migration increases junction frequencies, lengthening the tail of the JFS (Figure 4C,D). No migration or very low migration leads to a high proportion of fixed junctions, producing a spectrum that is u-shaped or even right-skewed. The occurrence of fixed junctions is a strong indicator of limited migration. For a deme size of 3,000, no fixed junctions are found in any simulation with a migration rate above 0.0004, and fixed junctions are rare at migration rates between 0.0001 and 0.0004.

When migration rates are very low, all junctions may eventually fix (as in the case of no migration) or an equilibrium featuring a mixture of polymorphic and fixed junctions might be reached. Populations simulated with a migration rate of 0.000 001 (and a deme size of 3,000) appear to achieve an equilibrium in which 68% of junctions are fixed and populations simulated with a migration rate of 0.000 01 (and a deme size of 3,000) appear to achieve an equilibrium in which 19% of junctions are fixed.

Effects of Deme Size on Ancestry

The last demographic parameter we explored was deme size. Increasing deme size generally increases the number of junctions (Figure 5A) as well as the heterogenicity (Figure 5B). Deme size strongly affects the shape of the JFS (Figure 6). Small demes harbor relatively more junctions at higher frequencies (a longer tail), and in some cases, a high proportion of fixed junctions (a u-shaped spectrum).

The effects of deme size depend on migration rate (Figure 5A,B; Figure 7). Increasing deme size enhances the potential for junction accumulation at intermediate and low migration rates (Figure 7A). In other words, junction numbers are less affected by changes in deme size when migration is common. This pattern may be driven by higher heterogenicity in larger demes (Figure 7B).

The Effect of Population Structure

While deme size is a convenient proxy for effective population size, deme number and migration rate also influence effective population size in a stepping-stone model (Barton and Whitlock, 1997; Whitlock and Barton, 1997). To better understand the impact of population structure on ancestry patterns, we conducted simulations under a "hybrid swarm" model, with only one hybrid deme receiving migrants from the source populations. We directly compared simulations with the same total hybrid population size: hybrid swarm simulations had a single deme of 5,000 individuals and stepping-stone simulations had five demes with 1,000 individuals in each deme (Figure 8). While the results follow many of the same trends, it is clear that population structure affects ancestry patterns. For example, increasing migration in a stepping-stone model can raise the junction number more than in the hybrid swarm model (Figure 8C). Furthermore, the stepping-stone model allows junctions to increase in frequency more than the hybrid swarm model, producing a relatively right-skewed junction frequency spectrum (Figure 8E,F). These patterns suggest that results from a hybrid swarm model will be difficult to generalize to populations with the type of structure that characterizes natural hybrid zones.

Discussion

Hybrid zones test the progression of speciation between diverging lineages, potentially leading to reinforcement of reproductive barriers (completion of speciation) or to fusion (reversal of speciation) (Abbott et al., 2013; Coyne and Orr, 2004). The study of hybrid zones provides insights into these alternative outcomes as well as the genetic and evolutionary processes that drive speciation in general (Barton and Hewitt, 1985; Harrison, 1990). Demographic factors are key contributors to both the speciation process and the dynamics of hybrid populations. The intensity and spatial pattern of gene flow between populations depends on migration. The ability of hybrid populations to persist and the effectiveness of selection within them are determined by deme size. As the approach of using genomic patterns observed in natural hybrids to identify the incompatible mutations that isolate nascent species continues to grow in popularity (Gompert et al., 2017; Payseur and Rieseberg, 2016), it is important to consider how demography alters hybrid genomes.

Our simulations reveal that migration and deme size combine to leave detectable footprints in patterns of ancestry in hybrid zones. A novel metric, the junction frequency spectrum (JFS), illustrates how competition between gene flow and drift (controlled by the interactions between migration and deme size) dictates the dynamics of ancestry in the genome. Increasing migration reduces junction accumulation, shortening the tail of the junction frequency spectrum. With little migration, drift within demes fixes junctions. At intermediate migration rates and after enough time, junction frequencies arrive at equilibria reflecting a balance between gene flow and drift. Collectively, these observations indicate strong connections between migration, population size and the shape of the junction frequency spectrum.

Our findings also demonstrate that the relationship between demographic parameters and ancestry can differ depending on parameter values. Although previous work showed that increasing

migration reduces junction number (Hvala et al., 2018) (or equivalently, expands ancestry tract length; Gravel, 2012), we found that higher migration can lead to *more* junctions when migrants are too infrequent to prevent junction accumulation but still frequent enough to contribute to the genetic variation present in the deme. Migrants are likely to carry variants that have drifted to low frequencies in central demes, due to the independent effects of drift across hybrid demes (Barton and Whitlock, 1997). This diversity allows the central deme to maintain a higher level of heterogenicity than it would without migration. Heterogenicity is the substrate for junction formation. This effect is present but much weaker in the simulations under the hybrid swarm model, likely because all migrant chromosomes are unadmixed, narrowing the window between migration increasing heterogenicity and swamping out junctions that have been formed in the hybrids.

The role of demographic history in shaping ancestry patterns was examined in previous theoretical work. Analyzing an isolated hybrid population, Janzen et al. (2018) found that smaller population size, biased starting ratios of ancestries, and non-uniform recombination all slow the formation of junctions by reducing heterogenicity. Modeling an admixture zone over continuous space, Sedghifar et al. (2015) reported that including nearest-neighbor migration impedes the decay of admixture linkage disequilibrium, likely due to the repeated introduction of unadmixed chromosomes from the periphery of the population. By jointly considering gene flow and drift in a stepping-stone model, our study complements Janzen et al. (2018) (which ignored gene flow) and Sedghifar et al. (2015) (which ignored drift). In addition to observing separate effects of gene flow and drift that qualitatively match those in Janzen et al. (2018) and Sedghifar et al. (2015), we demonstrate that these two processes interact to shape ancestry.

Our conclusions are accompanied by caveats and opportunities for extension. First, we expect our assumption of neutrality to be violated in most hybrid zones between divergent lineages, at least for those parts of the genome responsible for reproductive isolation. Selection against hybrids distorts allele

frequency clines (Barton, 1979a; Payseur, 2010) and maintains longer ancestry tracts with fewer junctions than expected under neutrality (Baird et al., 2003; Barton, 1983; Hvala et al., 2018; Sedghifar et al., 2016), suggesting that the effects of demography we documented should be examined in models with selection. The effect of selection is likely to vary across demographic histories. In several of the scenarios examined here, drift is strong and may readily overcome the effects of selection. Strong drift drives genetic patterns in some natural hybrid zones (e.g. McFarlane et al., 2021), and may be particularly relevant in zones where the hybrid populations are small or patchy.

Although the stepping-stone model we studied captures important aspects of hybrid zone structures, actual hybrid zones can take a variety of forms. One example is a mosaic or patchy population structure (Harrison and Rand, 1989). Depending on the connection between mosaic hybrid populations and source populations, these types of hybrid zones could be even more strongly affected by drift, leading to a higher proportion of common junctions over time. There can also be variation in the relative rates of migration from each of the source populations (e.g. Field, Ayre, Whelan, & Young, 2011). In these situations, it is possible that other metrics, such as ancestry proportion, would be stronger indicators of the migration rate. We might expect to see heterogenicity deflated due to a bias towards one parental type, leading to a decrease in junction formation, as seen in an isolated hybrid population (Janzen et al., 2018). Based on our results, considering the population structure of a given hybrid zone will be critical to interpreting its ancestry patterns.

Hybrid zones are dynamic. Our simulations assumed that deme sizes and migration rates are constant over long periods of time, an assumption that is likely to be violated in natural hybrid zones (Barton, 1979b; Buggs, 2007; Wielstra, 2019). The possibility that demographic parameters vary over time should be considered when interpreting ancestry patterns from hybrid zones.

Recombination produces junctions, suggesting that recombination rate shapes the ancestry signatures that demography leaves along chromosomes. We assumed that crossovers appear at a single

rate, independently of one another. Variation in recombination rate along a chromosome (Haenel et al., 2018; Nachman, 2002; Yu et al., 2001) as well as crossover interference (Berchowitz and Copenhaver, 2010)—both widespread phenomena—should further increase heterogeneity in junction patterns conferred by demography.

Despite these caveats, our findings emphasize the potential for using ancestry patterns to reconstruct demographic history in hybrid zones. Existing statistical methods enable the probabilistic inference of fine-scale ancestry switching along chromosomes from genomic data (Baran et al., 2012; Browning and Browning, 2011; Corbett-Detig and Nielsen, 2017; Guan, 2014; Price et al., 2009; Wegmann et al., 2011). Ancestry patterns in admixed populations have often been used to pinpoint the timing of initial gene flow, especially in humans (Corbett-Detig and Nielsen, 2017; Hellenthal et al., 2014; Henn et al., 2012; Liang and Nielsen, 2014; Moorjani et al., 2011; Patterson et al., 2012). In contrast, few analytical frameworks have been developed to characterize demographic history in populations with structures typical of hybrid zones. This gap is surprising, given that gene flow is usually the primary subject of interest when students of speciation examine hybrid zones. In addition, the effective population size of a metapopulation is shaped by both deme size and migration (Maruyama and Kimura, 1980; Whitlock and Barton, 1997). Hybrid zones with smaller demes, less migration, or both are expected to experience more drift. These ideas suggest that the common practice of ignoring population size when drawing evolutionary inferences from hybrid zones in the context of speciation could be misleading.

By identifying summary statistics that are sensitive to migration rate and population size, we have taken a first step toward developing an analytical framework for the reconstruction of demographic history from genomic data in hybrid zones. We view the junction frequency spectrum as an especially informative summary of ancestry. Inference of demographic history could follow two paths. First, simulation results such as ours could be used to guide mathematical theory that connects

junction patterns to demographic parameters, leading to formulae that could be used for parameter estimation. For example, the junction frequency spectrum appears to follow an exponential distribution under a range of conditions. Second, inference could proceed by searching by simulation for parameter combinations that produce similar junction patterns to those observed in hybrid zone data, through Approximate Bayesian Computation or related approaches. To mitigate effects of linked selection on inference, genomic regions with few genes and high recombination rates could be chosen. The reconstruction of demographic history could provide a baseline for detecting selection by scanning genomes from hybrid zones. As genomic datasets from hybrid zones become more readily available, the inference of demographic history will be an important step toward understanding the dynamics of hybrid zones and the process of speciation.

Table 2.1. Tested values of demographic parameters.

Parameters	Values Tested
Deme size	100; 500; 1,000; 3,000; 5,000
Generations of admixture	100; 500; 1,000; 2,000; 4,000; 6,000; 8,000; 10,000; 12,000; 14,000; 16,000; 18,000; 20,000; 22,000; 24,000; 26,000; 28,000; 32,000; 36,000 40,000; 44,000; 48,000; 52,000; 56,000; 60,000
Migration Rate	0; 1e-8; 1e-7; 1e-6; 1e-5; 2e-5; 4e-5; 6e-5; 8e-5; 1e-4; 2e-4; 4e-4; 6e-4; 8e-4; 1e-3; 1e-2
Recombination Rate	0.51
Deme Number	1; 5

Figure 2.1. Stepping-stone model assumed in individual-based simulations. Five hybrid populations (demes) exchange migrants at the same rate. Source populations contribute to the hybrid demes but do not receive migrants.

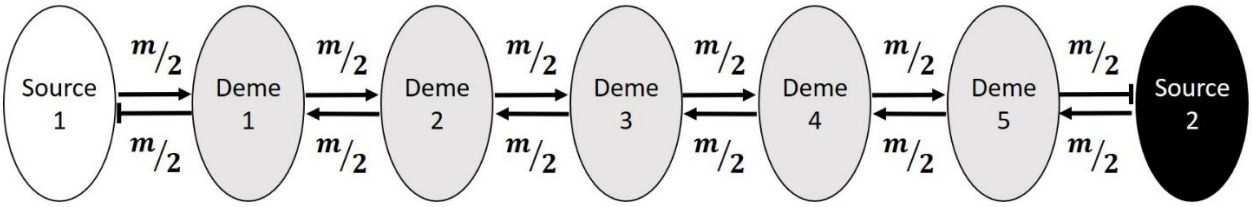


Figure 2.2. Descriptors of ancestry. A) Junctions are transition points between ancestries along chromosomes. B) Heterogenicity is the proportion of the genome that is heterozygous for ancestry.

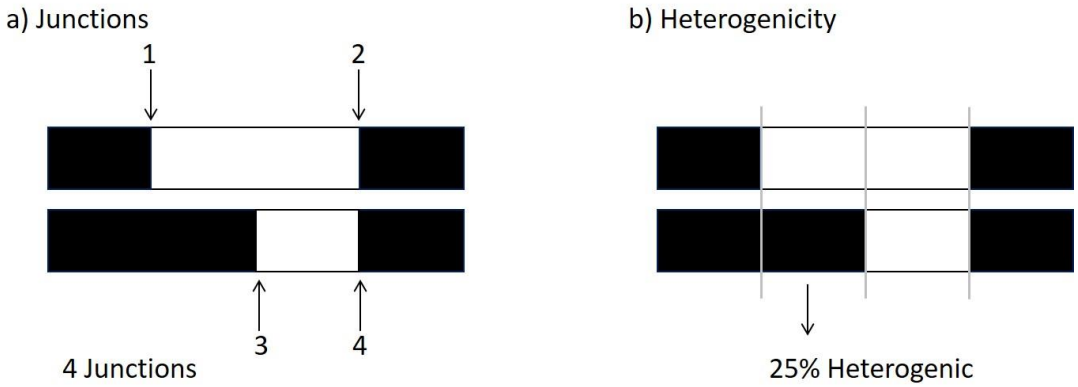


Figure 2.3. Ancestry over time. In simulations with a deme size of 3,000, mean junction number in the sample (A) and mean heterogenicity in the sample (B) are shown as the mean of 100 replicates. Bars represent one standard deviation above or below the mean. Average junction frequency spectra are shown for three migration rates at generation 2,000 (C), 16,000 (D), 30,000 (E), and 60,000 (F). In panel A, the gray line represents the expected number of junctions for the no-migration case based on Janzen et al. (2018).

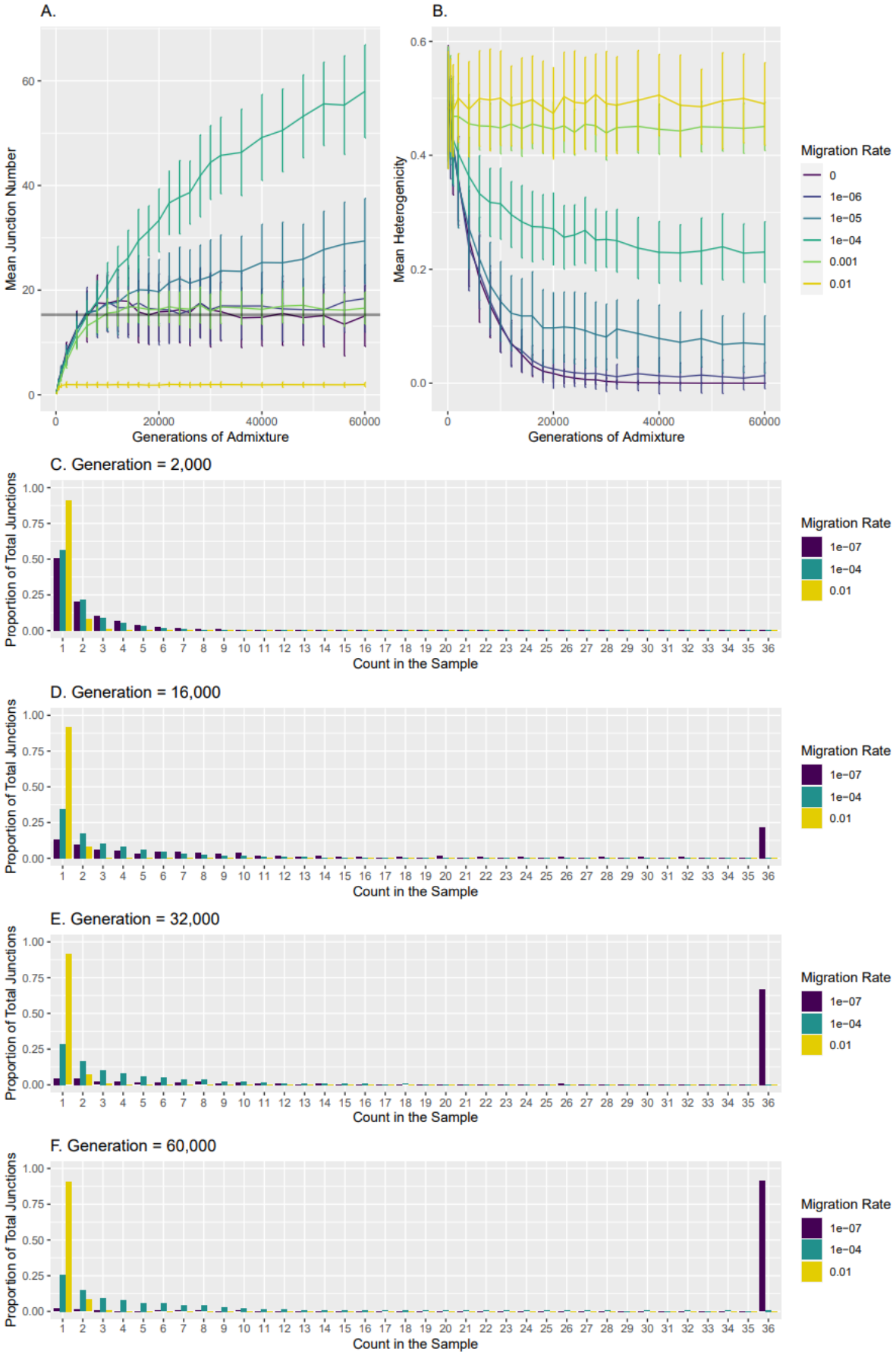

A.
Mean Junction Number
Generations of Admixture
B.
Mean Heterogenicity
Generations of Admixture
Migration Rate
0
1e−06
1e−05
1e−04
0.001
0.01
C. Generation = 2,000
D. Generation = 16,000
E. Generation = 32,000
F. Generation = 60,000
Proportion of Total Junctions
Count in the Sample
Migration Rate
1e−07
1e−04
0.01

Figure 2.4. Impact of migration on ancestry. In simulations with a deme size of 3,000 at generation 60,000, mean junction number in the sample (A) and mean heterogenicity in the sample (B) are shown as the mean of 100 replicates. Bars represent one standard deviation above or below the mean. Average junction frequency spectra are shown for the colored points: a migration rate of 0.000 01 (C) and 0.0006 (D), which have similar mean junction numbers (29.4 and 25.8, respectively).

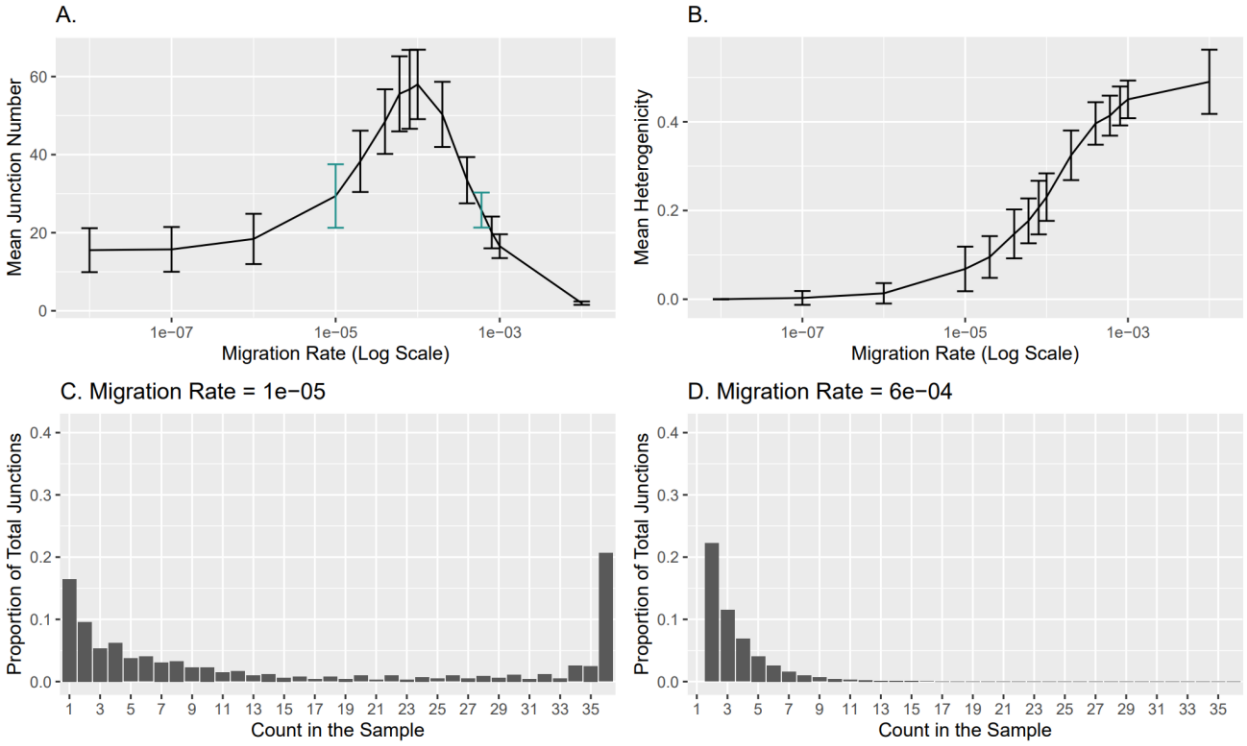


Figure 2.5. Impact of deme size on junction number and heterogenicity. In simulations at generation 30,000, mean junction number in the sample (A) and mean heterogenicity in the sample (B) are shown as the mean of 100 replicates. Bars represent one standard deviation above or below the mean.

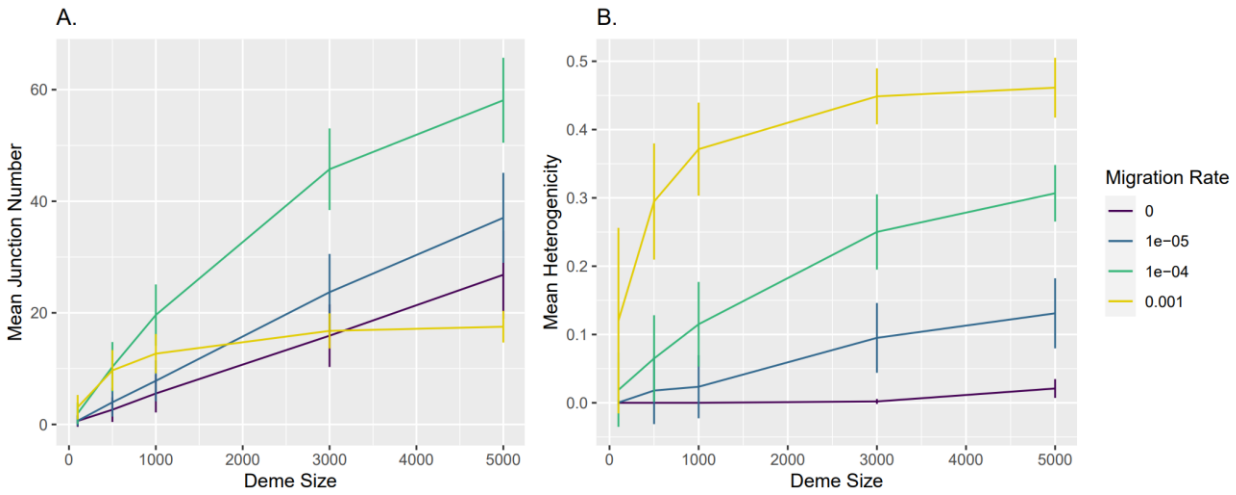


Figure 2.6. Impact of deme size on the junction frequency spectrum. Average junction frequency spectra are shown for a deme size of 500 (A) and 5,000 (B) for two migration rates and two time points.

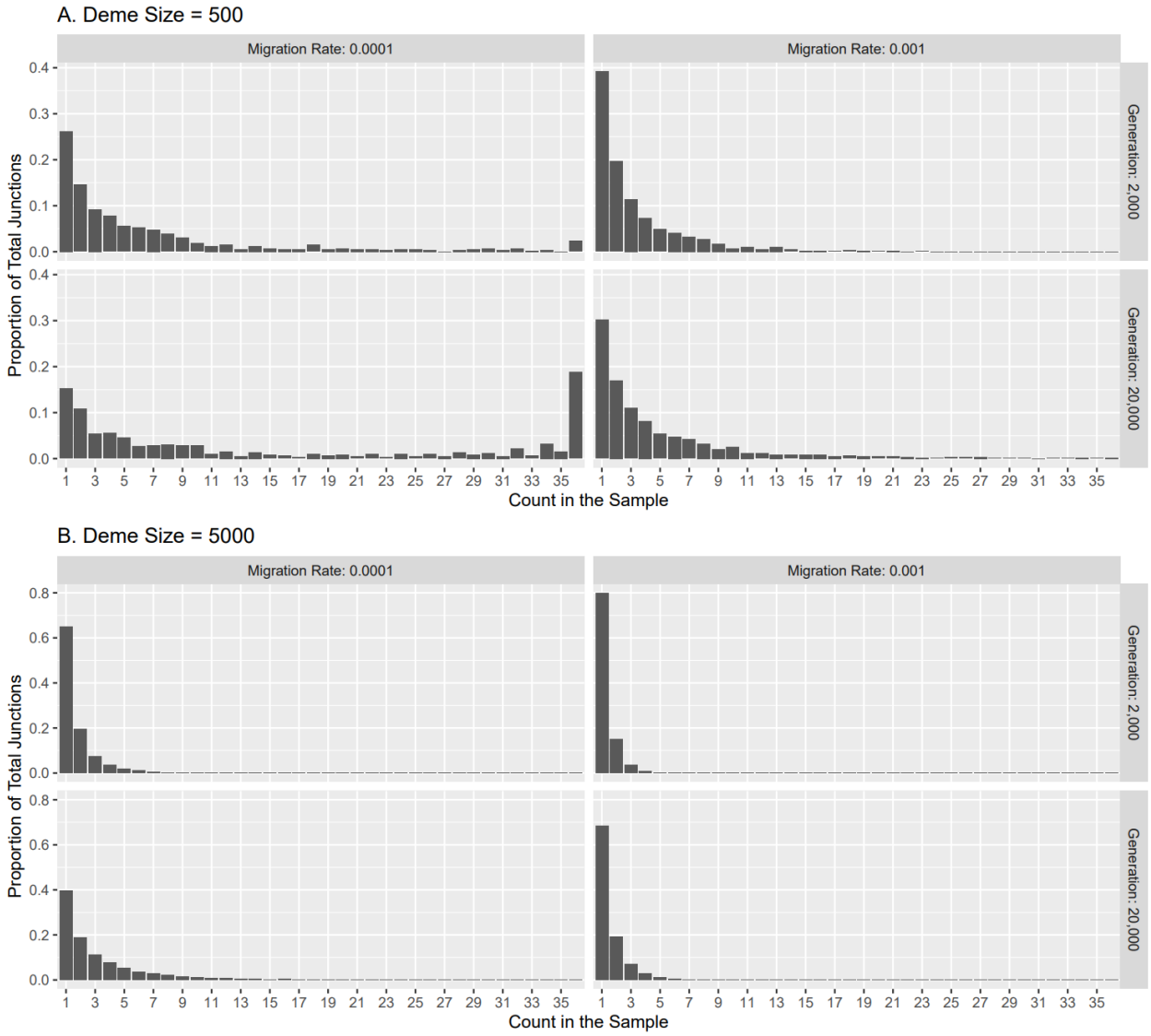


Figure 2.7. Interactions between migration rate and deme size impact ancestry patterns. In simulations at generation 30,000, mean junction number in the sample (A) and mean heterogenicity in the sample (B) are shown as the mean of 100 replicates. Bars represent one standard deviation above or below the mean.

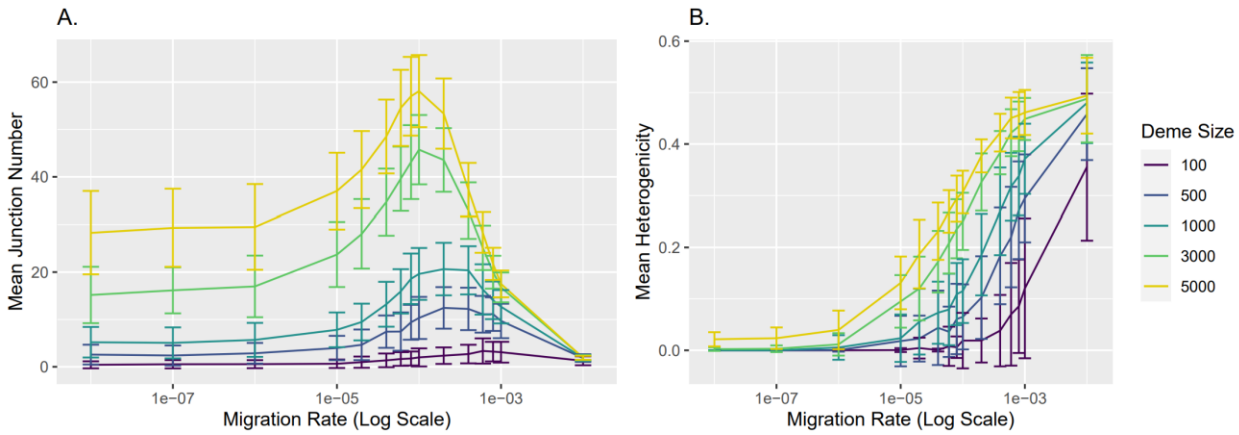
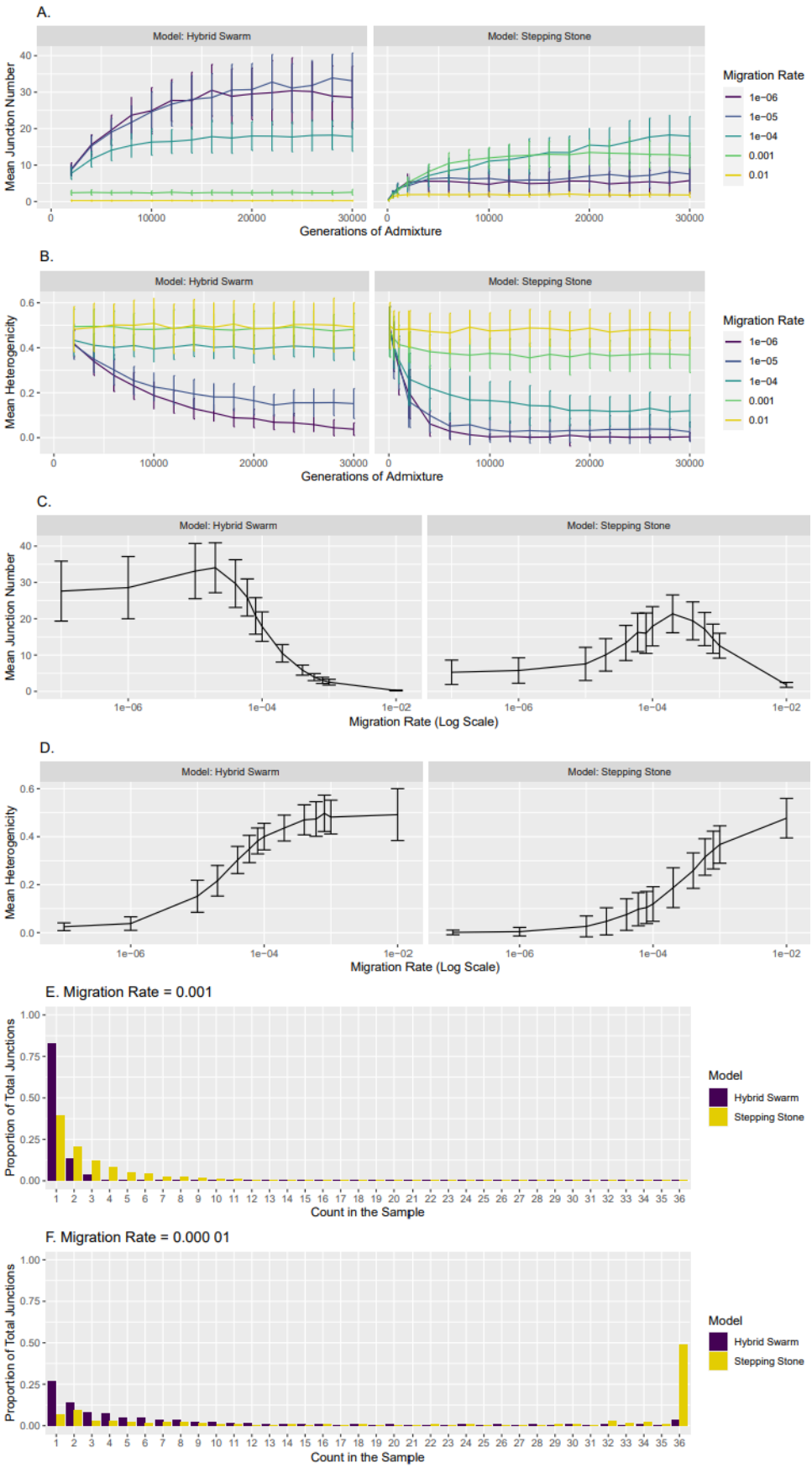


Figure 2.8. Comparison of ancestry patterns from hybrid swarm and stepping-stone models. For simulations at several migration rates with a total hybrid population of 5,000 individuals, mean junction number in the sample (A) and mean heterogenicity in the sample (B) are shown as the mean of 100 replicates. In simulations at generation 30,000 and a total hybrid population of 5,000 individuals, mean junction number in the sample (C) and mean heterogenicity in the sample (D) are shown as the mean of 100 replicates. Bars represent one standard deviation above or below the mean. Average junction frequency spectra are shown for a high migration rate (E) and a low migration rate (F) for both models at generation 30,000 with a total hybrid population of 5,000 individuals.


A.
Model: Hybrid Swarm
Model: Stepping Stone
Mean Junction Number
Generations of Admixture
Migration Rate
1e−06
1e−05
1e−04
0.001
0.01
B.
Mean Heterogenicity
C.
Migration Rate (Log Scale)
D.
E. Migration Rate = 0.001
Proportion of Total Junctions
Count in the Sample
Model
Hybrid Swarm
Stepping Stone
F. Migration Rate = 0.000 01

Acknowledgements

This research was supported by NSF Grant DEB1353737, NIH Grant R01GM120051, and NIH Grant R35GM139412 to BAP. MEF was also supported by NIH Graduate Training Grant in Genetics T32 GM007133. This research was performed using the *compute* resources and assistance of the Center for High Throughput Computing (CHTC) in the Department of Computer Sciences at UW-Madison. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. We thank John Novembre and members of the Payseur lab for advice during this project.

Data Accessibility

The scripts and data from this study have been deposited on Dryad (https://doi.org/10.5061/dryad.3tx95x6gk).

Chapter 3

Inferring demographic history using Approximate Bayesian Computation based on ancestry

Abstract

Understanding the complex history of hybrid zones is a key step in understanding the process of speciation between hybridizing taxa. Ancestry is a powerful indicator of the history of hybrid zones. Here, we try to develop a simulation-based framework to infer the history of a hybrid zone using summaries of ancestry. Cross-validation measurements indicate our framework is unable to reliably estimate parameters of interest, and exploration of partial least-squares components (which are used to correct for correlations between summary statistics) suggests there is little information within the simulation table about the summary statistics. We provide our thoughts on why this method was not successful and suggestions for future directions.

Introduction

Secondary contact zones are complex populations that form where the ranges of two partially isolated taxa meet. When the reproductive barriers between those parental taxa are incomplete, hybrid offspring are produced. Continued interbreeding of hybrids can turn the contact zones into "natural laboratories" where different combinations of the parental taxa are formed and tested (Hewitt, 1988). Under these conditions, hybrid genomes are records of the progression of speciation between the two parental taxa—formed by both the demography of hybridization and selection. Unfortunately, our desire to learn about selection can be hampered by our lack of understanding about the dynamics of the population (Payseur and Rieseberg, 2016). Investigating the demographic history of a hybrid zone can help us to contextualize the genomic patterns we observe in hybrids, and thus improve our understanding of speciation.

Our previous work has shown that ancestry can be a powerful descriptor of demographic history in hybrid zones (Frayer and Payseur, 2021; Hvala et al., 2018). As independent lineages interbreed, the genomes of their hybrid offspring become increasingly complex mixtures of the original lineages, or ancestries. The pattern of ancestry across the hybrid genomes can be inferred, and break points between ancestries can be identified. These breakpoints, or junctions, represent historical recombination events in hybrid individuals, and, once formed, they behave like point mutations carried forward in the hybrid population (Fisher, 1949, 1954). Junctions can accumulate in frequency across a population, and their frequencies can be compiled into a frequency spectrum that is sensitive to demographic history (Frayer and Payseur, 2021). Additional summaries of the ancestry patterns such as the proportion of the genome that is heterozygous for ancestry (heterogenicity) and the proportion of the genome that comes from each parental ancestry are also sensitive to demography.

Because ancestry patterns are sensitive to demography, it should be possible to use them to infer aspects of demographic history. Approximate Bayesian Computation (ABC) is a flexible, simulation-based method for model comparison and parameter estimation (Beaumont, 2010; Sunnåker et al., 2013). While most model-based inference methods depend on complex likelihood calculations, ABC bypasses these calculations by using simulations to estimate the posterior distributions for parameters of interest. For the basic rejection algorithm, simulations are performed under a given model using parameters drawn from a prior distribution for each parameter of interest, and a posterior distribution is generated from the subset of simulations that are accepted based on the distance between the summary statistics for that simulation and the observed data. The estimate from the posterior distribution can be improved by performing regression adjustment (Beaumont et al., 2002). Other algorithms for ABC have also been proposed, such as the MCMC approach (Marjoram et al., 2003).

ABC lends itself well to the study of hybrid zones, which often have complex population structures and histories. Even relatively simple models, such as the stepping stone model (Kimura and Weiss, 1964), can quickly become intractable in a likelihood framework. ABC has been used previously to model complex admixture histories in taxa such as *Mytilus* mussels (Fraïsse et al., 2018; Roux et al., 2016) and humans (Fortes-Lima et al., 2021).

The ABC framework can exploit many types of simulations and summary statistics, making it a natural choice for inference of the demographic history of a hybrid population using summaries of ancestry. Here, we attempt to use the statistics explored in our previous work (Frayer and Payseur, 2021) to develop an inference method using genomic data from a hybrid zone to infer demographic history.

Methods

Simulations

We chose to focus on the stepping stone model (Kimura and Weiss, 1964) because it captures population substructure that exists in many hybrid zones, unlike simpler hybrid swarm or single pulse models (Frayer and Payseur, 2021). We focused on estimating four critical parameters of this model: the number of generations of admixture, migration rate, deme size, and recombination rate. Migration rate was assumed to be constant over time and equal between demes. We chose to infer recombination rate because it would give us more flexibility in selecting regions of various recombination rates from our samples of interest.

ABC was performed using a rejection algorithm. While there have been many advances in alternative approaches to ABC that reduce the computational load (Beaumont et al., 2009; Wegmann et al., 2009), the forward simulations required to generate complex junction patterns are relatively slow. Therefore, we chose to generate a single large pool of simulations that we could draw from to perform many tests of the pipeline. To improve our estimates, we adjusted the posterior distribution using the ABC-GLM algorithm from ABCToolbox (Wegmann et al., 2010), based on the regression adjustment proposed by Beaumont et al. (2002).

We drew 1 million parameter combinations from our prior distributions, and ran simulations using forqs (Kessner and Novembre, 2014) following the approach of Frayer and Payseur (2021). The population was modeled as a stepping-stone, with 5 hybrid demes of finite size with a constant migration rate between them, and two unadmixed source populations that donate but do not receive migrants. For each parameter combination, we simulated 100 independent 1Mb windows with no selection. At the end of the simulation, 18 individuals were sampled from the central deme for calculation of summary statistics. This number was chosen to reflect the available sample size of hybrid

house mouse genomes that we hoped to use to perform demographic inference (see Chapter 4). Because these forward simulations create large datasets, we chose to extract only data about ancestry from the simulations.

Summary Statistics

We calculated several summary statistics with the goal of trying different combinations and choosing the most informative set. We focused on summaries of ancestry that describe diploid data, since it was not possible to get phase from our data set of interest. We included the mean and the variance of junction number, the proportion of heterogenicity, and hybrid index per 1Mb locus. We also included the total number of junctions (all junctions identified in all individuals at all loci) and the number of unique junctions (counting identical junctions occurring on multiple chromosomes only once).

The Junction Frequency Spectrum (JFS) describes junction sharing across the sample. Junctions that occur at the same locus in different individuals are likely to be identical by descent, and counting the occurrences of junctions within the same loci produces a frequency spectrum similar to the site frequency spectrum of polymorphisms (Frayer and Payseur, 2021). Our previous work suggested that the most informative summaries were the values of the frequency bins themselves. We calculated the JFS as both raw counts and proportions.

Testing the method

After the table of simulations was created, cross-validation of parameter estimation was performed using the R package 'abc' (Csilléry et al., 2012). Cross-validations were run with many different combinations of tolerance levels (0.0005, 0.005, 0.01, and 0.005), estimation approaches (standard rejection method, local linear regression), and subsets of summary statistics. The cross-

validation method randomly selects simulations from within the overall table and treats them as observed data to estimate parameters. If the estimation performs well, the cross-validation will show a close relationship between the true and estimated values, and low prediction error rates.

We further explored our simulation data by identifying partial least-squares (PLS) components as outlined in ABCToolbox (Wegmann et al., 2010). PLS components are transformations of calculated summary statistics that can be used in place of those statistics to minimize the effects of correlations between them. While the summaries of ancestry that we chose do have unique responses to demography, they are not entirely independent. This is especially true for the frequency bins of the JFS (the number of junctions that occur 15 times in the population is not likely to be independent from the number of junctions that occur 16 times in the population). We looked at the root mean squared error of prediction (RMSEP) to determine the amount of information about the parameters contained in the PLS components. PLS components were estimated using the R package 'pls' (Hovde Liland et al., 2021) and code provided by ABCToolbox (Wegmann et al., 2010).

Finally, inference of parameters on simulated data was attempted using ABCEstimator and the GLM estimation approach from ABCToolBox (Wegmann et al., 2009, 2010). Data for testing was simulated using *forqs* under the same model as the ABC simulations, and perfect knowledge of ancestry was assumed.

Results

The final table of simulations contained 862,380 parameter combinations (with 100 1Mb loci simulated for 18 individuals for each combination). The distributions of summary statistics in the table were highly skewed (Figure 1), likely as a result of the log uniform distributions used for the priors.

Cross-validation showed a high prediction error regardless of the settings and summary statistics used. An illustrative example of the cross-validations is shown in Figure 2. In this case, all summary statistics were used. The inference was performed with the local linear regression method with log transformations of the log uniform variables for three tolerance rates. Deme size has the lowest prediction error rate (for a tolerance of 0.05, error rate=0.229; for a tolerance of 0.01, error rate=0.204; for a tolerance of 0.005, error rate=0.200). Prediction error rates were very high for the number of generations (for a tolerance of 0.05, error rate=0.538; for a tolerance of 0.01, error rate=0.545; for a tolerance of 0.005, error rate=0.515), recombination rate (for a tolerance of 0.05, error rate=0.759; for a tolerance of 0.01, error rate=0.768; for a tolerance of 0.005, error rate=0.821), and migration rate (for a tolerance of 0.05, error rate=0.852; for a tolerance of 0.01, error rate=0.936; for a tolerance of 0.005, error rate=0.928). There was no consistent pattern in which tolerance level provided the lowest prediction error.

Across all cross-validations, migration rate and recombination rate were particularly difficult to estimate. Across nearly all combinations of statistics and settings tried, migration rate was the most difficult to estimate (prediction error values around 1) and deme size was the most reliable estimate (prediction error values around 0.2). While estimates performed with local linear regression are consistently better than those from the standard rejection method, there is little meaningful improvement from any other setting.

PLS components showed very little power in the dataset to predict these parameters (Figure 3). Adding additional PLS components should decrease the error in estimation, with diminishing returns,

and the smallest number of PLS components that produce the minimum RMSEP should be used for estimation. The minimum RMSEP was greater than 0.9 for both migration rate and recombination rate, and greater than 0.6 for both deme size and number of generations. This indicates that even a large number of PLS components would not produce reliable estimates for these parameters.

Inference of parameters on simulated data was also attempted using simulated data where perfect knowledge of ancestry was assumed. Because this approach minimizes natural sources of error (such as sequencing or ancestry inference errors, or divergence from the demographic model), it should have allowed the best possible chance for accurate estimation. However, estimates for these data were highly variable and rarely reflected the true simulated values. After these analyses, it was determined that inference on real data using this pipeline would not produce meaningful estimates.

Discussion

Overall, the ABC method we devised performed poorly for estimating the parameters of interest. This was surprising given the strong relationship between the parameters and the summary statistics we observed previously (Frayer and Payseur 2021). Below, we discuss several potential explanations.

First, our previous simulations exploring variation in recombination rate were limited. For Chapter 2, recombination rate was fixed at 0.51cM/Mb, the average for house mouse. While the simulations we performed that varied recombination rate showed a strong effect of recombination rate on the summary statistics, it is possible that there was not enough information to jointly estimate recombination rate and the other parameters.

Second, our prior distributions may have been poorly chosen. We used a log uniform distribution for several of the priors, leading to a strong concentration of our simulations in a certain region of the parameter space. The distributions of summary statistics in our simulation table were consequently highly skewed. This likely limited our predictive power. Furthermore, the range of some of the priors, especially migration rate, were very large. This is known to affect power in some cases (Sunnåker et al., 2013).

A recently developed method for ABC inference of admixture histories showed promising results (Fortes-Lima et al., 2021). They also used forward simulations because complex admixture in the coalescent is intractable. This method focused on summaries of the distribution of admixture. While their method performed well in their target populations (recently admixed humans), the authors note that there is a limit to the age of admixture for which these summaries can be useful. It is likely that our target population (hybrid house mice) is too old for methods based on ancestry proportions to be effective. The authors indicate that summary statistics based on admixture-LD may be more informative, but are much more computationally intense, and often require phasing. We concur with these

observations and support the use of junctions (which avoid phasing). Indeed, other methods that have successfully used ABC for inference in a hybrid context have been focused on much more recent admixture (Wang et al., 2019).

Future Directions

We still believe that the ancestry-based summary statistics we evaluated are useful indicators of demography, but we think that their power may be limited to more specific segments of the parameter space. For a future attempt at ABC inference with summary statistics of ancestry, we recommend removing recombination rate as a parameter and using either a fixed recombination rate or an approach with “high”, “medium”, and “low” recombination bins. We also recommend uniform distributions for parameters besides migration rate, as our use of a log-uniform distribution may have concentrated our simulations in too narrow an area of the parameter space. It may also be useful to limit the range of the migration rate prior distribution, although determining a reasonable prior is difficult due to the complexity of connecting migration in these simulations to a real-world measure. Finally, the ancestry-based statistics we explored here could provide complementary information to sequence-based statistics. Future efforts could combine these approaches within one ABC inference.

Table 3.1. Prior Distributions.

Parameter	**Range**	**Distribution Type**
Generations of Admixture	100—30,000	Log Uniform
Migration Rate	1e-8 – 1e-2	Log Uniform
Deme Size	50—5000	Log Uniform
Recombination Rate	0.3—3.0	Uniform

Figure 3.1. Distributions of Summary Statistics. The summary statistics all have highly skewed distributions, which may contribute to our lack of predictive power. Below are histograms of the values across the simulated dataset for A) total junction number, B) mean heterogenicity, c) the first category of the junction frequency spectrum (singleton junctions), D) and the last category of the junction frequency spectrum (fixed junctions).

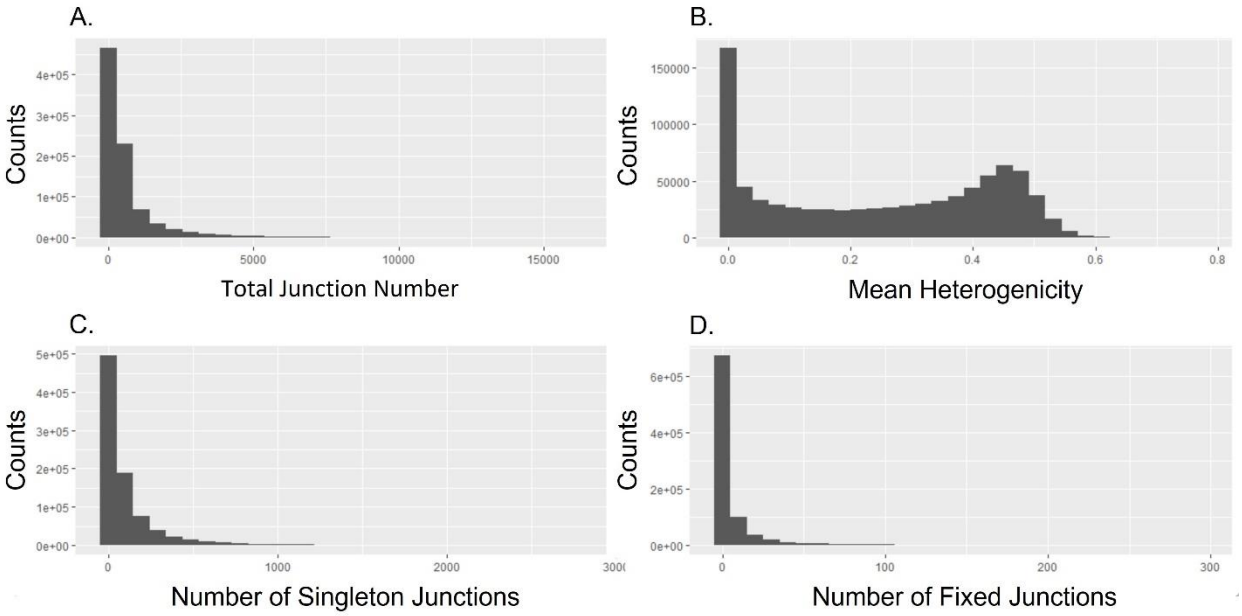


Figure 3.2. Cross-validation results. Below are results for cross-validation on the estimate of the parameters performed using all of the summary statistics and three tolerance rates (0.005 shown in red, 0.01 shown in orange, and 0.05 shown in yellow). These cross-validations were run using the local linear regression method with log transformations of the log-uniform variables.

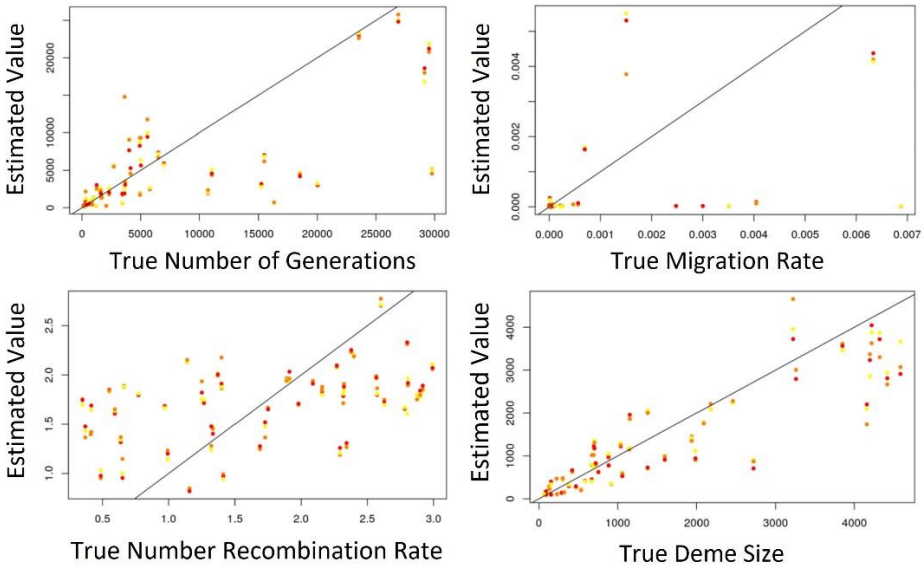
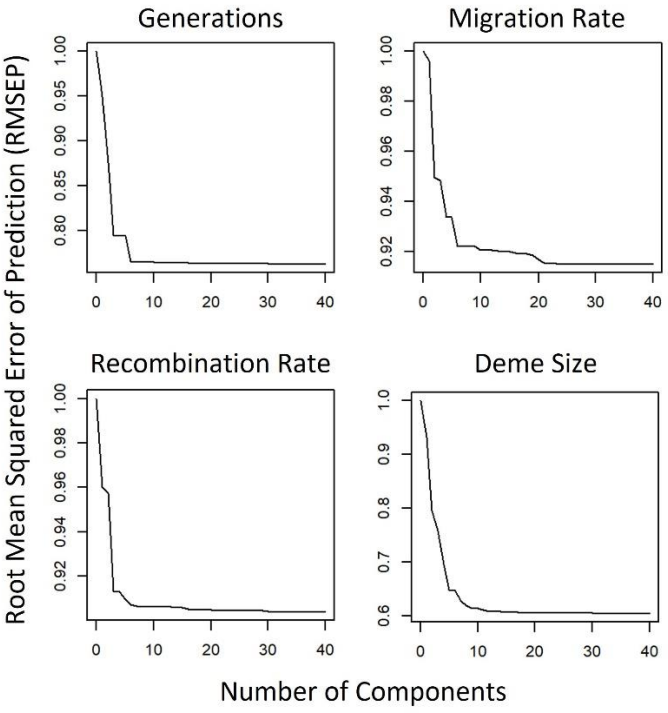


Figure 3.3. Partial least-squares components. The plots below show the root mean squared error of prediction (RMSEP) for 40 PLS components for each parameter.



Chapter 4

Genomic ancestry patterns in a classic hybrid zone

This chapter will be submitted for publication as a peer reviewed article.

Frayer, Megan E., Leslie M. Turner, Bettina Harr, and Bret A. Payseur. “Genomic ancestry patterns in a classic hybrid zone.” *In prep.*

Contributions were as follows: M.E.F. and B.A.P. designed the study. L.M.T. and B.H. provided the mice. M.E.F. conducted analyses. M.E.F. and B.A.P. wrote the paper.

Supplementary information can be found in Appendix A.

Abstract

The history of a hybrid zone is recorded in the ancestry of hybrid genomes. Selection and demography driven by the process of speciation between the parental taxa shape these genomes and the mixing of ancestries within them. Junctions, break points between regions of different ancestry, are the results of historical recombination events that occurred in hybrids. However, their application in hybrid zones has been limited. In this study, we infer ancestry and junctions across 40 whole genomes in mice from two populations from the European house mouse hybrid zone between *Mus musculus domesticus* and *M. m. musculus*. We identify hundreds of thousands of junctions and describe the patterns of sharing across individuals and populations with the first observations of the junction frequency spectrum from natural populations. Despite being just 7.4 km apart, we find that the two populations have unique ancestries and junction patterns, suggesting that they have unique and partially independent histories. The timing of formation of the hybrid populations is inferred to be a few thousand years using both junction density as well as the frequency spectrum of SNPs. Finally, we identify regions of the genome that are outliers for ancestry patterns and potential candidates for incompatibility alleles between the parental subspecies in this hybrid zone.

Introduction

Hybrid genomes offer a unique opportunity to explore species boundaries and the evolutionary process of speciation. They are the product of mating between distinct lineages, and their genomes are shaped by selective and demographic processes driven by the strength of reproductive isolation between the parental taxa. For example, differential selection at loci within hybrid genomes can lead to differences in introgression at those loci (Barton, 1979a) and unique patterns of admixture linkage disequilibrium within hybrid genomes (Schumer and Brandvain, 2016). The strength of reproductive isolation may influence the extent of hybridization and thus the size, number or migration rate of hybrid populations, which will in turn affect the admixture found within hybrid genomes (Gompert et al., 2017). Regions of extensive hybridization (known as hybrid zones) allow these dynamics to play out over time and space, creating a pool of genomes where the impacts of speciation are tested in new and unique ways (Hewitt, 1988).

Local ancestry is a powerful descriptor of admixture history. It has been used to explore ancient admixture in humans (e.g. Sankararaman et al. 2016), the dynamics of historical human movement (e.g. Hellenthal et al. 2014), and to look for loci under selection (e.g. Jeong et al. 2018). It's application to hybrid zones has been more limited (Chiou et al., 2021; Schumer et al., 2018). Local ancestry along a genome can be described by lengths of tracts of continuous ancestry, or by junctions, the break points between those tracts. First explored by Fisher (1949, 1954) in the context of inbreeding, junctions are the direct result of a recombination event in a hybrid individual. Thus, they can be considered like point mutations that are specific to the history of hybridization, rather than originating in either parental population. This may make them particularly well suited to the context of hybrid zones. A desirable feature of junction identity is that it can be determined without phase. Approaches that rely on track lengths require phasing, which can be difficult in hybrid genomes where statistical phasing may rely on

the same patterns generated by admixture and confound our inference of ancestry. This feature may facilitate the opportunity for application of junctions in more taxa.

Junctions are responsive to demography and selection. Junction density increases over time (Baird, 1995; Baird et al., 2003; Liang and Nielsen, 2014; Pool and Nielsen, 2009). An isolated population will rapidly eliminate heterogenicity (heterozygous ancestry) from the population and approach an equilibrium junction number (Chapman and Thompson, 2002). Migration slows the breakdown of tracts (Gravel, 2012; Pool and Nielsen, 2009), and can raise the amount of heterogenicity in an equilibrium population (Frayer and Payseur, 2021), thereby increasing junction density. The size of hybrid populations (Janzen et al., 2018) or the number of hybrid demes (Frayer and Payseur, 2021) can also affect the rate of junction formation and the time required to reach an equilibrium. The frequency of junctions can be counted and compiled into a frequency spectrum, which is also sensitive to demography. More junctions are expected to become fixed in hybrid populations that are small or experience less migration (Frayer and Payseur, 2021). Selection against incompatibilities in a hybrid population can lead to regions of reduced junction density around the selected locus (Hvala et al., 2018) or increased tract length (Sedghifar et al., 2016). The expected pattern depends on the type of selection—underdominance is expected to affect a larger region of the chromosome than epistasis (Hvala et al., 2018), and in some cases small increases in junction density occur adjacent to, but not overlapping with, selected loci (Sedghifar et al., 2016).

In this study, we investigate junctions in two populations of mice from a classic hybrid zone. A hybrid zone between the house mouse subspecies *Mus musculus musculus* (hereafter referred to as *musculus*) and *M. m. domesticus* (hereafter *domesticus*) runs from Norway to the Black Sea (Boursot et al., 1993; Jones et al., 2010). This hybrid zone was likely formed less than 6,000 years ago via secondary contact (Cucchi et al., 2011). The zone is narrow along its width (20-40km; Sage, Whitney, and Wilson

1986; Turner, Schwahn, and Harr 2012), suggesting that it may be a tension zone controlled by a balance between migration and selection against hybrids (Barton and Hewitt, 1985).

Many types of reproductive isolation have been described between *musculus* and *domesticus*. Mice collected from the hybrid zone show reduced fertility (Albrechtová et al., 2012; Turner et al., 2012). Laboratory crosses between *musculus* and *domesticus* have revealed many fertility defects, including abnormalities in sperm shape, motility, and count (Good et al., 2008; Schwahn et al., 2018; White et al., 2011). Hybrids show dysfunction in meiosis driven by an incompatibility involving *Prdm9* (Forejt et al., 2021), as well as aberrant expression patterns (Mack et al., 2016). Despite all of these factors maintaining isolation, and evidence of reduced fitness in hybrids, the mice in the hybrid zone are highly recombinant. No F_1 hybrid has ever been found in the hybrid zone, suggesting contact between the parental taxa no longer occurs.

House mice make ideal candidates for the investigation of ancestry junctions. The large body of hybrid zone work provides context for interpreting junction patterns. Additionally, mice are a genetic model organism—one of the oldest and best characterized models for biomedical research (Phifer-Rixey and Nachman, 2015). This allows us to leverage a high quality reference genome (Church et al., 2009), detailed recombination maps (Cox et al., 2009), and extensive genome annotation resources (Blake et al., 2011). High quality recombination data is especially essential to the investigation of junctions as junctions are a subset of recombination events.

In this study, we infer junctions in two hybrid populations from a transect across the hybrid zone in Bavaria that was defined by Sage et al. (1986). We map ancestry and investigate the density of these junctions across the genome. We gain insights into the complex demographic history of the zone by investigating the sharing of junctions within and between hybrid populations. We use junction density to infer the age of the hybrid populations, and the frequency spectrum of SNPs to infer population size and

migration rate. Finally, we identify candidate incompatibility regions based on unusual patterns of ancestry, which may be contributing to the process of speciation between *musculus* and *domesticus*.

Methods

Sample Collection

Hybrid mice were collected from several populations along the east-west transect of the hybrid zone in Bavaria (Turner et al., 2012). This transect was first defined by Sage et al. (1986b). For this study, two populations were chosen because they had relatively large sample sizes and were located in the central part of the transect: one collected from Neufahrn (hereafter FS), and one collected from Hohenbachern (hereafter HO). Based on a small panel of fixed differences, FS individuals were expected to have a hybrid index around 0.58 and HO individuals were expected to have a hybrid index around 0.68 (Turner et al., 2012).

Genome Sequencing and Processing

Eighteen mice from FS and 22 mice from HO were selected for whole genome resequencing. For 30 of these mice, we used DNA extractions from Turner, Schwahn, and Harr (2012). For the remaining 10 mice, we extracted genomic DNA from livers using the Puregene Kit (Qiagen, Hilden, Germany). DNA quality for all samples was confirmed on a 1% agarose gel. Sequencing library preparation was completed using TruSeq Nano WGS library prep (Illumina Inc., San Diego, CA, USA) in the University of Wisconsin – Madison Biotechnology Center. Sequencing was completed at the University of Minnesota Genomics Center. Libraries were checked for concentration (Quant-iT PicoGreen; Invitrogen, Waltham, MA, USA) and fragment size (Agilent Bioanalyzer; Agilent Technologies, Inc., Santa Clara, CA, USA). Two pools of 20 sequences each were created and validated by Kapa qPCR (KAPA Biosystems, Wilmington, MA, USA). Libraries were sequenced to ~44X individual coverage across several lanes of two Illumina NovaSeq S4 2x150-bp flow cells, with over 2 billion reads per lane. Images were processed using the standard Illumina pipeline by the University of Minnesota Informatics Institute. All libraries had a mean

quality score above Q30. Quality of reads was checked using fastqc (v0.11.7). Additionally, genome sequences for 32 parental individuals (8 each from 2 populations of *musculus* and 2 populations of *domesticus* from outside the hybrid zone) were downloaded as fastqs (Harr et al., 2016) and re-processed alongside the hybrid genomes (Figure 1).

Reads were mapped to the *Mus musculus* reference genome GRCm38.p6 (downloaded 6/26/2018) using bwa mem (v.0.7.12). Alignments were sorted and indexed using samtools (v. 1.7). Variants were called using the GATK Haplotype caller (v.3.7.0) with java (v. 1.8.0). We first called variants individually (GVCF) and then jointly across all mice (hybrid and parental populations). Given the bias towards *domesticus* in the reference sequence, this joint calling step was expected to increase support for alternative alleles common in the *musculus* population at the risk of under-calling singletons. We felt that this was an acceptable risk because our primary objective was to examine ancestry, for which singletons are uninformative. We filtered SNPs using the GATK best practices guidelines for hard filtering. We did not attempt to filter using VSQR because available datasets for mice are biased towards *domesticus*.

SNP positions in centiMorgans were estimated from the recombination map generated from heterogeneous stock mice (Cox et al., 2009) and position conversion to the GRCm38 reference genome provided by Karl Broman (https://github.com/kbroman/CoxMapV3). CentiMorgan positions of SNPs located between markers on the genetic map were interpolated assuming a linear increase in cM position with Mb position. Ends of the chromosomes beyond the recombination map were removed. For chromosome 14, a larger portion at the start of the chromosome was removed due to a conflict in site positions between reference genome versions.

Ancestry Inference

Local ancestry was inferred using Ancestry HMM (Corbett-Detig and Nielsen, 2017). Our variant set was reduced to biallelic SNPs with no missing data in mice from the populations of *domesticus* and *musculus* from outside the hybrid zone. The structure of our reference populations (8 individuals each from two populations per subspecies) could introduce bias due to small sample sizes and within subspecies population structure. To mitigate these biases, ancestry was called using all four combinations of individual *musculus* and *domesticus* source populations (ALxMC, ALxTP, etc.), and then a union set of sites with the same inferred ancestry across all four inferences was selected. This approach reduced spurious ancestry switching caused by incomplete information from the source population. For the X chromosome, males and females were analyzed separately, and males were treated as haploids.

Sites were filtered to those for which the likelihood of the most likely ancestry exceeded 0.95, and junctions were inferred from those calls. Junction location could be anywhere between the two adjacent SNPs that have different ancestries; we chose to assign junctions to the location of the first SNP with a new ancestry (the telomeric-most probable location for the junction). We calculated the hybrid index as the proportion of the genome with *domesticus* ancestry. We calculated the heterogenicity as the proportion of sites that are heterozygous for ancestry.

Finally, we examined the frequency spectrum of junctions. We defined junctions as shared when the ancestry changed in the same direction (i.e. *musculus* to *domesticus*) at the same site. Because our data is unphased, we considered the locations of junctions in a diploid space. However, we could still assign junction identity in a haploid manner. A location where diploid ancestry changes from *musculus* to heterozygous requires that there is a junction on one of the two chromosomes where ancestry changes from *musculus* to *domesticus*. Likewise, a location where diploid ancestry changes from *musculus* to *domesticus* requires two junctions, one on each chromosome. These “double junctions”

themselves can be markers of demography that behave similarly to heterogenicity (Frayer and Payseur, 2021). Locations where ancestry changes from *musculus* to *domesticus* on one chromosome and *domesticus* to *musculus* at the same location on the opposite chromosome are not visible in diploid data, but previous work suggests these locations are relatively rare (Frayer and Payseur, 2021). While assigning the identity of a junction is simple, assigning the location of a junction is difficult. Our approach of assigning junctions to the telomeric-most location will miss some shared junctions that are not assigned to the same site due to differential ancestry information, but it will also count some junctions that are not actually shared. However, we favored the simplest approach in the absence of a good model for assigning these locations. We looked at the impact on junction sharing when junctions are assigned to the centromeric-most position for chromosome 19, and found the differences were minimal and seemingly random (Supplemental Figure 1).

To examine variation in ancestry along chromosomes, the autosomes were separately split into windows based on physical (1Mb) or genetic (1cM) distance. These window sizes were chosen for a few reasons. First, our estimates of recombination rate are most accurate at larger scales. Second, a larger scale allows for more variation in junction pattern among windows. To reduce errors in ancestry calls, we removed windows with unusually low SNP density (fewer than 300 SNPs) or unusually high density (greater than 75,000 SNPs), as well as windows with low differentiation between the source populations.

We searched for predicted correlations between ancestry patterns and genomic features. We estimated recombination rates for each window using the cM positions described above. We also looked at gene density by calculating the proportion of sites within each window that overlapped with genes in the NCBI RefSeq track from the UCSC Genome Browser database (“ncbiRefSeq” downloaded as bed file). To mitigate the effects of uncertainty in the estimation of recombination rate, we also

compared ancestry metrics in the windows with the highest 10% and lowest 10% of recombination rates along the autosomes.

As a point of comparison, we also inferred ancestry using a heuristic approach based on fixed differences. We combined the two populations from each subspecies, extracted fixed differences between them, and assigned ancestry in the hybrids.

Demography and Population Structure

We investigated the structure of our population samples using SMARTPCA (v6.1.4), filtering SNPs to remove missing data and SNPs in linkage disequilibrium with plink (v2). We investigated relatedness and identity-by-descent sharing using IBIS (Seidman et al., 2020). Other methods for inferring relatedness that ignore SNP position may artificially inflate relatedness estimates for house mice (Harr et al., 2016).

We used the site frequency spectrum (SFS) of SNPs to estimate basic parameters of demography using the simulation-based maximum likelihood procedure implemented in fastsimcoal2 (v2.709). We estimated the site frequency spectrum from 1,062 5kb windows across the autosomes that were predicted to be intergenic (at least 100kb away from genes) and independent (at least 25kb away from each other). We separately estimated the folded SFS for the FS population (from 606,351 SNPs) and for the HO population (from 303,352 SNPs). In subsequent analyses, we ignored singletons because this adjustment improved the fit of our model. We fixed parameters that have reliable estimates in the literature, including the effective population sizes of *musculus* and *domesticus* (153,000 and 212,000, respectively; Phifer-Rixey et al., 2020) and mutation rate (6e-9; Milholland et al., 2017). The two hybrid populations were analyzed separately using the same general model, which was used as a starting point to compare the fit of several variations on the model: allowing two migration rates (a separate rate from each source population), allowing variable divergence time, allowing variable initial hybridization

proportions, and combinations of these variations. Models were compared using the Akaike Information Criterion (AIC). The best estimates (highest log likelihood) from the best model (lowest AIC) were used to generate parametric bootstrapping confidence intervals following the fastsimcoal2 manual.

We used the junctions package (v2.0.3) in R (v4.0.2) to separately estimate the age of the two hybrid populations using both our inferred number of junctions and our ancestry calls (Janzen and Miró Pina, 2022; Janzen et al., 2018). For both types of estimations, we obtained separate estimates for each chromosome from each individual; we present the distribution of these estimates. For estimates based on the junctions we inferred, we used the command `estimate_time()`. For estimation based on our ancestry calls, we used the command `estimate_time_diploid()` and provided a version of our AncestryHMM ancestry calls with no missing or ambiguous calls in any individual.

Outliers

Several types of selection are expected to leave a signature on the junction patterns in individuals from a hybrid zone (Hvala et al., 2018). Generally, selection against hybrid incompatibilities will reduce junction numbers around the loci under selection, with some types of selection (e.g. single-locus underdominant selection) leading to greater reductions than others (e.g. epistatic selection against a DMI). Ideally, we would pinpoint these signatures in hybrid mice by identifying genomic regions with ancestry metrics that depart from those expected given their demographic history. Along these lines, we attempted to use approximate Bayesian computation to reconstruct demographic history from ancestry metrics for these samples, but our approach failed to generate reliable estimates for the demographic parameters of interest (Chapter 3). For this reason, we adopted an outlier approach to identify regions that show ancestry patterns in the tails of genome-wide distributions. The distribution of ancestry statistics within the genome is in part shaped by selection, and the tails of the distribution are expected

to be enriched for loci under selection, although we cannot estimate the expected rate of false positives in the absence of a demographic model that fits ancestry for the majority of the genome.

Looking in the 1Mb and 1cM windows as described above, we identified an initial set of potential outlier regions by collecting the highest 1% and lowest 1% of windows for each ancestry summary statistic: hybrid index (the proportion of sites with *domesticus* ancestry), heterogenicity (the proportion of sites heterozygous for ancestry), mean junctions (the number of junctions averaged over individuals), total junctions (all junctions identified in the window, counted separately even if they are identical junctions), unique junctions (the number of non-identical junctions identified in the window), singleton junctions (the number of junctions that appear on only one chromosome), and fixed junctions (the number of junctions that are found on every chromosome, which was only measured for the HO population). In addition, for hybrid index and heterogenicity, a normal distribution was fit to the data and windows further than 2.5 standard deviations from the mean were also selected as outliers. These outlier analyses were conducted separately using 1Mb and 1cM windows.

These windows were then surveyed for content and overlap. We compared windows within and between populations, and looked for correlations between statistics, recombination rate, and genic site density (defined as the proportion of sites that overlap with at least one known gene). We then intersected the windows to find those that were identified across multiple metrics and merged to join sequential windows into larger windows. The largest clusters of windows were selected as regions of interest.

Results

Sequences

Over 400 million reads were generated per sequence, for an average coverage of 44X (Table 1). After processing and filtering, we had over 47 million SNPs. We checked for relatedness among mice sampled from the same locality using IBIS (Seidman et al., 2020) to infer identity-by-descent (IBD) windows. While some IBD blocks were identified between pairs of individuals, there was no relatedness identified above the 4th degree. There was more IBD sharing among HO individuals than among FS individuals.

One individual from the FS population had unusually high sequence heterozygosity, a high number of indels, and a low number of singletons. We redid joint SNP calling and conducted all subsequent analyses without this mouse.

Basic Patterns of Ancestry

Ancestry inference revealed a complex and varied landscape of junctions across the genome. We identified 882,395 total junctions across the autosomes of both populations. On average, there are 10.9 junctions/Mb among FS individuals and 9.7 junctions/Mb among HO individuals. This was a surprising number of junctions, but we think it is conservative. The alternative approach of calling ancestry based on fixed differences infers more junctions than the union-call model-based approach (17.4 junctions/Mb in FS and 15.2 junctions/Mb in HO), likely as a result of spurious switches due to incomplete source population information. The vast majority of “extra” blocks were defined by only one SNP, which would be expected if certain fixed differences in our sample are not truly fixed in the source populations of the hybrid zone.

In both populations, ancestry differences are greater between chromosomes than between individuals (Figure 3). While junction number is correlated with length of the chromosome in centiMorgans (as expected because junctions are a subset of recombination events), this contrast cannot be fully explained by differences in recombination rate since this pattern persists when averaging over genetic length. While junction density, heterogenicity, and hybrid index vary across chromosomes, the junction frequency spectrum is similar in shape across chromosomes. This could suggest that drift or selection are swamping out some effects of migration in these populations, thereby allowing pieces of the genome to drift towards a fixed pattern of ancestry (as expected in a closed population; Chapman and Thompson 2002).

While individual variation is smaller than chromosomal variation, there is still quite a bit of variation between individuals. In the FS population, individual hybrid indices range from 0.444 to 0.654 and heterogenicity ranges from 0.307 to 0.478. Hybrid individuals contain unique ancestry information and may result from distinct hybridization histories. The relationship between ancestry and hybridization history is also supported by variation in junction sharing among individuals. The amount of pairwise junction sharing between individuals correlates with the sharing of IBD blocks. There is unique information to be gained from the individual genomes that would be lost if they were pooled, or if smaller sample sizes were used as representatives, particularly in the FS population.

The most consistent pattern across all analyses is the striking difference between the two hybrid populations. Despite being only 7.41km apart, there are differences in all basic ancestry metrics between the two populations (Figure 2). HO has fewer junctions per individual (individual average of 21,471 versus 24,120; Wilcoxon rank-sum test $W=374$, $p<0.0001$), fewer unique junctions in the population (56,561 versus 105,128; $W=20$, $p<0.0001$), lower heterogenicity (individual average of 0.19 vs 0.37; $W=374$, $p<0.0001$), a higher hybrid index (individual average of 0.79 vs 0.57; $W=0$, $p<0.0001$), and increased junction sharing relative to the FS population (indicated by a right-skewed JFS; Figure 2C).

About 15% of unique junctions are shared between populations (Figure 2D), which indicates some common history between the populations after the onset of hybridization. However, the proportion of junctions that are unique to each population differ (79.5% of FS junctions and 62% of HO junctions). Each of these metrics suggests that the HO population is more isolated than the FS population, with more migrants moving from HO to FS than the reverse.

Ancestry at a finer scale

To look at finer scale patterns, the genome was broken into 1Mb windows and 1cM windows. These window sizes are small enough to observe deviations from the global means, but large enough to contain multiple junctions. After filtering windows with extreme SNP densities or low differentiation, we were left with 2,250 1Mb windows and 1,312 1cM windows. The distributions of ancestry metrics for both populations in 1Mb and 1cM windows are shown in Figure 4. For hybrid index and heterogenicity, distributions in the HO population are much more skewed, in line with our observations at a broader scale. While the junction statistics are more similar between the two populations, the means of each population are significantly different for all statistics. In 1cM windows as at the broader scale, the hybrid index is greater in HO than FS (Wilcoxon rank-sum test $W=375935$, $p<0.0001$), heterogenicity is lower in HO than FS ($W=1349495$, $p<0.0001$), mean junction number is lower ($W=948300$,$p<0.0001$), total junction number is higher ($W=776043$, $p<0.001$), unique junctions is lower ($W=1261658$,$p<0.0001$) and singleton junctions is lower ($W=1360953$, $p<0.0001$).

We find that junction number correlates with recombination rate at this scale as it does on the chromosomal scale. In the FS population, the Spearman's rho for recombination rate vs mean junction number is 0.402 ($p<0.001$). There is a significant correlation using unique junctions as well, and for both statistics in the HO population. While we do find the expected positive correlation between recombination rate and the proportion of genic sties in a window ($rho=0.210$; $p<0.001$), we do not find a

correlation between genic site density and junction density (for mean junction number in the FS population; rho=-0.010; p-value=0.62). In both populations, a linear regression treating mean junction number as the dependent variable and including both recombination rate and genic site density as independent variables indicates that recombination rate has a much stronger effect, but genic site density does have small negative effect (F_{FS}=202.9, p<0.001; F_{HO}=133.1, p<0.001). This pattern is also present and significant if unique junction number is used. Junctions within genes could lead to chimeric proteins, so selection against such junctions could be expected (Baird, 2006).

We do not find a significant correlation between recombination rate and minor parent ancestry at this scale in either population. There is also no difference between the distributions of minor parent ancestry for the windows in the top and bottom 10% of recombination rates. A positive correlation between recombination rate and minor parent ancestry is expected when selection against incompatibilities is present (Barton and Bengtsson, 1986), so it is surprising that we do not find it here.

Demography of the Hybrid Populations

We used the model-free approach smartPCA (Patterson et al., 2006; Price et al., 2006) to check for population structure. We found that the hybrid populations cluster distinctly from each other (Supplemental Figure 2). FS clusters intermediate to the parents, while HO clusters closer to *domesticus*, as expected based on the hybrid index.

We estimated migration rate, timing of hybridization, and hybrid population size from the site frequency spectrum of SNPs assuming a simple admixture model and using fastsimcoal2 (Figure 5A). After comparing several variations of our model, we find that the most-supported model differs between the two populations. For the FS population, the best model adds divergence time as a variable, and for the HO population, variation in both divergence time and admixture proportion are best. Although the likelihood was significantly improved by this model, estimates across all models were

roughly similar. Because our observed site frequency spectra are very uneven, none of the simulated site frequency spectra were a particularly close match (Figure 5B). However, we proceeded with estimating the parameters under the best model for each population (Table 2). The estimates between the two populations are distinct, although this is reasonable given the differences in observed frequency spectra. The confidence intervals for many of the estimates are large, and some do not contain the estimates. In these cases, the estimates are in the extreme tails of the confidence intervals generated by the parametric bootstrapping procedure. This reflects the difficulty in estimating the confidence intervals—the imperfect fit of the model has caused inflation of the confidence intervals. The population size estimate for HO (27,265; 95% confidence interval 7,566-20,830) is much larger than that for FS (1,705; 95% confidence interval 1,873-23,722), but the confidence interval for HO is completely contained in that for FS. The estimate for migration rate into the FS population (1.8e-3; 95% confidence interval 4e-7-2.6e-3) is roughly twice that for the HO population (9.7e-4; 95% confidence interval 4e-7-3.1e-3), but again, the confidence intervals are very large. The estimated timing for the onset of hybridization is smaller for FS (673; 95% confidence interval 727-7,910) than HO (2,710; 95% confidence interval 709-2,107), but the confidence intervals suggest the age is better estimated for HO than FS. In both cases, fsc2 estimates a divergence time between *musculus* and *domesticus* on the order of 1.3 million generations. This is much earlier than the most recent estimates in the literature (124,000- Phifer-Rixey, et al., 2020; 226,000- Fujiwara et al., 2022) and even the older estimates (628,000- Geraldes et al., 2008). Interestingly, fsc2 estimates an initial admixture proportion much closer to 50/50 for the HO population (0.498, 95% confidence interval 0.436-0.558). This is surprising because, for other estimates, we had assumed a starting proportion equal to the current hybrid index (20% *musculus* and 80% *domesticus*).

Using the `estimate_time()` function in the R package `junctions`, we estimated the age of the hybrid zone based on the inferred number of junctions separately for each chromosome and each

individual. We find a wide distribution of estimates (Figure 6A), with an overall mean of 3,481 generations for the FS population and 3,095 generations for the HO population. Using the data-based `estimate_time_diploid()` function leads to slightly lower estimates (Figure 6B). A few individual chromosomes have estimates that are extreme outliers (not shown in Figure 6). The FS population has a mean of 1,471 and a median of 1,400; the HO population has a mean of 2,443 and a median of 2,225. Following our observations in ancestry metrics, there is far less variation in estimates of hybrid zone age between individuals than between chromosomes.

Outlier Scan for Selection on Autosomes

To identify candidate regions that may be under selection against incompatibilities, we identified the top and bottom 1% of windows for each population for each ancestry metric. We chose windows based on ranking—if there were many windows with the same value for a given statistic, none of the windows were considered outliers. In the HO population, only one tail had identifiable windows for most statistics due to the extreme skew of the distributions (the lower tail for hybrid index, the upper tail for the rest of the metrics). We identified approximately 23 windows for each distribution tail (Table 3). Some of these windows were identified through multiple approaches or in multiple populations—in total, we found 372 unique 1cM or 1Mb windows across the autosomes.

Outlier windows are distributed across the genome, but not evenly across chromosomes. Across all statistics and populations, chromosome 10 had the most unique windows identified as outliers (41) and chromosome 17 had the least (8). Patterns we observed genome-wide also appear in these windows, such as higher recombination rate in 1 Mb windows with higher junction densities.

An overlap between outlier windows in FS and HO could be an indicator of shared evolutionary pressures driving those windows into the tails of the distribution. In general, there is overlap in outlier windows between the populations. Outlier windows in the FS population have a non-random rank in the

HO population (Supplemental Figure 3). There is a significant difference in HO population rank between non-outliers and upper tail 1cM windows from the FS population for hybrid index (Wilcoxon rank-sum test W= 11,311, $p<0.05$) and mean (W= 16,120, $p<0.0001$), unique (W= 14,244, $p<0.0001$), total (W=16,128, $p<0.0001$) and singleton junction number (W=12,577, $p<0.0001$). There is a significant difference for lower tail windows for mean (W=639, $p<0.0001$), total (W=624, $p<0.0001$) and unique junction number (W=587, $p<0.0001$). No singleton junction lower tail was identified for the FS population. While heterogenicity is not significantly different, it does trend in the expected direction. An alternative cause of the overlap between populations could be gene flow. If gene flow were driving this pattern, we might expect to see a difference in the proportion of junctions that are shared between the two populations in outlier vs non-outlier windows. We do find a significant difference in the proportion of shared junctions for both tails of the hybrid index distribution (only one of which was significant for HO rank; Upper tail W=8522, $p<0.05$; Lower tail W=1537, $p<0.01$), but none of the other statistics. Finally, some of the overlap may be driven by shared genome organization (Langdon et al., 2022). While our use of 1cM windows should rule out the effects of recombination rate, other aspects of genome organization could play a role.

Most windows are only considered outliers based on one statistic (206/372 unique windows). However, there were many windows that were considered outliers based on multiple statistics (Figure 7). For example, there is a 1cM window on chromosome 7 and another on chromosome 3 that were identified as high junction density windows by all four junction statistics in both populations. Both windows are relatively large on a physical scale. One interpretation could be selection for junctions in these windows. Alternatively, it is possible that these windows are regions where recombination rate is poorly described. Hybrid mice may experience aberrant patterns of recombination due to rapidly evolving recombination rates within and between *M. musculus* subspecies (Dumont et al., 2011; Peterson and Payseur, 2021). It is possible that the high density of junctions indicates that these hybrids

experience more recombination in this region than the heterogenous stock mice. These regions may be of interest for further study in this regard.

There are several places in the genome where windows cluster—adjacent or overlapping windows are outliers for one or more statistics. We separately aggregated physical and genetic windows to look for the largest clusters. We looked in detail at clusters over 6Mb in length, 4 in the FS population and 1 in the HO population (Table 4). The largest cluster was found on chromosome 10 in the FS population, where there is an 11.9Mb cluster of 1cM windows identified by multiple approaches (Supplemental Figure 4). This region is mainly characterized by low heterogenicity and low hybrid index (indicating an excess of *musculus* ancestry), identified through both the 1% outlier approach and the 2.5 SD approach. Part of this window is also in the lower tail for mean, unique, and total junctions. The first half of this region has a fairly low rate of recombination, making the excess of *musculus* ancestry in this region particularly interesting.

The largest window in the HO population was actually a single 1cM window found on chromosome 5, and this cluster is also identified in the FS population (Figure 8). This 6.3Mb region is in the upper tail of mean, unique and total junctions in the FS population, and the upper tail of mean and total junctions in the HO population. This window is quite variable in the FS population, but not in the HO population. It also overlaps with a 1Mb window in the HO population that is in the lower tail of hybrid index, with an excess of *musculus* ancestry. It is notable that this region is close to the centromere, typically a location of lower recombination in mice. While it is not clear what might cause selection directly for higher junction numbers, an increase in junction number around certain incompatibilities has been predicted in the theoretical literature (Hvala et al., 2018; Sedghifar et al., 2016). This could also apply to the cluster on chromosome 13, which is characterized by a high number of singleton junctions in the FS population (Supplemental Figure 5).

The largest cluster identified among the physical windows is a 6Mb region on chromosome 14 in the FS population. This region shows high heterogenicity in both the 1% outlier approach and the 2.5SD approach (Supplemental Figure 6). An overlapping 1cM window is in the upper tail for singleton junctions. The majority of this region is considered heterozygous.

X Chromosome

The X chromosome plays a large role in speciation (Coyne and Orr, 2004), due in part to its suite of unique attributes that also impact its ancestry patterns.

The X is hemizygous in males. We inferred ancestry on male and female X chromosomes separately, treating the males as haploids. In both populations, females have more junctions than males: (FS female average = 307, FS male average = 83, W=72, $p < 0.0001$; HO female average = 204, HO male average = 74, W=120, $p < 0.0001$). This is expected because they have two X chromosomes, but there is also more variability in the number of junctions between females. Additionally, FS females have closer to 3 times the number of junctions as FS males, where the ratio is closer to double in the HO population. This could be an artifact of the way junctions were called— treating ancestry as haploid or diploid may affect the probability assigned to ancestry calls, leading to an inflation in junction calling in females or a deflation of junction calling in males. It could also be a result of our small sample size (40 female X chromosomes vs 19 male X chromosomes). Females and males have similar hybrid indices. Males could not have heterogenicity in our analyses, and we did not attempt a detailed inference of ancestry in the pseudo-autosomal region.

To compare the X to the autosomes, we focused on comparisons within females. We compared the X chromosomes to chromosome 3, which has a similar physical length (169Mb and 160Mb respectively). The X chromosomes have far fewer junctions than the female chromosome 3 (Figure 8A; Wilcoxon rank-sum test W= 400, p-value < 0.0001). The X tends to have very low heterogenicity in the

HO population. In fact, there are two individuals in the HO population who appear to have almost no heterogenicity on the X. These individuals do not appear to be males based on the coverage of their sex chromosomes, so we infer that they have two very similar copies of the X. The X chromosome showed reduced junction number compared to all of the autosomes. However, this reduction may be completely explained by the lower recombination rate of the X. At a chromosomal scale, female X chromosomes fall within the range of the other chromosomes in terms of junctions/cM (Figure 8B).

The FS population individuals have a large amount of *musculus* ancestry on the X chromosome, whereas the HO population have largely *domesticus* X chromosomes. This is especially interesting given the role the *musculus* X plays in hybrid sterility among F_1 hybrids in the lab. Elements of the *musculus* X and a heterozygous *Prdm9* locus often cause hybrid sterility (Forejt et al., 2021; Lustyk et al., 2019). While it is difficult to infer *Prdm9* alleles from short-read sequencing data, our ancestry data suggests that 5 out of 17 individuals from the FS population have heterogenic ancestry at *Prdm9* and homozygous *musculus* ancestry at *Hstx2* (the interacting region of the *musculus* X, defined by Lustyk et al., 2019).

Further demographic modeling is needed to fully understand the expectations for ancestry on the X chromosome, so we did not directly compare the X to the autosomes in terms of outliers.

Discussion

Throughout the 70 years of research in the European house mouse hybrid zone, the hybrids found have been highly recombinant. Our ancestry maps highlight this highly recombinant nature. In this study, we have presented an overview of genomic ancestry in two hybrid house mouse populations from this classic hybrid zone. We have shown that ancestry patterns in these genomes are very complex and have identified thousands of ancestry junctions. We have presented the first estimation of junction frequency spectra from a natural hybrid zone. We presented estimates of several demographic parameters and identified patterns that are consistent with a strong impact of demographic history on these hybrid genomes. Finally, we have presented candidate loci that are consistent with expectations for selection.

Age of the Hybrid Zone

Our results indicate that the hybrids in the center of the hybrid zone are the result of at least a few thousand generations of hybridization. This is in line with earlier estimates based on archeological data that place a ceiling of 6000 years on the formation of the hybrid zone (Cucchi et al., 2005). Mice at the north-south center of the zone are expected to be of intermediate age, with younger mice to the north and older mice to the south. This places our estimate nicely with those from farther north (250 generations; Hunt and Selander, 1973; Sage et al., 1993).

We generated three estimates of age for these populations. For the FS population, the estimates range from 673 generations to 3,481 generations. For the HO population, the estimates range from 2,225 to 3,095 generations. The two hybrid populations have different estimates in all of the analyses. The estimate is higher in the HO population from fastsimcoal2 and the estimate_time_diploid() approach. From the estimate_time() approach, it is lower in the HO population, although the differences are less extreme. It is possible that founding of these populations occurred at very different times within

the overall establishment of the hybrid zone. If we think these populations should be similar in age, the difference could be explained by the way our data appears to violate the assumptions of the models. HO has a lower junction density, so if the two populations experienced no migration after initial admixture, we would infer that the HO population is younger. However, two populations of the same age that experience different levels of migration could reach different junction densities (Frayer and Payseur 2021). Intermediate amounts of migration can lead to junction densities well above the density expected in a closed population, and high migration can lead to junction densities well below the expectation. The `estimate_time_diploid()` method within the junctions package follows the same model as the `estimate_time()` method, but infers the junctions based on pairwise ancestry. Because the HO experiences an increase of junction sharing between the chromosomes of the same individual, and greatly reduced heterogenicity, it could appear older than the more varied FS population. Furthermore, because we needed to remove sites with missing data in any individual, the `estimate_time_diploid()` approach utilized less data than was used to infer junctions. The fsc2 results are difficult to interpret due to the difficulty in inferring confidence intervals. While the best estimate for the FS population was smaller, the confidence interval is actually larger.

All three of the models assumed a single hybrid population. We know from previous work that demic or stepping-stone structure will affect the accumulation of junctions differently than a hybrid swarm or single population model (Frayer and Payseur, 2021; Hvala et al., 2018). In the case of fastsimcoal2, adding additional demes increases the number of variables and makes the fitting of the model more difficult, so we chose not to attempt it. It would be ideal to reconstruct demography with a more complex population model, although the exact structure of house mouse populations is poorly understood. Despite relatively large effective population sizes (and presumably large census population sizes), mice may have a demic population structure with local extinction and colonization events (Berry

and Bronson, 1992; Berry et al., 1982; Dallas et al., 1995; Wang et al., 2011), which would have a strong effect on the efficiency of selection in their populations.

Hybrid zone dynamics

All of our analyses indicate that the FS and HO populations have different histories, despite both being from the central hybrid zone and less than 8km apart. There are differences in all ancestry patterns at multiple scales (whole genome, chromosomes, 1Mb and 1cM windows). HO exhibits reduced variation in sequence and ancestry patterns among individuals. There are clear differences in the frequency spectra of SNPs and junctions between the populations. This has two important implications. First, it highlights the importance of sampling multiple hybrid populations. Variation between hybrid populations has been previous demonstrated in this (e.g. Janoušek et al., 2012) and other hybrid zones (e.g. Mandeville et al., 2017). Particularly in house mice, which have a huge geographic range compared to their dispersal ability, variation among hybrid populations is an important component of speciation between parental taxa. The outcome of speciation could be changed by geographic variation in selection against hybrids (Cutter, 2012; Gompert et al., 2017). Second, it highlights the existence of population structure within the house mouse hybrid zone beyond the clinal structure. Understanding this population structure is essential for interpreting population genetic data that we collect from this zone. Attempts to investigate hybrid population size have been particularly rare in this zone.

In this study, we used fastsimcoal2 to estimate hybrid population size and migration rate. The hybrid population size inferred for the two populations was very different. The population size inferred for the HO population was much larger, despite our other evidence pointing towards a smaller, more insulated history for this population. It does, however, predict a lower migration rate into the zone, which is in line with our other data. Unfortunately, these fastsimcoal2 estimates have very large confidence intervals, making them somewhat difficult to interpret. The difficulty in fitting this model in

fsc2 might come from a mismatch between our simple population structure model and the real history of the mice. The SFS generated from these data are highly unusual and there are not many models that can sufficiently capture their shape.

There is not a lot of information on the sizes of hybrid populations within the zone, but there have been many estimates of gene flow. Gene flow has been estimated in the context of cline estimation (e.g. Teeter et al., 2008; Gompert and Buerkle 2011; Janousek et al., 2015), although it is difficult to directly compare these estimations at specific loci to the general migration rate we estimated here. Some studies have suggested that gene flow is asymmetrical (Dallas et al., 1995). While there is evidence for asymmetrical gene flow at specific loci in other studies (e.g. Ďureje et al., 2012; Janoušek et al., 2012), allowing two migration rates did not improve the fit of the model in our estimation. Dispersal rates of hybrid mice are estimated to be on the order of 0.5-1km per generation (Macholán et al., 2007; Raufaste et al., 2005), but dispersal rates are not necessarily equivalent to migration rates.

Evolutionary Forces Acting in the Hybrid Zone

Our analyses support a strong role for demographic history in shaping these hybrid populations. In addition to the large differences between the two hybrid populations discussed above, we found no reduction in junctions on the X beyond those driven by recombination. Such a reduction is expected if ancestry patterns are driven by selection against incompatibilities, which are predicted to be enriched on the X chromosome (Coyne and Orr, 1989; Masly and Presgraves, 2007). The X is known to have reduced introgression in this transect, but that reduction varies between loci (Payseur et al., 2004; Teeter et al., 2008, 2010). Junction patterns do not show evidence that such selection extends broadly across the chromosome.

A positive correlation between minor parent ancestry and recombination rate has been observed in other hybrid populations (Sankararaman et al., 2016; Schumer et al., 2018). This pattern is

expected if there is strong selection against incompatibilities—more minor parent ancestry will escape the effects of selection on linked deleterious sites if recombination is working more rapidly to break that linkage (Barton and Bengtsson, 1986). We did not find this correlation in either of our populations. While it may not be appropriate to look for such a correlation in the FS population, which is quite close to the 0.5 hybrid index, the HO population has similar admixture proportions to those of populations where such correlations were found (Schumer et al., 2018). If there is selection against minor parent ancestry in these hybrid populations, it may be weak, or at least weak enough to be swamped out by migration. Negative correlations between these parameters have been predicted under conditions of unequal genetic load (Kim et al., 2018), as well as positive selection or incompatibilities with asymmetric dominance (Duranton and Pool, 2022). It could be that the populations in our study are experiencing conflicting forces that counteract each other and eliminate the expected correlation. This provides further support that variation in ancestry patterns across the genomes may be driven by both demography and selection.

We should reconcile this difference with the findings of Janousek et al. (2015), who found a positive correlation between introgression and recombination in mice from this same transect of the hybrid zone. There are several differences between our studies which could explain the discrepancy. First, the sample sizes are very different. We have a small sample of individual genomes (17-22 full genomes per population), whereas Janousek et al. used a large sample of a small number of markers (1,316 markers for 432 individual mice). Importantly, these markers were selected from a panel of fixed differences, which could bias the resulting estimates of introgression (Wang et al., 2011). Second, the measures of introgression were different. We measured minor parent ancestry as a simple proportion, whereas Janousek *et al.* used β, the measure of introgression from the Bayesian Genomic Cline model (Gompert and Buerkle, 2011). While they found a correlation in the Bavarian transect that we have studied here, they did not find a correlation in every transect, suggesting that there is some variability.

Third, we measured this correlation in single populations, as opposed to introgression across the cline. It is possible that the effects of selection against minor parent ancestry could be swamped out by demographic factors at the level of individual populations.

While selection is often thought of as the major driver in hybrid populations, there are many reasons why demography should play a strong role. Hybrid zones occur at range edges, where reduced population density is common (Buggs, 2007). Selection itself is also expected to reduce hybrid population densities when hybrids experience reduced fertility and/or inviability. If hybrid populations are small, this increases the likelihood that drift is strong. Drift will reduce the efficiency of selection (Kimura et al., 1963) and can lead to steeper clines, even at neutral loci (Polechová and Barton, 2011). Drift may play a role in variation in other hybrid zones (Mandeville et al., 2017; McFarlane et al., 2021) and is likely to be involved in polymorphism among incompatibility loci (Cutter, 2012).

Unusual Ancestry Patterns

Our outlier approach identified many interesting regions. Since we were not able to take a model-based approach, it is likely that some of these windows are not under selection, and that some loci under selection for ancestry were missed. In the HO population, the skewed distribution of junction densities means that we could not identify any windows that were outliers for low junctions, which is the most common expected junction pattern for an incompatibility locus under selection (Hvala et al., 2018). However, we were still able to identify such outliers for the FS population, and it is still likely that the windows we identified are enriched for loci under selection.

A major caveat of our ancestry inference is our assumptions about the rate of recombination. It is essential to understand the recombination landscape in order to understand junctions because they are a subset of recombination events. This might make the junction approach unsuitable for some systems, but there is also a need for greater understanding the recombination landscape in hybrid zones

(Payseur and Rieseberg, 2016). While mice have a relatively well understood recombination landscape, that landscape can evolve rapidly in terms of both numbers of crossovers (Dumont and Payseur, 2011; Peterson and Payseur, 2021) and hotspot positions (Smagulova et al., 2016). Furthermore, hybrids themselves may have unusual patterns of recombination due to the differences between parental populations. A misspecification of recombination rate would change our interpretations of any of these junction metric outliers. In principle, one can use junctions to make inferences about hybrid recombination (Wegmann et al., 2011), although we did not attempt it here due to the age of our hybrid population and our sample size. While this caveat is important to keep in mind, the fact that junction metrics strongly correlate with recombination rate in our dataset suggests that our assumptions about recombination rate were reasonable.

Hybrid populations between incipient species with partial reproductive isolation will experience selection due to that reproductive isolation. Identifying regions of the genome that harbor such loci is an important goal in understanding the genomics of speciation because these loci may directly influence the outcome of speciation when hybridization occurs (Campbell et al., 2018; Ravinet et al., 2017). Barrier loci between the parental taxa may show exceptional patterns relative to the rest of the genome in hybrids. We have shown that ancestry patterns are variable across the hybrid genomes we sampled. Selection is expected to impact patterns of ancestry by creating biases in ancestry proportions or reducing the density of junctions (Hvala et al., 2018). We have noted several regions of the genome that show such patterns. We have chosen to focus our attention on the largest clusters of outlier windows, but in principle any of the windows we identified could potentially be linked to barrier loci.

The Content of Loci of Interest

Ancestry at the locus on chromosome 10 is consistent with selection against hybrids. First, it shows reduced junction density, a signature of selection (Hvala et al., 2018). Second, it shows biased

ancestry and reduced heterogenicity, which one might expect if an incompatible locus from another ancestry has been removed. Third, it overlaps with QTL for reduced hybrid fertility phenotypes such as reduced testes area (Schwahn et al., 2018), and reduced testes weight and sperm proximal tail bend (White et al., 2011). Finally, this region contains several genes that are expressed in the testes or have known roles in spermatogenesis. The genes *Shc2* and *Chst11* were identified by Morgan et al. (2020) as significantly differentially expressed in fertile vs sub-fertile hybrid mice. Several spermatogenesis genes in this region are predicted to be under positive selection in *Mus* (Dorus et al., 2010).

Under most conditions, regions of high junction density are not expected to harbor incompatibilities. However, Hvala et al. (2018) found that in the case of a tightly-linked, dominant-dominant DMI, a peak of junction density may occur due to selection for recombination between the DMI loci. Furthermore, certain types of selection generate regions of high junction density bordering the junction density trough around the incompatibility (e.g. positive selection; Hvala et al., 2018). Sedghifar et al. (2016) also noted regions of reduced block length around selected loci. In our data, the strong skew of the junction density distribution makes it easier to identify increases in junction density than reductions, which may explain why many of our focal regions trend in that direction.

Several outlier regions show such increases in junction density: the clusters on chromosome 5 and 13, as well as the windows with the most “hits” (found to be an outlier across the most statistics) found on chromosomes 3 and 7. The locus at chromosome 5 does not overlap with any known QTL, but it does contain many genes, a few of which were identified by Loire et al. as candidates for involvement in assortative mating between *musculus* and *domesticus*. The locus at chromosome 13 overlaps a testis weight QTL identified by Dzur-Gejdosova et al. (2012). The windows with the most “hits” (on chromosomes 3 and 7) also have high junction densities. The window on chromosome 7 overlaps with a QTL influencing sperm head shape in F_1 hybrids (White et al., 2011).

The locus at chromosome 14 is interesting because it is mainly characterized by high heterogenicity. This is a necessary prerequisite for increasing junction density (Chapman and Thompson, 2002), so it might result from similar selection pressures that lead to high junction density. Junction density is high across this window, although it is not in the upper tail of the distribution. The locus at chromosome 14 overlaps with a novel interaction hotspot identified by Turner and Harr (2014). We believe that any of these regions could be good candidates for further investigation.

Summary and Future Prospects

Our study represents a step forward in the practical application of a junction framework in hybrid zones. While there are still many difficulties in junction identification to be overcome, we have shown that junctions and junction sharing can be investigated in hybrid populations. Junctions can act as a complementary framework for ongoing studies that focus on correlations in ancestry. This approach may reveal new aspects of hybridization history that could contribute to our understanding of the process of speciation between hybridizing incipient species.

Table 4.1. Summary of genome sequences.

Population	Location	Subspecies	Number of Individuals	Average Coverage	Source
MC	France	*M. m. domesticus*	8	23.1x	Harr et al. 2016
TP	Germany	*M. m. domesticus*	8	22.2x	Harr et al. 2016
FS	Hybrid Zone	Hybrids	17	46.4x	This study
HO	Hybrid Zone	Hybrids	22	42.9x	This study
CR	Czech Republic	*M. m. musculus*	8	24.4x	Harr et al. 2016
AL	Kazakhstan	*M. m. musculus*	8	24.9x	Harr et al. 2016

Table 4.2. Parameter estimates from best fastsimcoal2 model.

Population		FS	HO
Best Model		Variable divergence time	Variable divergence time + variable initial admixture proportions
Parameter Estimates (95% confidence intervals)	Hybrid population size (number of chromosomes)	1,705 (1,873-23,722)	27,265 (7,566-20,830)
	Migration rate from source populations	0.0018 (4e-7-2.6e-3)	0.00097 (4e-7-3.1e-3)
	Time of hybrid population foundation	673 (727-7,910)	2,710 (710-2,108)
	Divergence time of *musculus* and *domesticus*	1,502,947 (888,147-2,100446)	1,341,120 (1,271,835-1,464,759)
	Initial admixture proportion of hybrid population	-	0.498 (0.436-0.558)
Likelihoods	Maximum likelihood of observed SFS	-29120605	-372710.8
	Maximum likelihood of estimated SFS	-29172200	-378943.6

Table 4.3. Number of outlier windows identified using each approach.

Pop	Approach	Window Type	Tail	Fixed Junctions	Heterogenicity	Hybrid Index	Mean Junctions	Singleton Junctions	Total Junctions	Unique Junctions
FS	1% Outlier	Genetic	Lower	0	12	12	9	0	13	11
FS	1% Outlier	Genetic	Upper	0	13	13	13	13	13	13
FS	1% Outlier	Physical	Lower	0	22	22	24	0	22	21
FS	1% Outlier	Physical	Upper	0	23	23	23	24	23	23
FS	2.5 SD	Genetic	Lower	0	8	7	0	0	0	0
FS	2.5 SD	Genetic	Upper	0	4	0	0	0	0	0
FS	2.5 SD	Physical	Lower	0	0	13	0	0	0	0
FS	2.5 SD	Physical	Upper	0	14	0	0	0	0	0
HO	1% Outlier	Genetic	Lower	0	0	12	0	0	0	0
HO	1% Outlier	Genetic	Upper	11	13	23	13	13	13	14
HO	1% Outlier	Physical	Lower	0	0	22	0	0	0	0
HO	1% Outlier	Physical	Upper	23	23	0	23	22	23	22

Table 4.4. Largest outlier window clusters.

Window	Length (bp)	Pop[1]	Scale[2]	TJN[3]	UJN[4]	HI[5]	Het[6]	Outlier Statistics in this region
chr14:81000000-87000000	6000000	FS	Mb	854	247	0.613	0.698	Upper tail of heterogenicity; 2.5 standard deviations above mean heterogenicity
chr13:85070079-91239482	6169403	FS	cM	717	229	0.874	0.219	Upper tail of hybrid index; Upper tail of singleton junctions
chr5:5395560-11694302	6298742	FS	cM	1182	295	0.219	0.388	Upper tail of mean junctions; Upper tail of total junctions; Upper tail of unique junctions
chr5: 5395560-11694302	6298742	HO	cM	1125	106	0.512	0.084	Upper tail of mean junctions; Upper tail of total junctions
chr10:76443695-88351768	11908073	FS	cM	1171	146	0.041	0.025	Lower tail of hybrid index; Lower tail of heterogenicity; Lower tail of mean junctions; Lower tail of total junctions; Lower tail of unique junctions; 2.5 standard deviations below mean hybrid index; 2.5 standard deviations below mean heterogenicity

[1]Population in which the cluster was identified
[2]Scale of the windows in which the cluster was identified (physical vs. genetic)
[3]Total number of junctions in the window
[4]Number of unique junctions in the window
[5]Hybrid index across the window
[6]Heterogenicity across the window

Figure 4.1. Sample Locations. This map shows the approximate location of the hybrid zone (red line) and the locations of the 6 populations considered in this study. MC and TP are the domesticus source populations. CR and AL are the musculus source populations. FS/HO are the two hybrid populations.

Figure 4.2. Ancestry patterns compared between FS and HO for A) hybrid index versus heterogenicity, B) mean junction number per individual and unique junction number for each chromosome, C) the single-population junction frequency spectrum, and D) the joint junction frequency spectrum between populations.

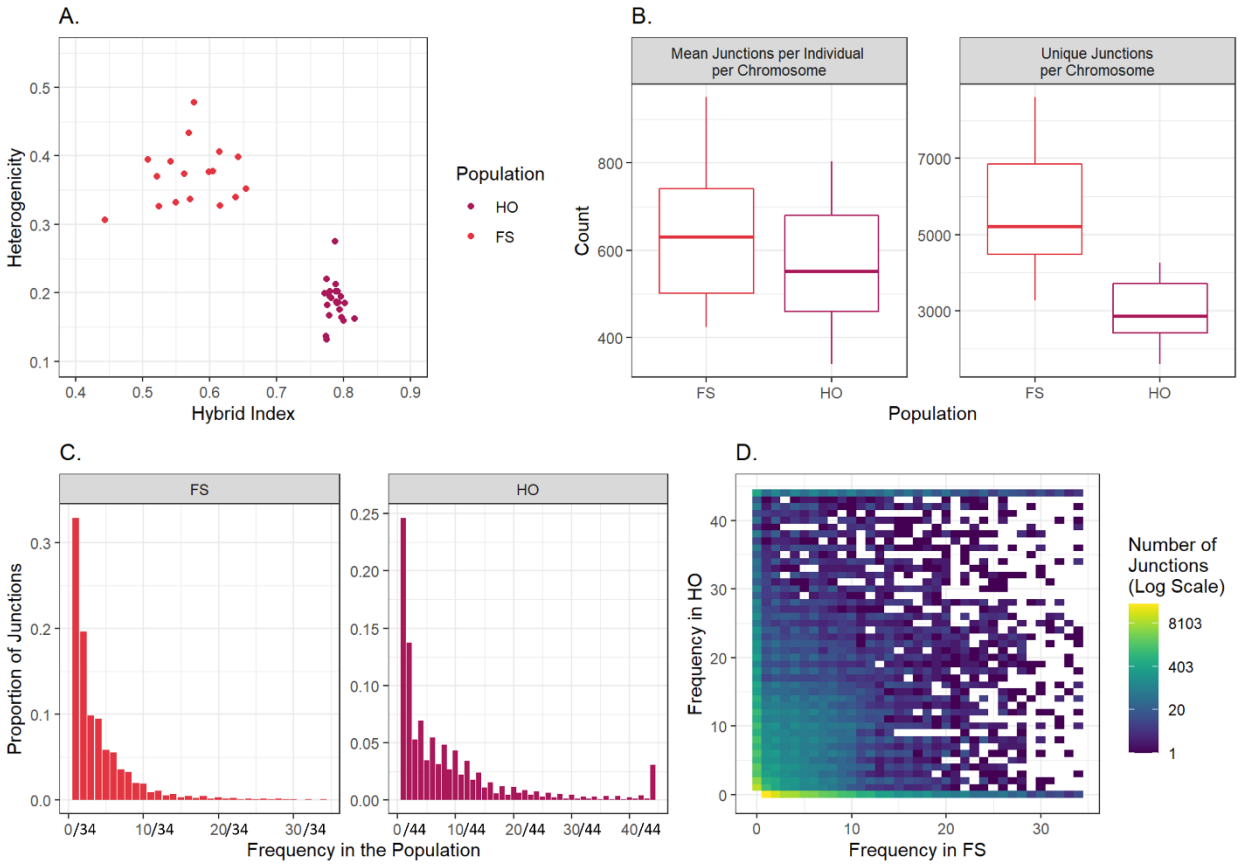


Figure 4.3. Variation in junction number among chromosomes (A) and among individuals (B).

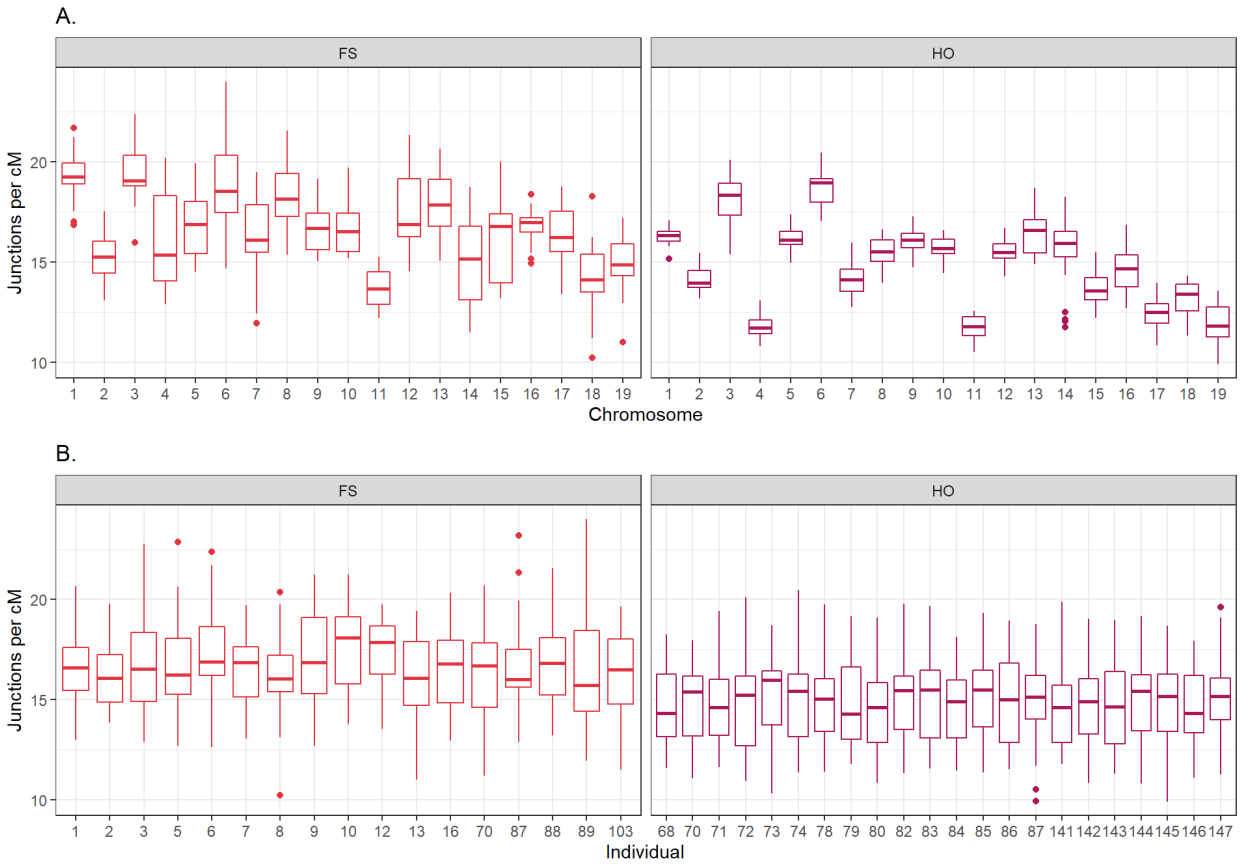


Figure 4.4. Genomic distributions of ancestry summary statistics measured in 1Mb windows and 1cM windows. Distributions for each population shown for A) hybrid index, B) heterogenicity, C) mean junctions per individual, D) total junctions, E) unique junctions, and F) singleton junctions.

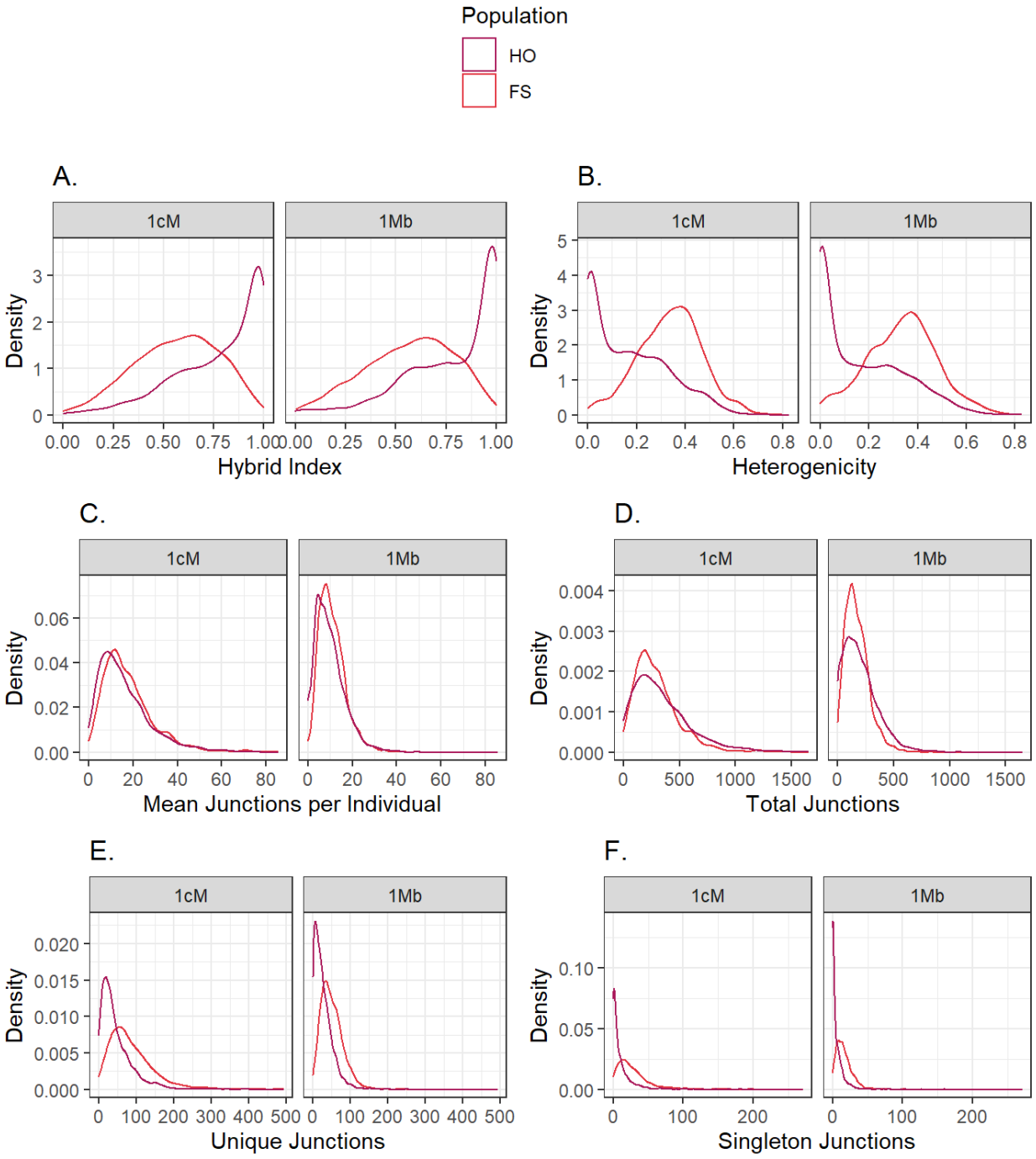
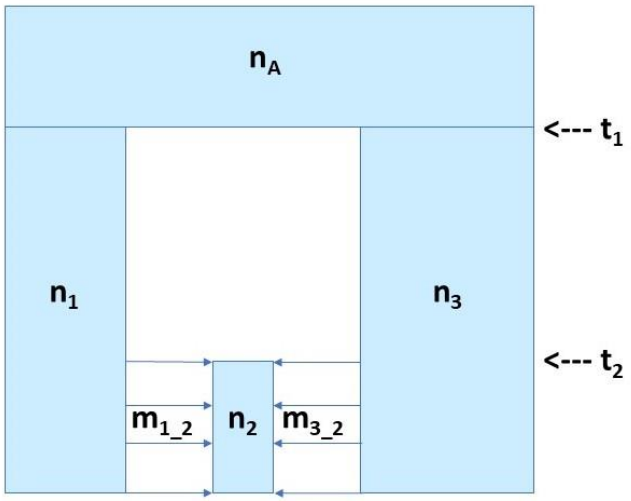


Figure 4.5. Summary of demographic analyses based on the SFS. A) The general model used for demographic analyses based on the site frequency spectrum of SNPs. An ancestral population of size n_A splits into two distinct populations of size n_1 and n_3 at time t_1. At time t_2, a hybrid population of size n_2 is created from a mix of the two parental populations. There is constant migration at rate m from the source populations into the hybrid population, but no migration back into the source populations. B) The observed SFS and best estimated SFS for each population.

A.



B.

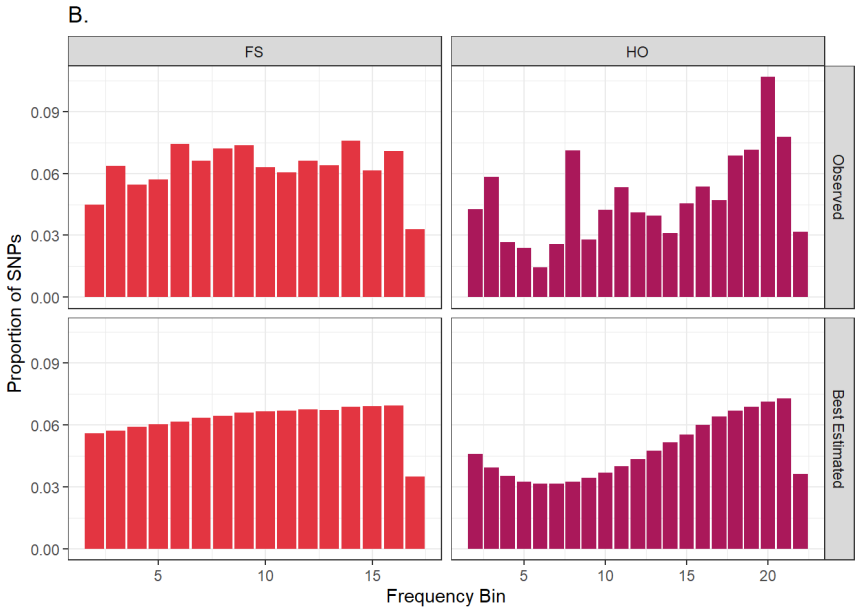


Figure 4.6. Distributions of estimated age of initial hybridization per individual per chromosome using A) the junction-based `estimate_time()` approach and B) the ancestry-based `estimate_time_diploid()` approach. Outliers above 4,000 generations have been removed from B for readability.

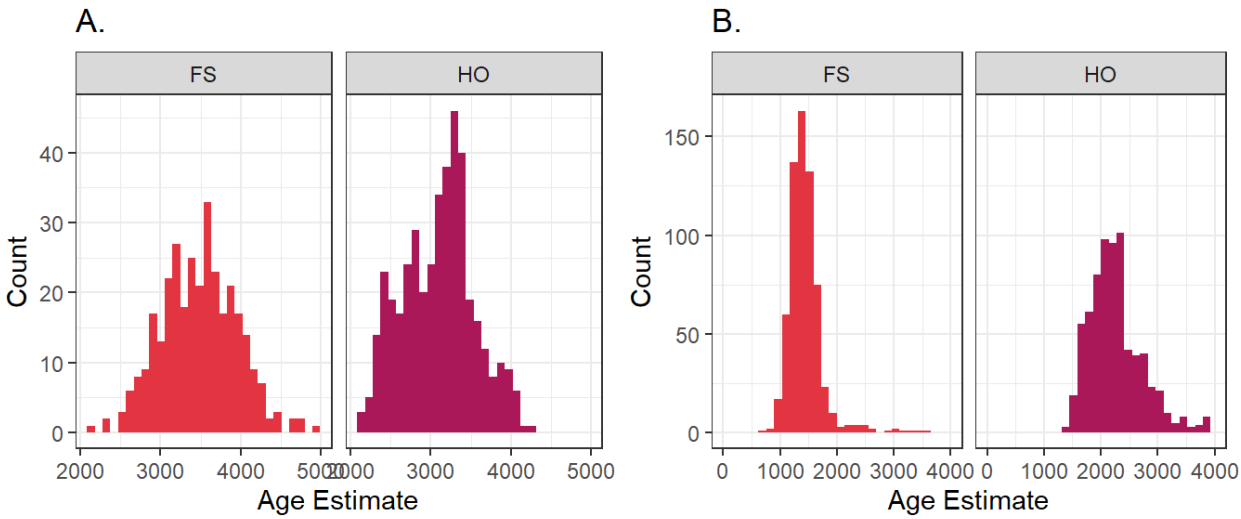


Figure 4.7. Outlier windows identified by several approaches (separated into rows).

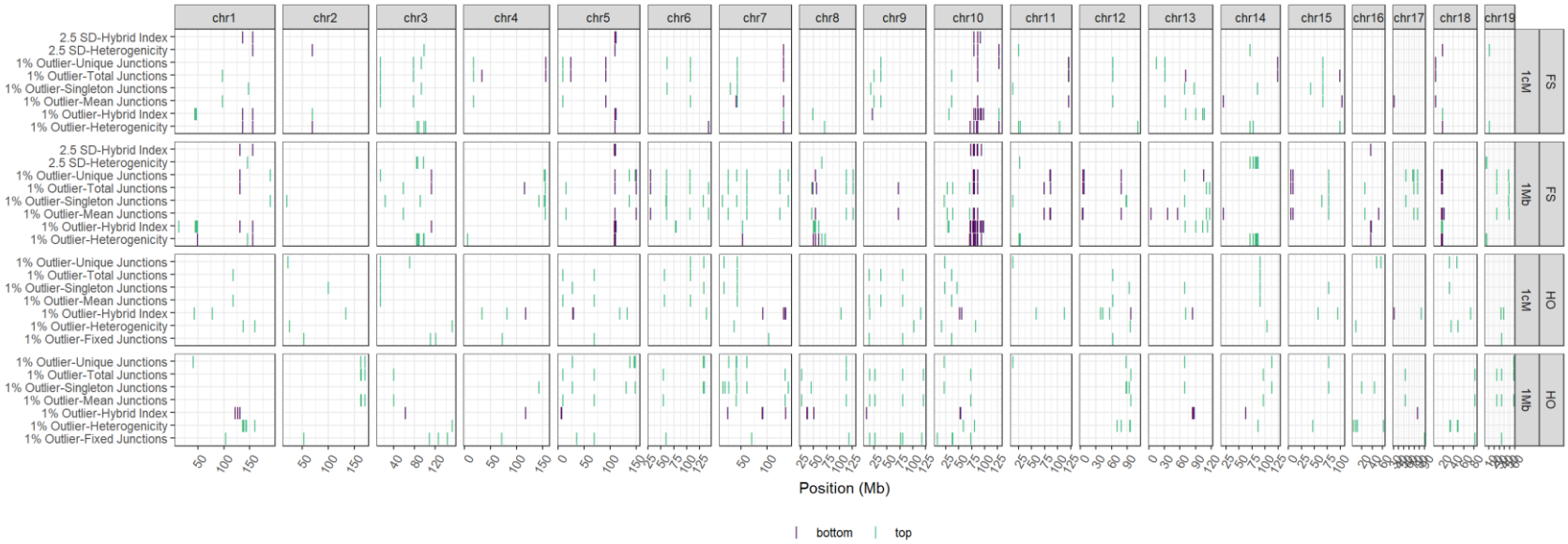


Figure 4.8. Region of interest on chromosome 5. Green lines indicate the edges of the region. A) Ancestry across the region in the FS population. B) Ancestry across the region in the HO population. C) Recombination rate shown as physical versus genetic position, and hybrid index and heterogenicity shown in 10kb windows.

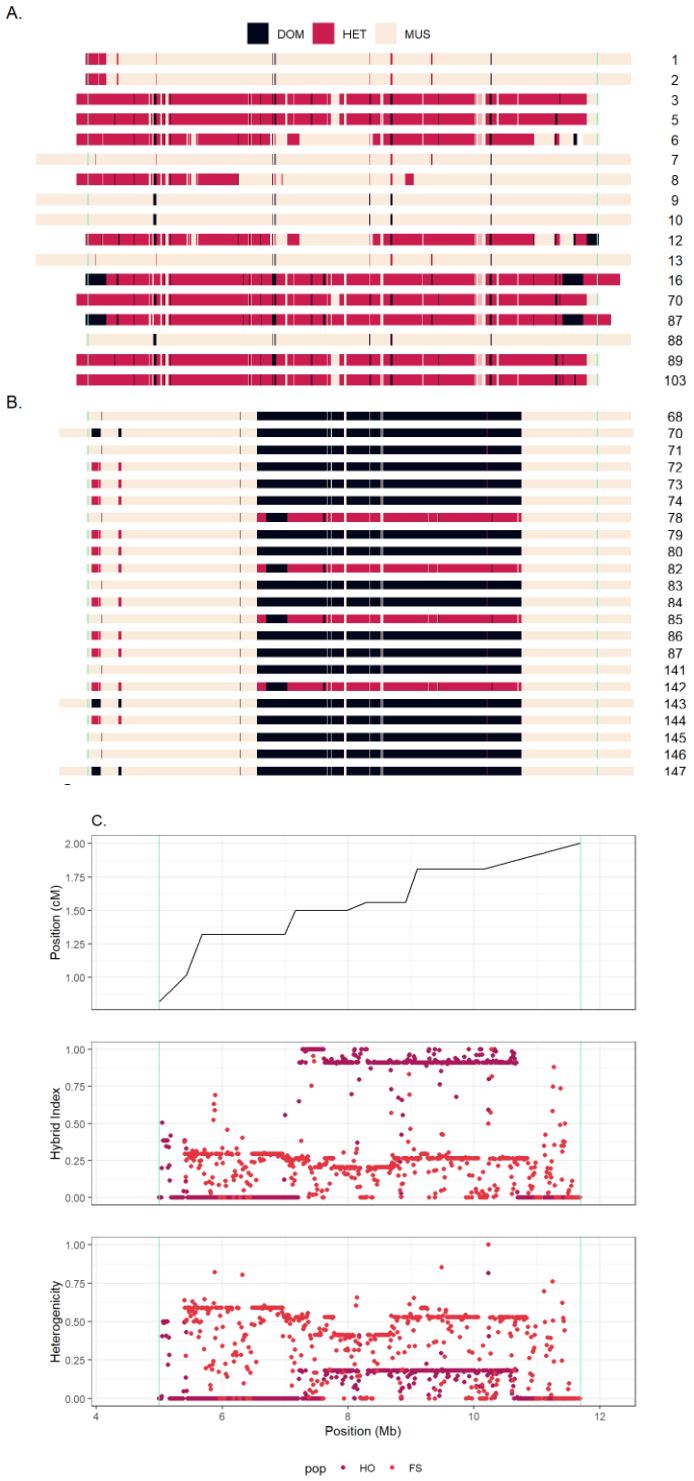


Figure 4.9. X chromosome ancestry. A) Ancestry maps of female X chromosomes versus chromosome 3 in the same individual are shown for a few individuals from each population. B) Junction counts are shown as a function of genetic length of chromosomes for each population. The black line shows the linear regression of these two variables.

Chapter 5

Conclusions

This dissertation aimed to improve our understanding of the behavior of genomic ancestry in hybrid zones, particularly ancestry junctions. In Chapter 2, I highlighted the importance of considering population structure for future studies of ancestry junctions. Complex hybrid zones where significant migration occurs between hybrid demes relative to migration from source populations cannot be adequately described by models of single hybrid populations. In Chapter 3, I attempted to infer demographic parameters under such a model. While my method was not successful, I presented potential pitfalls of my approach and suggestions for future work. In Chapter 4, I applied the junction framework to a classic hybrid zone. I showed that junctions can be readily identified in natural hybrid genomes and presented candidate incompatibility regions based on their genomic ancestry patterns.

The junction frequency spectrum as a metric for investigation of hybrid zones

The junction frequency spectrum (JFS) is a novel metric presented in Chapter 2 and estimated in both hybrid populations in Chapter 4. In principle, the JFS is similar to the site frequency spectrum (SFS) of SNPs. While the SFS is sensitive to admixture, the JFS may be more useful in the context of hybrid zones. First, it is specific to hybridization. Junctions can only form in the context of hybridization, and junctions that appear in multiple individuals imply common ancestry post-hybridization. Second, it is inherently unfolded. Allele frequency can either be calculated in terms of the minor allele frequency (folded) or the derived allele frequency (unfolded). The unfolded SFS is likely to be more informative, but estimating it requires polarization of the SNPs. Polarization may be particularly difficult for young species groups that provide conflicting signals for ancestral state (Keightley and Jackson, 2018). Incorrect polarization can generate a misleading SFS (Morton et al., 2009). On the other hand, junctions can only be derived and can be identified even when they are fixed. This may provide an advantage over the SFS, especially in the case of older hybrid populations.

Future theoretical work on the JFS should focus on understanding its sensitivity to selection and the development of inference methods. In particular, we were able to estimate the joint JFS for our two populations, but we have a limited theoretical basis for interpreting it. It is likely that a joint JFS between several hybrid populations could be highly informative about the history of gene flow between those populations. Future empirical work should focus on the fundamental challenge of junction identification, which would allow the application of the JFS in more taxa. In Chapter 4, we assumed junctions of the same identity were shared if they were assigned to the same location. This approach may be too conservative in some datasets and too permissive in others, depending on the specific procedure of ancestry inference and junction location assignment. A framework for probabilistic inference of junction sharing would greatly improve the prospects for estimating the JFS in more hybrid populations.

European House Mice: A classic hybrid zone

This dissertation complements the existing literature on the house mouse hybrid zone. In Chapter 4, we delineated genomic regions that show perturbations in ancestry we expect to see for loci targeted by natural selection. Some of these regions, such as the cluster on chromosome 10, have been identified previously in the literature. Identification of these regions as having biased ancestry patterns strengthens the case for involvement of these regions in reproductive isolation in nature. In some cases, the overlaps are with regions which caused subfertility in F_2 males from laboratory crosses of *musculus* and *domesticus* inbred strains. The connection between such regions and the natural hybrid zone is not always clear (Campbell et al., 2018), particularly when there are no early stage hybrids found in the hybrid zone. Our independent identification of these regions provides evidence that they may have affected hybrids within the zone, linking laboratory and natural studies. Other regions we identified that do not overlap with known QTL provide new avenues for investigation, such as the region on chromosome 5.

To my knowledge, the whole genome sequences analyzed here are the first to be reported from this classic hybrid zone. I have identified a huge amount of variation in both sequence and ancestry. The genomes generated for this dissertation will be a great resource for future studies of the hybrid zone. Despite over 70 years of study, there are still fundamental questions about the house mouse hybrid zone that are difficult to answer. The breadth of reproductive isolation that has been found in laboratory crosses is contrasted by the highly recombinant nature of the mice found in the zone—which is further showcased by the ancestry maps generated in Chapter 4 (all maps can be seen in Appendix B).

While there are outstanding questions to be answered about the progression of speciation in house mice, there is far more information about this zone than most known hybrid zones. The European house mouse hybrid zone has been studied from diverse genetic, geographic, ecological and behavioral angles. However, I view this as a strength, rather than a weakness, of this dissertation. This dissertation builds on the base of knowledge and provides opportunities for exploring new frameworks of variation, such as junctions, with more context. The fact that several outlier regions from Chapter 4 overlap with previously identified candidate loci is encouraging and suggests that our approach is successfully identifying regions under selection.

The importance of considering demography

While I have made strides towards utilizing the junction framework in a natural hybrid zone, more theoretical work is needed to understand ancestry under complex demographic histories. Obtaining additional predictions for baseline ancestry metrics in hybrid populations would allow us to better identify candidate regions. For example, I was unable to thoroughly investigate the X chromosome due to a lack of theoretical predictions for relative patterns on the X chromosome. Given the large role that the X chromosome plays in speciation (Coyne and Orr, 1989; Payseur et al., 2018), it would be valuable to develop a framework for interpreting junctions on sex chromosomes. That being

said, there have been recent efforts to incorporate more complicated models. Hvala *et al.* (2018) applied junctions to a stepping-stone model that incorporated migration and other aspects of population structure. This dissertation has built on that work to show that changes in demographic parameters within that structure can significantly affect the expected junction patterns. Sedghifar *et al.* (2015;2016) extended a basic model of ancestry tracts to a continuous population structure with nearest neighbor migration. Their results are distinct from those found in the discrete stepping-stone model, but this model is still an improvement over the single population model. Janzen *et al*. (2018) focused on a single population, but included the impact of small population size, thereby incorporating the effects of drift in a theoretical framework. All of these advances represent important building blocks towards a more comprehensive model of junctions/tracts that will allow us to better interpret findings in natural hybrid zones.

Beyond the framework of ancestry, there needs to be greater recognition of the effect of complex demography on hybrid zones in general. Hybrid zones may have complex demographic histories more often than other populations. They are likely to occur at range edges, where the population density may be more sparse and drift may be strong (Buggs, 2007). Hybrids are likely to experience more intense selection that could reduce population density. Variation is often reported within and between hybrid zones, such as: polymorphism of incompatibility loci within species (e.g. Mukaj et al. 2020), variation in the outcome of hybridization between independent meetings of species (e.g. Mandeville et al. 2017), or variation in introgression between transects of the same hybrid zone (e.g. Teeter et al. 2010). This variation further discourages the treatment of hybrid zones as uniform populations. Unfortunately, inferring complex demographic history is very difficult. In some cases, it can even be circular—information needed to infer demography may be unknown without first having some knowledge of demography (Johri et al., 2022). For example, in many taxa, recombination rates needed to infer demography are estimated from patterns of linkage disequilibrium, which are themselves

influenced by demography. Unfortunately, this difficulty often leads aspects of demography to be overlooked. Even commonly used approaches can make opposite assumptions about parameters as simple as the amount of migration. For example, LD-based DMI scans are prone to high false positive rates when there is ongoing parental migration (Schumer and Brandvain, 2016), whereas cline models assume such migration (Barton, 1979a).

Consideration of demography may be especially important in the house mouse hybrid zone. House mice are known to have a complicated population structure. Despite having high heterozygosity compared to other rodents (Berry, 1986; Fujiwara et al., 2022), mice are sometimes naturally inbred. Most adult mice stay within the subpopulation they were born into, although there is plenty of mixing with nearby demes (Berry and Jakobson, 1974). Field studies have suggested that demes of mice are relatively small (as few as 4; Berry and Jakobson 1974). It is believed that these small populations are unstable and mice may repeatedly or seasonally recolonize certain locations (Berry et al., 1982; Dallas et al., 1995; Wang et al., 2011). This structure would create opportunities for drift and recurrent admixture within the hybrid population. While this issue is not settled (see Baker 1981), this complex population structure could clearly impact the suitability of mice for models assuming large, panmictic populations, despite having a large effective population size. Population structures such as these could play an important role in the dynamics of a hybrid zone (Barton, 1979b), including in house mice where the hybrid zone has yet to be associated with an environmental barrier (Baird and Macholan, 2012).

How might this structure affect the accumulation of junctions? Although it was beyond the scope of this dissertation to model such a complex population structure, the stepping-stone model investigated in Chapter 2 may provide some insights. In Chapter 2, I found that there are essentially three “zones” of migration rates with unique effects on junction accumulation. The highest migration rates reduce junction number, presumably because migration of unbroken haplotypes is swamping out the effect of recombination in generating new junctions. At the lowest migration rates, junctions

accumulate at the same rate that they would in a closed population, with almost negligible effects from migration. At intermediate migration rates, however, more junctions can be accumulated than in a closed population. This appears to be due to a balance between migration providing enough new genetic material to increase heterogenicity (required for the formation of new junctions), without replacing too many junctions with unadmixed tracts. Importantly, this effect is much greater under the stepping-stone model than the hybrid swarm model with a single hybrid deme. This is likely because it is very rare for unadmixed individuals to make it to the central demes at low to intermediate migration rates. Thus, the migrants into the central demes are likely to carry junctions of their own—junctions that are unique due to the semi-independent processes of junction formation and drift occurring in the adjacent demes. I predict that this effect would be even more exaggerated in a model with recurrent extinction and recolonization. When new demes form, they would be colonized by individuals who were already highly recombinant. If there was little or no migration during the lifetime of the deme, the ancestries of the founding individuals would move towards a new fixed ancestry pattern until the extinction of the deme. Migrants leaving the deme would bring unique junctions from that deme to any new deme they colonize.

While we do not have a theoretical basis to test this hypothesis on the data from Chapter 4, the differences between our two populations are consistent with such a model. An extinction-recolonization model is likely to have increased variability between populations in ancestry, population age, and population size. The HO and FS populations have very different overall ancestries as measured by junction count, hybrid index and heterogenicity. The ages we estimated for the two populations varied, but they were different from each other under all three methods. Multiple lines of evidence suggest the populations have different sizes including differences in the number of fixed junctions, which was strongly associated with population size and isolation in Chapter 2. While these observations do not

exclude other population structures, they support continued investigation of the extinction-recolonization model.

Future Directions

In conclusion, junctions are an untapped source of hybrid zone information with considerable potential for revealing the genomic consequences of hybridization. While their identification can be challenging, the primary strength of junctions is their specificity to hybridization—a confidently identified junction is the direct result of admixture. This allows junctions to provide a complementary view of history to that which is drawn from sequence variation alone. Additional theoretical work should be focused on demographic inference to properly develop baseline ancestry predictions for inference of selection. Efforts to improve estimation of junction location would also be beneficial, particularly in genomes with a lower density of markers than occurs in house mice. This dissertation provides evidence that these are avenues worth pursuing.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J.E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C.A., Buggs, R., et al. (2013). Hybridization and speciation. Journal of Evolutionary Biology *26*, 229–246. https://doi.org/10.1111/j.1420-9101.2012.02599.x.

Albrechtová, J., Albrecht, T., Baird, S.J.E., Macholán, M., Rudolfsen, G., Munclinger, P., Tucker, P.K., and Piálek, J. (2012). Sperm-related phenotypes implicated in both maintenance and breakdown of a natural species barrier in the house mouse. Proceedings of the Royal Society B: Biological Sciences *279*, 4803–4810. https://doi.org/10.1098/rspb.2012.1802.

Auffray, J.-C., Alibert, P., Renaud, S., Orth, A., and Bonhomme, F. (1996). Fluctuating asymmetry in Mus musculus subspecific hybridization. In Advances in Morphometrics, L.F. Marcus, M. Corti, A. Loy, G.J.P. Naylor, and D.E. Slice, eds. (New York: Plenum Press), pp. 275–283.

Baird, S.J.E. (1995). A simulation study of multilocus clines. Evolution *49*, 1038–1045. https://doi.org/10.1111/j.1558-5646.1995.tb04431.x.

Baird, S.J.E. (2006). Fisher's markers of admixture. Heredity *97*, 81–83. https://doi.org/10.1038/sj.hdy.6800850.

Baird, S.J.E., and Macholan, M. (2012). What can the Mus musculus musculus/M. m. domesticus hybrid zone tell us about speciation? In Evolution of the House Mouse, M. Macholan, S.J.E. Baird, P. Munclinger, and J. Pialek, eds. (New York: Cambridge University Press), p.

Baird, S.J.E., Barton, N.H., and Etheridge, A.M. (2003). The distribution of surviving blocks of an ancestral genome. Theoretical Population Biology *64*, 451–471. https://doi.org/10.1016/S0040-5809(03)00098-4.

Baker, A.E.M. (1981). Gene flow in house mice: Introduction of a new allele into free-living populations. Evolution *35*, 243–258. https://doi.org/10.2307/2407835.

Balard, A., and Heitlinger, E. (2022). Shifting focus from resistance to disease tolerance: A review on hybrid house mice. Ecology and Evolution *12*, e8889. https://doi.org/10.1002/ece3.8889.

Bank, C., Bürger, R., and Hermisson, J. (2012). The limits to parapatric speciation: Dobzhansky–Muller incompatibilities in a continent–island model. Genetics *191*, 845–863. https://doi.org/10.1534/genetics.111.137513.

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics *28*, 1359–1367. https://doi.org/10.1093/bioinformatics/bts144.

Barton, N.H. (1979a). Gene flow past a cline. Heredity *43*, 333–339. https://doi.org/10.1038/hdy.1979.86.

Barton, N.H. (1979b). The dynamics of hybrid zones. Heredity *43*, 341–359. https://doi.org/10.1038/hdy.1979.87.

Barton, N.H. (1983). Multilocus clines. Evolution *37*, 454–471. https://doi.org/10.1111/j.1558-5646.1983.tb05563.x.

Barton, N.H., and Bengtsson, B.O. (1986). The barrier to genetic exchange between hybridising populations. Heredity *57*, 357–376. https://doi.org/10.1038/hdy.1986.135.

Barton, N.H., and Gale, K.S. (1993). Genetic analysis of hybrid zones. In Hybrid Zones and the Evolutionary Process, R.G. Harrison, ed. (New York: Oxford University Press), pp. 13–45.

Barton, N.H., and Hewitt, G.M. (1985). Analysis of hybrid zones. Annu. Rev. Ecol. Syst. *16*, 113–148. https://doi.org/10.1146/annurev.es.16.110185.000553.

Barton, N.H., and Whitlock, M.C. (1997). The evolution of metapopulations. In Metapopulation Biology, I. Hanski, and M.E. Gilpin, eds. (San Diego: Academic Press), pp. 183–210.

Beaumont, M.A. (2010). Approximate Bayesian Computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics *41*, 379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621.

Beaumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate Bayesian Computation in population genetics. Genetics *162*, 2025–2035. .

Beaumont, M.A., Cornuet, J.-M., Marin, J.-M., and Robert, C.P. (2009). Adaptive approximate Bayesian computation. Biometrika *96*, 983–990. https://doi.org/10.1093/biomet/asp052.

Berchowitz, L.E., and Copenhaver, G.P. (2010). Genetic interference: Don't stand so close to me. Curr Genomics *11*, 91–102. https://doi.org/10.2174/138920210790886835.

Berry, R.J. (1986). Genetics of insular populations of mammals, with particular reference to differentiation and founder effects in British small mammals. Biological Journal of the Linnean Society *28*, 205–230. https://doi.org/10.1111/j.1095-8312.1986.tb01754.x.

Berry, R.J., and Bronson, F.H. (1992). Life history and bioeconomy of the house mouse. Biological Reviews *67*, 519–550. https://doi.org/10.1111/j.1469-185X.1992.tb01192.x.

Berry, R.J., and Jakobson, M.E. (1974). Vagility in an island population of the house mouse. Journal of Zoology *173*, 341–354. https://doi.org/10.1111/j.1469-7998.1974.tb04119.x.

Berry, R.J., Cuthbert, A., and Peters, J. (1982). Colonization by house mice: an experiment. Journal of Zoology *198*, 329–336. https://doi.org/10.1111/j.1469-7998.1982.tb02079.x.

Bhattacharyya, T., Reifova, R., Gregorova, S., Simecek, P., Gergelits, V., Mistrik, M., Martincova, I., Pialek, J., and Forejt, J. (2014). X Chromosome control of meiotic chromosome synapsis in mouse inter-subspecific hybrids. PLOS Genetics *10*, e1004088. https://doi.org/10.1371/journal.pgen.1004088.

Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., Eppig, J.T., and the Mouse Genome Database Group (2011). The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Research *39*, D842–D848. https://doi.org/10.1093/nar/gkq1008.

Boursot, P., Auffray, J.-C., Britton-Davidian, J., and Bonhomme, F. (1993). The evolution of house mice. Annual Review of Ecology and Systematics *24*, 119–152. .

Braverman, J.M., Hudson, R.R., Kaplan, N., Langley, C.H., and Stephad, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics *140*, 783–796. .

Bridle, J.R., and Vines, T.H. (2007). Limits to evolution at range margins: when and why does adaptation fail? Trends in Ecology & Evolution *22*, 140–147. https://doi.org/10.1016/j.tree.2006.11.002.

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J.G., and Bustamante, C.D. (2012). PCAdmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol *84*, 343–364. https://doi.org/10.3378/027.084.0401.

Britton-Davidian, J., Fel-Clair, F., Lopez, J., Alibert, P., and Boursot, P. (2005). Postzygotic isolation between the two European subspecies of the house mouse: estimates from fertility patterns in wild and laboratory-bred hybrids. Biological Journal of the Linnean Society *84*, 379–393. https://doi.org/10.1111/j.1095-8312.2005.00441.x.

Browning, S.R., and Browning, B.L. (2011). Haplotype phasing: existing methods and new developments. Nature Reviews Genetics *12*, 703–714. https://doi.org/10.1038/nrg3054.

Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S.A., et al. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. U.S.A. *107*, 786–791. https://doi.org/10.1073/pnas.0909559107.

Buerkle, C.A., and Rieseberg, L.H. (2008). The rate of genome stabilization in homoploid hybrid species. Evolution *62*, 266–275. https://doi.org/10.1111/j.1558-5646.2007.00267.x.

Buggs, R.J.A. (2007). Empirical study of hybrid zone movement. Heredity *99*, 301–312. https://doi.org/10.1038/sj.hdy.6800997.

Bycroft, C., Fernandez-Rozadilla, C., Ruiz-Ponte, C., Quintela, I., Carracedo, Á., Donnelly, P., and Myers, S. (2019). Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. Nat Commun *10*, 1–14. https://doi.org/10.1038/s41467-018-08272-w.

Campbell, C.R., Poelstra, J.W., and Yoder, A.D. (2018). What is Speciation Genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. Biological Journal of the Linnean Society *124*, 561–583. https://doi.org/10.1093/biolinnean/bly063.

Chapman, N.H., and Thompson, E.A. (2002). The effect of population history on the lengths of ancestral chromosome segments. Genetics *162*, 449–458. .

Charlesworth, B., Coyne, J.A., and Barton, N.H. (1987). The relative rates of evolution of sex chromosomes and autosomes. The American Naturalist *130*, 113–146. https://doi.org/10.1086/284701.

Chiou, K.L., Bergey, C.M., Burrell, A.S., Disotell, T.R., Rogers, J., Jolly, C.J., and Phillips-Conroy, J.E. (2021). Genome-wide ancestry and introgression in a Zambian baboon hybrid zone. Molecular Ecology *30*, 1907–1920. https://doi.org/10.1111/mec.15858.

Church, D.M., Goodstadt, L., Hillier, L.W., Zody, M.C., Goldstein, S., She, X., Bult, C.J., Agarwala, R., Cherry, J.L., DiCuccio, M., et al. (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. PLOS Biology *7*, e1000112. https://doi.org/10.1371/journal.pbio.1000112.

Corbett-Detig, R., and Nielsen, R. (2017). A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy. PLOS Genetics *13*, e1006529. https://doi.org/10.1371/journal.pgen.1006529.

Coughlan, J.M., and Matute, D.R. (2020). The importance of intrinsic postzygotic barriers throughout the speciation process. Philosophical Transactions of the Royal Society B: Biological Sciences *375*, 20190533. https://doi.org/10.1098/rstb.2019.0533.

Cox, A., Ackert-Bicknell, C.L., Dumont, B.L., Ding, Y., Bell, J.T., Brockmann, G.A., Wergedal, J.E., Bult, C., Paigen, B., Flint, J., et al. (2009). A new standard genetic map for the laboratory mouse. Genetics *182*, 1335–1344. https://doi.org/10.1534/genetics.109.105486.

Coyne, J.A., and Orr, H.A. (1989). Patterns of speciation in Drosophila. Evolution *43*, 362–381. https://doi.org/10.1111/j.1558-5646.1989.tb04233.x.

Coyne, J.A., and Orr, H.A. (2004). Speciation (Sunderland: Sinauer Associates).

Csilléry, K., François, O., and Blum, M.G.B. (2012). abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution *3*, 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x.

Cucchi, T., Vigne, J.-D., and Auffray, J.-C. (2005). First occurrence of the house mouse (Mus musculus domesticus Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. Biological Journal of the Linnean Society *84*, 429–445. https://doi.org/10.1111/j.1095-8312.2005.00445.x.

Cucchi, T., Bălăşescu, A., Bem, C., Radu, V., Vigne, J.-D., and Tresset, A. (2011). New insights into the invasive process of the eastern house mouse (Mus musculus musculus): Evidence from the burnt houses of Chalcolithic Romania. The Holocene *21*, 1195–1202. https://doi.org/10.1177/0959683611405233.

Cutter, A.D. (2012). The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. Trends in Ecology & Evolution *27*, 209–218. https://doi.org/10.1016/j.tree.2011.11.004.

Dallas, J.F., Dod, B., Bourso, P., Prager, E.M., and Bonhomme, F. (1995). Population subdivision and gene flow in Danish house mice. Molecular Ecology *4*, 311–320. https://doi.org/10.1111/j.1365-294X.1995.tb00224.x.

Darwin, C. (1859). On the origin of species (London, UK: John Murray).

De La Torre, A., Ingvarsson, P.K., and Aitken, S.N. (2015). Genetic architecture and genomic patterns of gene flow between hybridizing species of Picea. Heredity *115*, 153–164. https://doi.org/10.1038/hdy.2015.19.

Dean, M.D., and Nachman, M.W. (2009). Faster fertilization rate in conspecific versus heterospecific matings in house mice. Evolution *63*, 20–28. https://doi.org/10.1111/j.1558-5646.2008.00499.x.

Dobzhansky, T. (1936). Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. Genetics *21*, 113–135. .

Dobzhansky, T. (1937). Genetics and the Origin of Species (New York, NY: Columbia University Press).

Dorus, S., Wasbrough, E.R., Busby, J., Wilkin, E.C., and Karr, T.L. (2010). Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. Mol Biol Evol *27*, 1235–1246. https://doi.org/10.1093/molbev/msq007.

Dudek, K., Gaczorek, T.S., Zieliński, P., and Babik, W. (2019). Massive introgression of major histocompatibility complex (MHC) genes in newt hybrid zones. Molecular Ecology *0*. https://doi.org/10.1111/mec.15254.

Dumont, B.L., and Payseur, B.A. (2011). Evolution of the genomic recombination rate in murid rodents. Genetics *187*, 643–657. https://doi.org/10.1534/genetics.110.123851.

Dumont, B.L., White, M.A., Steffy, B., Wiltshire, T., and Payseur, B.A. (2011). Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. Genome Res. *21*, 114–125. https://doi.org/10.1101/gr.111252.110.

Duranton, M., and Pool, J.E. (2022). Interactions between natural selection and recombination shape the genomic landscape of introgression. Mol Biol Evol msac122. https://doi.org/10.1093/molbev/msac122.

Duranton, M., Bonhomme, F., and Gagnaire, P.-A. (2019). The spatial scale of dispersal revealed by admixture tracts. Evolutionary Applications *12*, 1743–1756. https://doi.org/10.1111/eva.12829.

Ďureje, Ľ., Macholán, M., Baird, S.J.E., and Piálek, J. (2012). The mouse hybrid zone in Central Europe: from morphology to molecules. Folia Zoologica *61*, 308–318. https://doi.org/10.25225/fozo.v61.i3.a13.2012.

Duvaux, L., Belkhir, K., Boulesteix, M., and Boursot, P. (2011). Isolation and gene flow: inferring the speciation history of European house mice. Molecular Ecology *20*, 5248–5264. https://doi.org/10.1111/j.1365-294X.2011.05343.x.

Dzur-Gejdosova, M., Simecek, P., Gregorova, S., Bhattacharyya, T., and Forejt, J. (2012). Dissecting the Genetic Architecture of F1 Hybrid Sterility in House Mice. Evolution *66*, 3321–3335. https://doi.org/10.1111/j.1558-5646.2012.01684.x.

Endler, J.A. (1977). Geographic variation, speciation, and clines (Princeton, N.J.: Princeton University Press).

Feldman, M.W., and Christiansen, F.B. (1974). The effect of population subdivision on two loci without selection. Genetics Research *24*, 151–162. https://doi.org/10.1017/S0016672300015184.

Field, D.L., Ayre, D.J., Whelan, R.J., and Young, A.G. (2011). Patterns of hybridization and asymmetrical gene flow in hybrid zones of the rare Eucalyptus aggregata and common E. rubida. Heredity *106*, 841–853. .

Fisher, R.A. (1949). The Theory of Inbreeding (Edinburgh: Oliver and Boyd).

Fisher, R.A. (1954). A fuller theory of "Junctions" in inbreeding. Heredity *8*, 187–197. https://doi.org/10.1038/hdy.1954.17.

Forejt, J., Jansa, P., and Parvanov, E. (2021). Hybrid sterility genes in mice (Mus musculus): a peculiar case of PRDM9 incompatibility. Trends in Genetics *37*, 1095–1108. https://doi.org/10.1016/j.tig.2021.06.008.

Fortes-Lima, C.A., Laurent, R., Thouzeau, V., Toupance, B., and Verdu, P. (2021). Complex genetic admixture histories reconstructed with Approximate Bayesian Computation. Molecular Ecology Resources *21*, 1098–1117. https://doi.org/10.1111/1755-0998.13325.

Fraïsse, C., Roux, C., Gagnaire, P.-A., Romiguier, J., Faivre, N., Welch, J.J., and Bierne, N. (2018). The divergence history of European blue mussel species reconstructed from Approximate Bayesian Computation: the effects of sequencing techniques and sampling strategies. PeerJ *6*, e5198. https://doi.org/10.7717/peerj.5198.

Frayer, M.E., and Payseur, B.A. (2021). Demographic history shapes genomic ancestry in hybrid zones. Ecology and Evolution *11*, 10290–10302. https://doi.org/10.1002/ece3.7833.

Fujiwara, K., Kawai, Y., Takada, T., Shiroishi, T., Saitou, N., Suzuki, H., and Osada, N. (2022). Insights into Mus musculus population structure across Eurasia revealed by whole-genome analysis. Genome Biol Evol *14*, evac068. https://doi.org/10.1093/gbe/evac068.

Galaverni, M., Caniglia, R., Pagani, L., Fabbri, E., Boattini, A., and Randi, E. (2017). Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing wolf population. Mol Biol Evol *34*, 2324–2339. https://doi.org/10.1093/molbev/msx169.

Gavrilets, S. (1997). Hybrid zones with Dobzhansky-type epistatic selection. Evolution *51*, 1027–1035. https://doi.org/10.2307/2411031.

Geraldes, A., Basset, P., Gibson, B., Smith, K.L., Harr, B., Yu, H.-T., Bulatova, N., Ziv, Y., and Nachman, M.W. (2008). Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Molecular Ecology *17*, 5349–5363. https://doi.org/10.1111/j.1365-294X.2008.04005.x.

Gompert, Z., and Buerkle, C.A. (2011). Bayesian estimation of genomic clines. Molecular Ecology *20*, 2111–2127. https://doi.org/10.1111/j.1365-294X.2011.05074.x.

Gompert, Z., and Buerkle, C.A. (2013). Analyses of genetic ancestry enable key insights for molecular ecology. Mol Ecol *22*, 5278–5294. https://doi.org/10.1111/mec.12488.

Gompert, Z., Mandeville, E.G., and Buerkle, C.A. (2017). Analysis of population genomic data from hybrid zones. Annual Review of Ecology, Evolution, and Systematics *48*, 207–229. https://doi.org/10.1146/annurev-ecolsys-110316-022652.

Good, J.M., Dean, M.D., and Nachman, M.W. (2008). A complex genetic basis to X-linked hybrid male sterility between two species of house mice. Genetics *179*, 2213–2228. https://doi.org/10.1534/genetics.107.085340.

Gravel, S. (2012). Population genetics models of local ancestry. Genetics *191*, 607–619. https://doi.org/10.1534/genetics.112.139808.

Gregorova, S., Gergelits, V., Chvatalova, I., Bhattacharyya, T., Valiskova, B., Fotopulosova, V., Jansa, P., Wiatrowska, D., and Forejt, J. (2018). Modulation of Prdm9-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice. ELife Sciences *7*, e34282. https://doi.org/10.7554/eLife.34282.

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. Genetics *196*, 625–642. https://doi.org/10.1534/genetics.113.160697.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLOS Genetics *5*, e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Haenel, Q., Laurentino, T.G., Roesti, M., and Berner, D. (2018). Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol. Ecol. *27*, 2477–2497. https://doi.org/10.1111/mec.14699.

Haldane, J.B.S. (1922). Sex ratio and unisexual sterility in hybrid animals. Journ. of Gen. *12*, 101–109. https://doi.org/10.1007/BF02983075.

Harr, B., Karakoc, E., Neme, R., Teschke, M., Pfeifle, C., Pezer, Ž., Babiker, H., Linnenbrink, M., Montero, I., Scavetta, R., et al. (2016). Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Scientific Data *3*, 160075. https://doi.org/10.1038/sdata.2016.75.

Harrison, R.G. (1990). Hybrid zones: windows on evolutionary process. In Oxford Surveys in Evolutionary Biology, D. Futuyma, and J. Antonovics, eds. (Oxford University Press, USA), pp. 69–128.

Harrison, R.G. (1993). Hybrids and hybrid zones: historical perspective. In Hybrid Zones and the Evolutionary Process, R.G. Harrison, ed. (New York: Oxford University Press), pp. 3–12.

Harrison, R.G., and Larson, E.L. (2014). Hybridization, introgression, and the nature of species boundaries. Journal of Heredity *105*, 795–809. https://doi.org/10.1093/jhered/esu033.

Harrison, R.G., and Rand, D.M. (1989). Mosaic hybrid zones and the nature of species boundaries. In Speciation and Its Consequences, J. Endler, and D. Otte, eds. (Sunderland, Massachusetts: Sinauer Associates), pp. 111–133.

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science *343*, 747–751. https://doi.org/10.1126/science.1243518.

Henn, B.M., Botigué, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlaoui-Zid, K., Zalloua, P.A., Moreno-Estrada, A., Bertranpetit, J., et al. (2012). Genomic ancestry of north Africans supports back-to-Africa migrations. PLOS Genetics *8*, e1002397. https://doi.org/10.1371/journal.pgen.1002397.

Hewitt, G.M. (1988). Hybrid zones-natural laboratories for evolutionary studies. Trends in Ecology & Evolution *3*, 158–167. https://doi.org/10.1016/0169-5347(88)90033-X.

Hovde Liland, K., Mevik, B.-H., and Wehrens, R. (2021). pls: Partial least squares and principal component regression.

Hunt, W.G., and Selander, R.K. (1973). Biochemical genetics of hybridisation in European house mice. Heredity *31*, 11–33. .

Hurst, J.L., Beynon, R.J., Armstrong, S.D., Davidson, A.J., Roberts, S.A., Gómez-Baena, G., Smadja, C.M., and Ganem, G. (2017). Molecular heterogeneity in major urinary proteins of *Mus musculus* subspecies: potential candidates involved in speciation. Scientific Reports *7*, 44992. https://doi.org/10.1038/srep44992.

Hvala, J.A., Frayer, M.E., and Payseur, B.A. (2018). Signatures of hybridization and speciation in genomic patterns of ancestry. Evolution *72*, 1540–1552. https://doi.org/10.1111/evo.13509.

Irwin, D.E. (2018). Sex chromosomes and speciation in birds and other ZW systems. Molecular Ecology *27*, 3831–3851. https://doi.org/10.1111/mec.14537.

Irwin, D.E. (2020). Assortative mating in hybrid zones Is remarkably ineffective in promoting speciation. The American Naturalist *195*, E150–E167. https://doi.org/10.1086/708529.

Janoušek, V., Wang, L., Luzynski, K., Dufková, P., Vyskočilová, M.M., Nachman, M.W., Munclinger, P., Macholán, M., Piálek, J., and Tucker, P.K. (2012). Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between Mus musculus musculus and M. m. domesticus. Molecular Ecology *21*, 3032–3047. https://doi.org/10.1111/j.1365-294X.2012.05583.x.

Janousek, V., Munclinger, P., Wang, L., Teeter, K.C., and Tucker, P.K. (2015). Functional organization of the genome may shape the species boundary in the house mouse. Molecular Biology and Evolution *32*, 1208–1220. https://doi.org/10.1093/molbev/msv011.

Janzen, T., and Miró Pina, V. (2022). Estimating the time since admixture from phased and unphased molecular data. Molecular Ecology Resources *22*, 908–926. https://doi.org/10.1111/1755-0998.13519.

Janzen, T., Nolte, A.W., and Traulsen, A. (2018). The breakdown of genomic ancestry blocks in hybrid lineages given a finite number of recombination sites. Evolution *72*, 735–750. https://doi.org/10.1111/evo.13436.

Jeong, C., Witonsky, D.B., Basnyat, B., Neupane, M., Beall, C.M., Childs, G., Craig, S.R., Novembre, J., and Rienzo, A.D. (2018). Detecting past and ongoing natural selection among ethnically Tibetan women at high altitude in Nepal. PLOS Genetics *14*, e1007650. https://doi.org/10.1371/journal.pgen.1007650.

Johri, P., Aquadro, C.F., Beaumont, M., Charlesworth, B., Excoffier, L., Eyre-Walker, A., Keightley, P.D., Lynch, M., McVean, G., Payseur, B.A., et al. (2022). Recommendations for improving statistical inference in population genomics. PLOS Biology *20*, e3001669. https://doi.org/10.1371/journal.pbio.3001669.

Jones, E.P., Van Der KOOIJ, J., Solheim, R., and Searle, J.B. (2010). Norwegian house mice (Mus musculus musculus/domesticus): distributions, routes of colonization and patterns of hybridization. Molecular Ecology *19*, 5252–5264. https://doi.org/10.1111/j.1365-294X.2010.04874.x.

Keightley, P.D., and Jackson, B.C. (2018). Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. Genetics *209*, 897–906. https://doi.org/10.1534/genetics.118.301120.

Kenney, A.M., and Sweigart, A.L. (2016). Reproductive isolation and introgression between sympatric Mimulus species. Mol Ecol *25*, 2499–2517. https://doi.org/10.1111/mec.13630.

Kessner, D., and Novembre, J. (2014). forqs: forward-in-time simulation of recombination, quantitative traits and selection. Bioinformatics *30*, 576–577. https://doi.org/10.1093/bioinformatics/btt712.

Kim, B.Y., Huber, C.D., and Lohmueller, K.E. (2018). Deleterious variation shapes the genomic landscape of introgression. PLOS Genetics *14*, e1007741. https://doi.org/10.1371/journal.pgen.1007741.

Kimura, M., and Weiss, G.H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics *49*, 561–576. .

Kimura, M., Maruyama, T., and Crow, J.F. (1963). The mutation load in small populations. Genetics *48*, 1303–1312. .

Langdon, Q.K., Powell, D.L., Kim, B., Banerjee, S.M., Payne, C., Dodge, T.O., Moran, B., Fascinetto-Zago, P., and Schumer, M. (2022). Predictability and parallelism in the contemporary evolution of hybrid genomes. PLoS Genet *18*, e1009914. https://doi.org/10.1371/journal.pgen.1009914.

Larson, E.L., Keeble, S., Vanderpool, D., Dean, M.D., and Good, J.M. (2017). The composite regulatory basis of the large X-effect in mouse speciation. Molecular Biology and Evolution msw243. https://doi.org/10.1093/molbev/msw243.

Larson, E.L., Vanderpool, D., Sarver, B.A.J., Callahan, C., Keeble, S., Provencio, L.L., Kessler, M.D., Stewart, V., Nordquist, E., Dean, M.D., et al. (2018). The evolution of polymorphic hybrid incompatibilities in house mice. Genetics *209*, 845–859. https://doi.org/10.1534/genetics.118.300840.

Latour, Y., Perriat-Sanguinet, M., Caminade, P., Boursot, P., Smadja, C.M., and Ganem, G. (2014). Sexual selection against natural hybrids may contribute to reinforcement in a house mouse hybrid zone. Proc. R. Soc. B *281*, 20132733. https://doi.org/10.1098/rspb.2013.2733.

Laukaitis, C.M., Critser, E.S., and Karn, R.C. (1997). Salivary androgen-binding protein (ABP) mediates sexual isolation in Mus musculus. Evolution *51*, 2000–2005. https://doi.org/10.2307/2411020.

Lavretsky, P., Janzen, T., and McCracken, K.G. (2019). Identifying hybrids & the genomics of hybridization: Mallards & American black ducks of Eastern North America. Ecology and Evolution *9*, 3470–3490. https://doi.org/10.1002/ece3.4981.

Leitwein, M., Gagnaire, P.-A., Desmarais, E., Berrebi, P., and Guinand, B. (2018). Genomic consequences of a recent three-way admixture in supplemented wild brown trout populations revealed by local ancestry tracts. Molecular Ecology *27*, 3466–3483. https://doi.org/10.1111/mec.14816.

Liang, M., and Nielsen, R. (2014). The lengths of admixture tracts. Genetics *197*, 953–967. https://doi.org/10.1534/genetics.114.162362.

Lindtke, D., and Buerkle, C.A. (2015). The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. Evolution *69*, 1987–2004. https://doi.org/10.1111/evo.12725.

Loire, E., Tusso, S., Caminade, P., Severac, D., Boursot, P., Ganem, G., and Smadja, C.M. (2017). Do changes in gene expression contribute to sexual isolation and reinforcement in the house mouse? Mol Ecol *26*, 5189–5202. https://doi.org/10.1111/mec.14212.

Lustyk, D., Kinský, S., Ullrich, K.K., Yancoskie, M., Kašíková, L., Gergelits, V., Sedlacek, R., Chan, Y.F., Odenthal-Hesse, L., Forejt, J., et al. (2019). Genomic structure of Hstx2 modifier of Prdm9-dependent hybrid male sterility in mice. Genetics genetics.302554.2019. https://doi.org/10.1534/genetics.119.302554.

Macholán, M., Munclinger, P., Šugerková, M., Dufková, P., Bímová, B., Božíková, E., Zima, J., and Piálek, J. (2007). Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. Evolution *61*, 746–771. .

Mack, K.L., Campbell, P., and Nachman, M.W. (2016). Gene regulation and speciation in house mice. Genome Res. *26*, 451–461. https://doi.org/10.1101/gr.195743.115.

Maheshwari, S., and Barbash, D.A. (2011). The genetics of hybrid incompatibilities. Annual Review of Genetics *45*, 331–355. https://doi.org/10.1146/annurev-genet-110410-132514.

Mandeville, E.G., Parchman, T.L., Thompson, K.G., Compton, R.I., Gelwicks, K.R., Song, S.J., and Buerkle, C.A. (2017). Inconsistent reproductive isolation revealed by interactions between Catostomus fish species. Evolution Letters *1*, 255–268. https://doi.org/10.1002/evl3.29.

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. (2003). Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci U S A *100*, 15324–15328. https://doi.org/10.1073/pnas.0306899100.

Maruyama, T., and Kimura, M. (1980). Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. PNAS *77*, 6710–6714. https://doi.org/10.1073/pnas.77.11.6710.

Masly, J.P., and Presgraves, D.C. (2007). High-resolution genome-wide dissection of the two rules of speciation in Drosophila. PLOS Biology *5*, e243. https://doi.org/10.1371/journal.pbio.0050243.

Mayr, E. (1942). Systematics and the Origin of Species, from the Viewpoint of a Zoologist (Harvard University Press).

McFarlane, S.E., Senn, H.V., Smith, S.L., and Pemberton, J.M. (2021). Locus-specific introgression in young hybrid swarms: Drift may dominate selection. Molecular Ecology *30*, 2104–2115. https://doi.org/10.1111/mec.15862.

Medina, P., Thornlow, B., Nielsen, R., and Corbett-Detig, R. (2018). Estimating the timing of multiple admixture pulses during local ancestry inference. Genetics *210*, 1089–1107. https://doi.org/10.1534/genetics.118.301411.

Mikula, O., and Macholán, M. (2008). There is no heterotic effect upon developmental stability in the ventral side of the skull within the house mouse hybrid zone. Journal of Evolutionary Biology *21*, 1055–1067. https://doi.org/10.1111/j.1420-9101.2008.01539.x.

Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., and Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. Nat Commun *8*, 15183. https://doi.org/10.1038/ncomms15183.

Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., and Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. PLOS Genetics *7*, e1001373. https://doi.org/10.1371/journal.pgen.1001373.

Morgan, K., Harr, B., White, M.A., Payseur, B.A., and Turner, L.M. (2020). Disrupted gene networks in subfertile hybrid house mice. Molecular Biology and Evolution *37*, 1547–1562. https://doi.org/10.1093/molbev/msaa002.

Morton, B.R., Dar, V.-N., and Wright, S.I. (2009). Analysis of site frequency spectra from Arabidopsis with context-dependent corrections for ancestral misinference. Plant Physiology *149*, 616–624. https://doi.org/10.1104/pp.108.127787.

Moyle, L.C., Muir, C.D., Han, M.V., and Hahn, M.W. (2010). The contribution of gene movement to the "Two Rules of Speciation." Evolution *64*, 1541–1557. https://doi.org/10.1111/j.1558-5646.2010.00990.x.

Muirhead, C.A., and Presgraves, D.C. (2016). Hybrid incompatibilities, local adaptation, and the genomic distribution of natural introgression between species. The American Naturalist *187*, 249–261. https://doi.org/10.1086/684583.

Mukaj, A., Piálek, J., Fotopulosova, V., Morgan, A.P., Odenthal-Hesse, L., Parvanov, E.D., and Forejt, J. (2020). Prdm9 intersubspecific interactions in hybrid male sterility of house mouse. Molecular Biology and Evolution *37*, 3423–3438. https://doi.org/10.1093/molbev/msaa167.

Muller, H.J. (1942). Isolating mechanisms, evolution and temperature. Biol Symp *6*, 71–125. .

Nachman, M.W. (2002). Variation in recombination rate across the genome: evidence and implications. Current Opinion in Genetics & Development *12*, 657–663. https://doi.org/10.1016/S0959-437X(02)00358-1.

Orr, H.A. (1989). Does postzygotic isolation result from improper dosage compensation? Genetics *122*, 891–894. https://doi.org/10.1093/genetics/122.4.891.

Orr, H.A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics *139*, 1805–1813. .

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and Eigenanalysis. PLOS Genetics *2*, e190. https://doi.org/10.1371/journal.pgen.0020190.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.

Payseur, B.A. (2010). Using differential introgression in hybrid zones to identify genomic regions involved in speciation. Molecular Ecology Resources *10*, 806–820. https://doi.org/10.1111/j.1755-0998.2010.02883.x.

Payseur, B.A., and Rieseberg, L.H. (2016). A genomic perspective on hybridization and speciation. Molecular Ecology *25*, 2337–2360. https://doi.org/10.1111/mec.13557.

Payseur, B.A., Krenz, J.G., and Nachman, M.W. (2004). Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution *58*, 2064–2078. .

Payseur, B.A., Presgraves, D.C., and Filatov, D.A. (2018). Sex chromosomes and speciation. Mol Ecol *27*, 3745–3748. https://doi.org/10.1111/mec.14828.

Peterson, A.L., and Payseur, B.A. (2021). Sex-specific variation in the genome-wide recombination rate. Genetics *217*, iyaa019. https://doi.org/10.1093/genetics/iyaa019.

Phifer-Rixey, M., and Nachman, M.W. (2015). Insights into mammalian biology from the wild house mouse Mus musculus. ELife *4*, e05959. https://doi.org/10.7554/eLife.05959.

Phifer-Rixey, M., Harr, B., and Hey, J. (2020). Further resolution of the house mouse (Mus musculus) phylogeny by integration over isolation-with-migration histories. BMC Evol Biol *20*, 120. https://doi.org/10.1186/s12862-020-01666-9.

Polechová, J., and Barton, N.H. (2011). Genetic drift widens the expected cline but narrows the expected cline width. Genetics *189*, 227–235. https://doi.org/10.1534/genetics.111.129817.

Pool, J.E. (2015). The mosaic ancestry of the Drosophila Genetic Reference Panel and the D. melanogaster reference genome reveals a network of epistatic fitness interactions. Mol Biol Evol *32*, 3236–3251. https://doi.org/10.1093/molbev/msv194.

Pool, J.E., and Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics *181*, 711–719. https://doi.org/10.1534/genetics.108.098095.

Presgraves, D.C. (2018). Evaluating genomic signatures of “the large X-effect” during complex speciation. Molecular Ecology *27*, 3822–3830. https://doi.org/10.1111/mec.14777.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet *38*, 904–909. https://doi.org/10.1038/ng1847.

Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLOS Genetics *5*, e1000519. https://doi.org/10.1371/journal.pgen.1000519.

Rand, D.M., and Harrison, R.G. (1989). Ecological genetics of a mosaic hybrid zone: Mitochondrial, nuclear, and reproductive differentiation of crickets by soil type. Evolution *43*, 432–449. https://doi.org/10.1111/j.1558-5646.1989.tb04238.x.

Raufaste, N., Orth, A., Belkhir, K., Senet, D., Smadja, C., Baird, S.J.E., Bonhomme, F., Dod, B., and Boursot, P. (2005). Inferences of selection and migration in the Danish house mouse hybrid zone. Biological Journal of the Linnean Society *84*, 593–616. https://doi.org/10.1111/j.1095-8312.2005.00457.x.

Ravinet, M., Faria, R., Butlin, R.K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. a. F., Mehlig, B., and Westram, A.M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J. Evol. Biol. *30*, 1450–1477. https://doi.org/10.1111/jeb.13047.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLOS Biology *14*, e2000234. https://doi.org/10.1371/journal.pbio.2000234.

Sage, R.D., Heyneman, D., Lim, K.-C., and Wilson, A.C. (1986a). Wormy mice in a hybrid zone. Nature *324*, 60–63. .

Sage, R.D., Whitney, J.B.I., and Wilson, A.C. (1986b). Genetic analysis of a hybrid zone between domesticus and musculus mice (Mus musculus complex): hemoglobin polymorphisms. In The Wild Mouse in Immunology, M. Potter, J.H. Nadeau, and M.P. Cancro, eds. (Springer Berlin Heidelberg), p.

Sage, R.D., Atchley, W.R., and Capanna, E. (1993). House mice as models in systematic biology. Systematic Biology *42*, 523. https://doi.org/10.2307/2992487.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357. https://doi.org/10.1038/nature12961.

Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Current Biology *26*, 1241–1247. https://doi.org/10.1016/j.cub.2016.03.037.

Schumer, M., and Brandvain, Y. (2016). Determining epistatic selection in admixed populations. Mol Ecol *25*, 2577–2591. https://doi.org/10.1111/mec.13641.

Schumer, M., Xu, C., Powell, D.L., Durvasula, A., Skov, L., Holland, C., Blazier, J.C., Sankararaman, S., Andolfatto, P., Rosenthal, G.G., et al. (2018). Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science *360*, 656–660. https://doi.org/10.1126/science.aar3684.

Schwahn, D.J., Wang, R.J., White, M.A., and Payseur, B.A. (2018). Genetic dissection of hybrid male sterility across stages of spermatogenesis. Genetics *210*, 1453–1465. https://doi.org/10.1534/genetics.118.301658.

Sedghifar, A., Brandvain, Y., Ralph, P., and Coop, G. (2015). The spatial mixing of genomes in secondary contact zones. Genetics *201*, 243–261. https://doi.org/10.1534/genetics.115.179838.

Sedghifar, A., Brandvain, Y., and Ralph, P. (2016). Beyond clines: lineages and haplotype blocks in hybrid zones. Mol Ecol *25*, 2559–2576. https://doi.org/10.1111/mec.13677.

Seidman, D.N., Shenoy, S.A., Kim, M., Babu, R., Woods, I.G., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., et al. (2020). Rapid, phase-free detection of long identity-by-descent segments enables effective relationship classification. Am J Hum Genet *106*, 453–466. https://doi.org/10.1016/j.ajhg.2020.02.012.

Smadja, C., and Ganem, G. (2002). Subspecies recognition in the house mouse: a study of two populations from the border of a hybrid zone. Behav Ecol *13*, 312–320. https://doi.org/10.1093/beheco/13.3.312.

Smadja, C., and Ganem, G. (2005). Asymmetrical reproductive character displacement in the house mouse. Journal of Evolutionary Biology *18*, 1485–1493. https://doi.org/10.1111/j.1420-9101.2005.00944.x.

Smagulova, F., Brick, K., Pu, Y., Camerini-Otero, R.D., and Petukhova, G.V. (2016). The evolutionary turnover of recombination hot spots contributes to speciation in mice. Genes Dev. *30*, 266–280. https://doi.org/10.1101/gad.270009.115.

Suarez-Gonzalez, A., Hefer, C.A., Christe, C., Corea, O., Lexer, C., Cronk, Q.C.B., and Douglas, C.J. (2016). Genomic and functional approaches reveal a case of adaptive introgression from Populus balsamifera (balsam poplar) in P. trichocarpa (black cottonwood). Mol Ecol *25*, 2427–2442. https://doi.org/10.1111/mec.13539.

Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian Computation. PLOS Computational Biology *9*, e1002803. https://doi.org/10.1371/journal.pcbi.1002803.

Suzuki, T.A., and Nachman, M.W. (2015). Speciation and reduced hybrid female fertility in house mice. Evolution *69*, 2468–2481. https://doi.org/10.1111/evo.12747.

Suzuki, H., Nunome, M., Kinoshita, G., Aplin, K.P., Vogel, P., Kryukov, A.P., Jin, M.-L., Han, S.-H., Maryanto, I., Tsuchiya, K., et al. (2013). Evolutionary and dispersal history of Eurasian house mice Mus musculus clarified by more extensive geographic sampling of mitochondrial DNA. Heredity *111*, 375–390. https://doi.org/10.1038/hdy.2013.60.

Szymura, J.M., and Barton, N.H. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads, Bombina bombina and B. variegata, near Cracow in southern Poland. Evolution https://doi.org/10.1111/j.1558-5646.1986.tb05740.x.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595. .

Tao, Y., and Hartl, D.L. (2003). Genetic dissection of hybrid incompatibilities between Drosophila simulans and D. mauritiana.: III. Heterogeneous accumulation of hybrid incompatibilities, degree of

dominance, and implications for Haldane Rule. Evolution *57*, 2580–2598. https://doi.org/10.1111/j.0014-3820.2003.tb01501.x.

Teeter, K.C., Payseur, B.A., Harris, L.W., Bakewell, M.A., Thibodeau, L.M., O'Brien, J.E., Krenz, J.G., Sans-Fuentes, M.A., Nachman, M.W., and Tucker, P.K. (2008). Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. *18*, 67–76. https://doi.org/10.1101/gr.6757907.

Teeter, K.C., Thibodeau, L.M., Gompert, Z., Buerkle, C.A., Nachman, M.W., and Tucker, P.K. (2010). The variable genomic architecture of isolation between hybridizing species of house mice. Evolution *64*, 472–485. https://doi.org/10.1111/j.1558-5646.2009.00846.x.

Tucker, P.K., Sage, R.D., Warner, J., Wilson, A.C., and Eicher, E.M. (1992). Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. Evolution *46*, 1146–1163. https://doi.org/10.1111/j.1558-5646.1992.tb00625.x.

Turelli, M., and Orr, H.A. (1995). The dominance theory of Haldane's rule. Genetics *140*, 389–402. https://doi.org/10.1093/genetics/140.1.389.

Turner, L.M., and Harr, B. (2014). Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. Elife *3*, e02504. .

Turner, L.M., Schwahn, D.J., and Harr, B. (2012). Reduced male fertility is common but highly variable in form and severity in a natural house mouse hybrid zone. Evolution *66*, 443–458. https://doi.org/10.1111/j.1558-5646.2011.01445.x.

Turner, L.M., White, M.A., Tautz, D., and Payseur, B.A. (2014). Genomic networks of hybrid sterility. PLoS Genetics *10*, e1004162. https://doi.org/10.1371/journal.pgen.1004162.

Ungerer, M.C., Baird, S.J.E., Pan, J., and Rieseberg, L.H. (1998). Rapid hybrid speciation in wild sunflowers. PNAS *95*, 11757–11762. https://doi.org/10.1073/pnas.95.20.11757.

Ursin, E. (1952). Occurrence of voles, mice, and rats (Muridae) in Denmark, with a special note on a zone of intergradation between two subspecies of the house mouse (Mus musculus L.). Vid. Medd. Dansk. Naturhist. Foren *114*, 217–244. .

Vara, C., Capilla, L., Ferretti, L., Ledda, A., Sánchez-Guillén, R.A., Gabriel, S.I., Albert-Lizandra, G., Florit-Sabater, B., Bello-Rodríguez, J., Ventura, J., et al. (2019). PRDM9 Diversity at Fine Geographical Scale Reveals Contrasting Evolutionary Patterns and Functional Constraints in Natural Populations of House Mice. Molecular Biology and Evolution *36*, 1686–1700. https://doi.org/10.1093/molbev/msz091.

Wang, D., Wang, Z., Kang, X., and Zhang, J. (2019). Genetic analysis of admixture and hybrid patterns of Populus hopeiensis and P. tomentosa. Scientific Reports *9*. https://doi.org/10.1038/s41598-019-41320-z.

Wang, H., Vieira, F.G., Crawford, J.E., Chu, C., and Nielsen, R. (2017). Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. Genome Res. *27*, 1029–1038. https://doi.org/10.1101/gr.204800.116.

Wang, L., Luzynski, K., Pool, J.E., Janoušek, V., Dufková, P., Vyskočilová, M.M., Teeter, K.C., Nachman, M.W., Munclinger, P., Macholán, M., et al. (2011). Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. Mol. Ecol. *20*, 2985–3000. https://doi.org/10.1111/j.1365-294X.2011.05148.x.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics *182*, 1207–1218. https://doi.org/10.1534/genetics.109.102509.

Wegmann, D., Leuenberger, C., Neuenschwander, S., and Excoffier, L. (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics *11*, 116. https://doi.org/10.1186/1471-2105-11-116.

Wegmann, D., Kessner, D.E., Veeramah, K.R., Mathias, R.A., Nicolae, D.L., Yanek, L.R., Sun, Y.V., Torgerson, D.G., Rafaels, N., Mosley, T., et al. (2011). Recombination rates in admixed individuals identified by ancestry-based inference. Nat Genet *43*, 847–853. https://doi.org/10.1038/ng.894.

White, M.A., Steffy, B., Wiltshire, T., and Payseur, B.A. (2011). Genetic dissection of a key reproductive barrier between nascent species of house mice. Genetics *189*, 289–304. https://doi.org/10.1534/genetics.111.129171.

Whitlock, M.C., and Barton, N.H. (1997). The effective size of a subdivided population. Genetics *146*, 427–441. .

Wielstra, B. (2019). Historical hybrid zone movement: More pervasive than appreciated. Journal of Biogeography *46*, 1300–1305. https://doi.org/10.1111/jbi.13600.

Wright, S. (1931). Evolution in Mendelian populations. Genetics *16*, 97–159. .

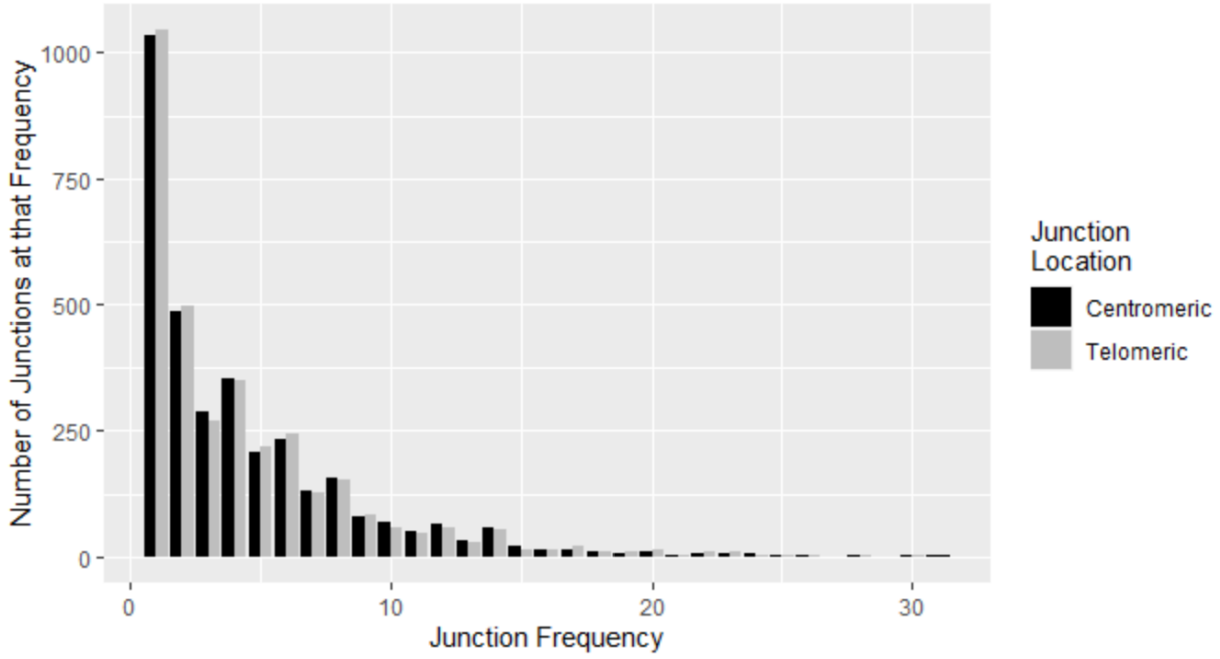
Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A.J., Deloukas, P., Olsen, A., Doggett, N.A., Ghebranious, N., Broman, K.W., et al. (2001). Comparison of human genetic and sequence-based physical maps. Nature *409*, 951–953. https://doi.org/10.1038/35057185.

Zimmerman, K. (1949). Zur Kenntnis der mitteleuropäischen Hausmäuse. Zool. Jb. Syst. *78*, 301–322. .

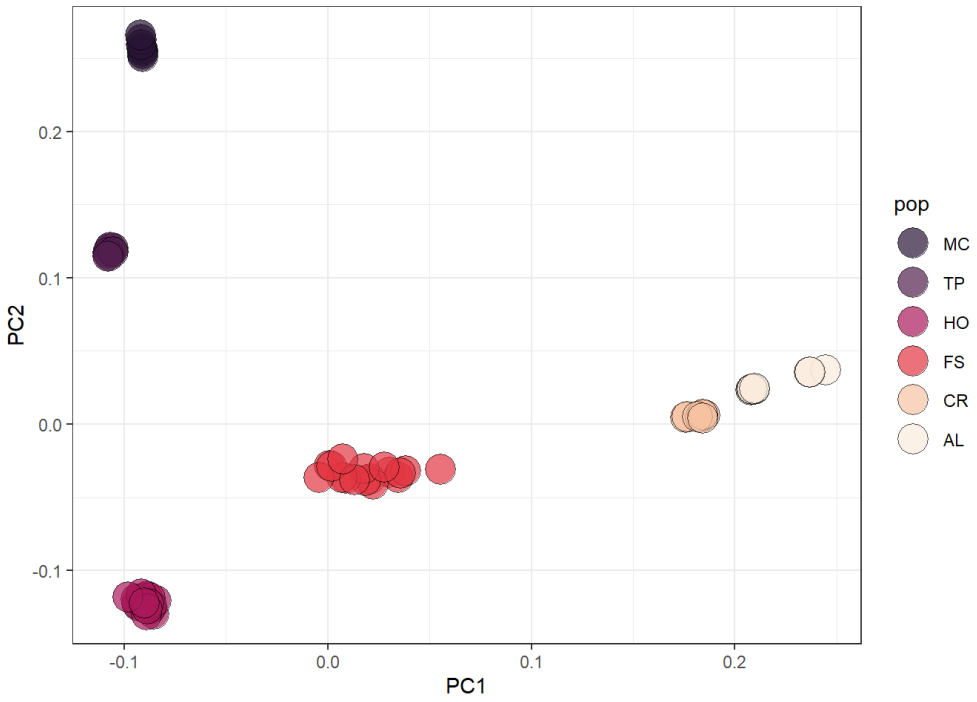
Appendix A

Supplementary Material for Chapter 4

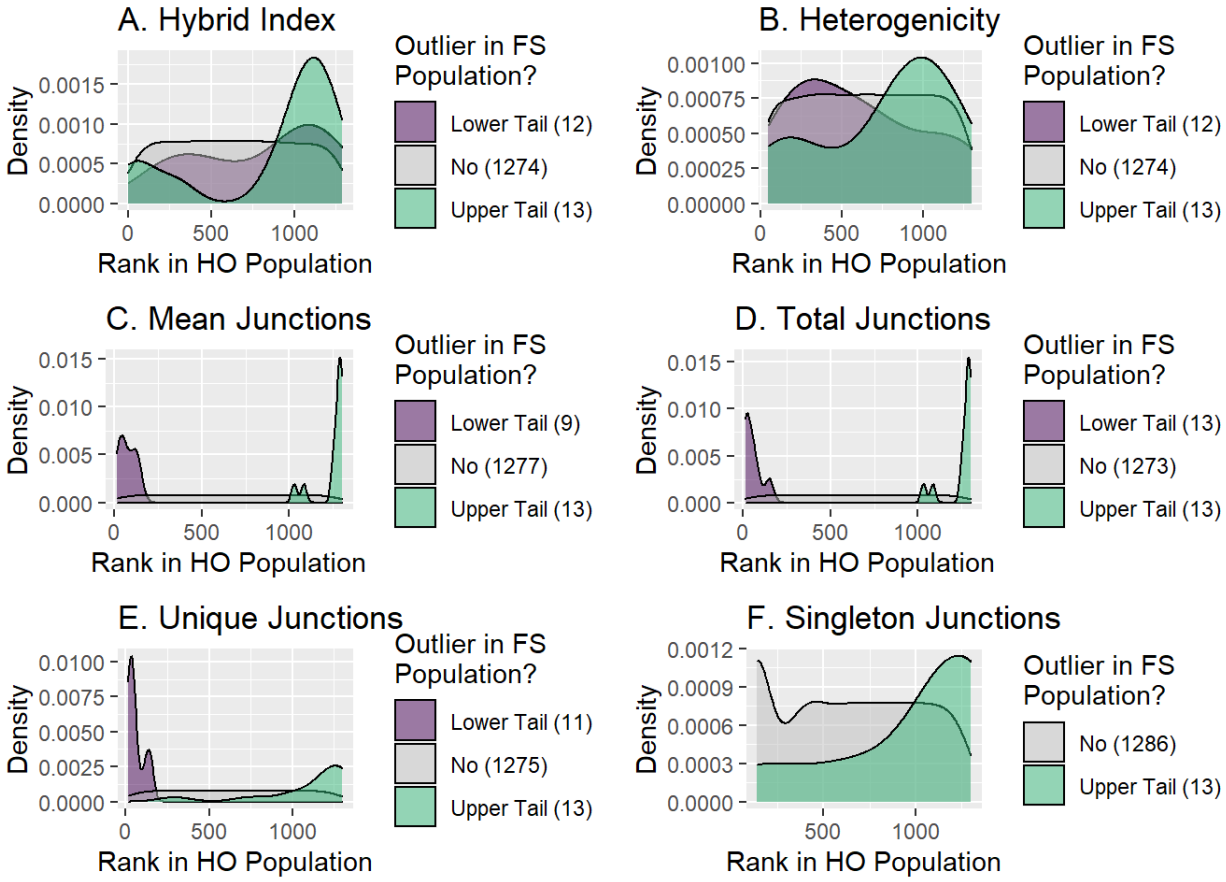
Supplemental Figure 4.1. Comparison of junction location assignment. Junction frequency spectra for chromosome 19 of the FS population are shown below when junctions are assigned to the telomeric-most probable location versus the centromeric-most probable location.



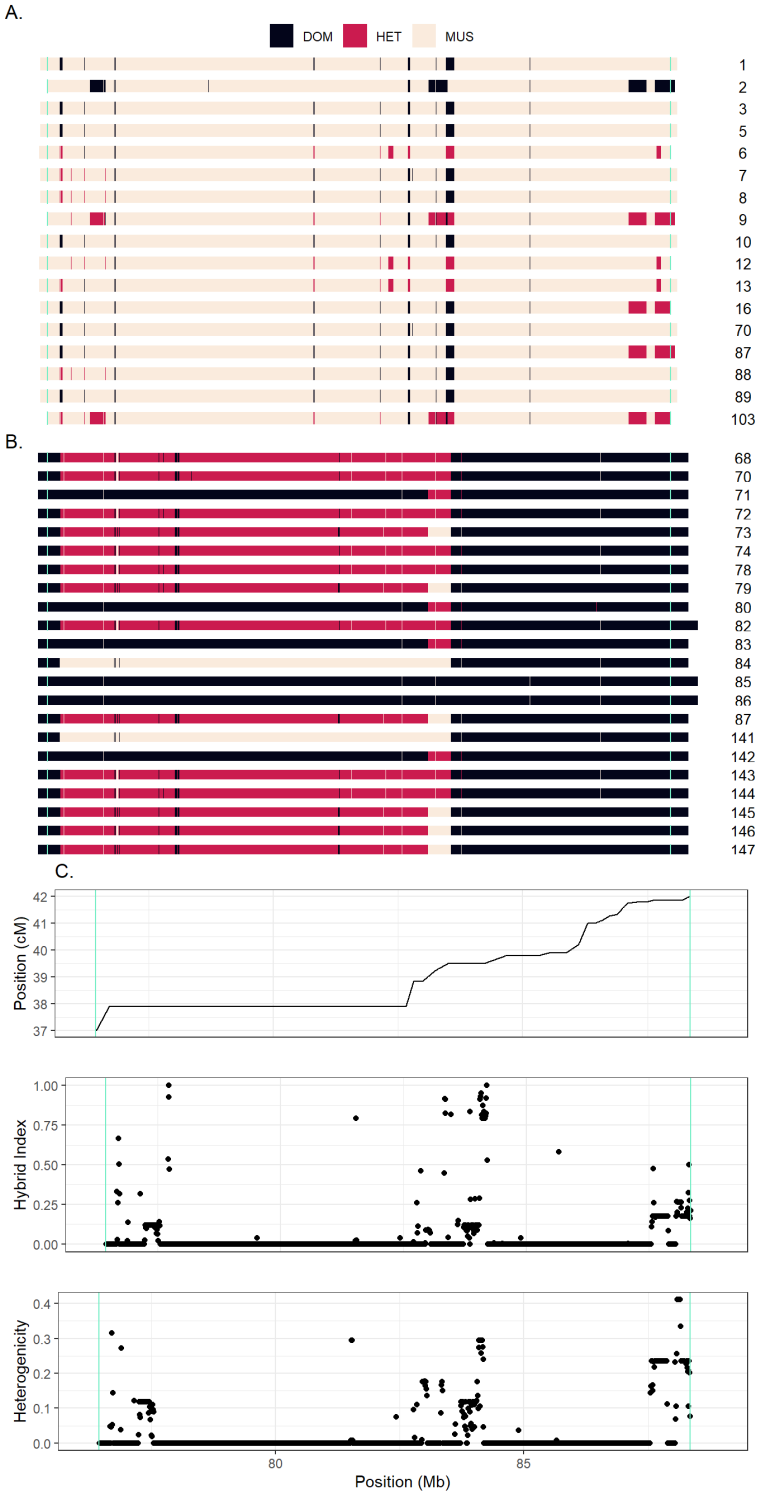
Supplemental Figure 4.2. PC1 versus PC2 for all six populations.



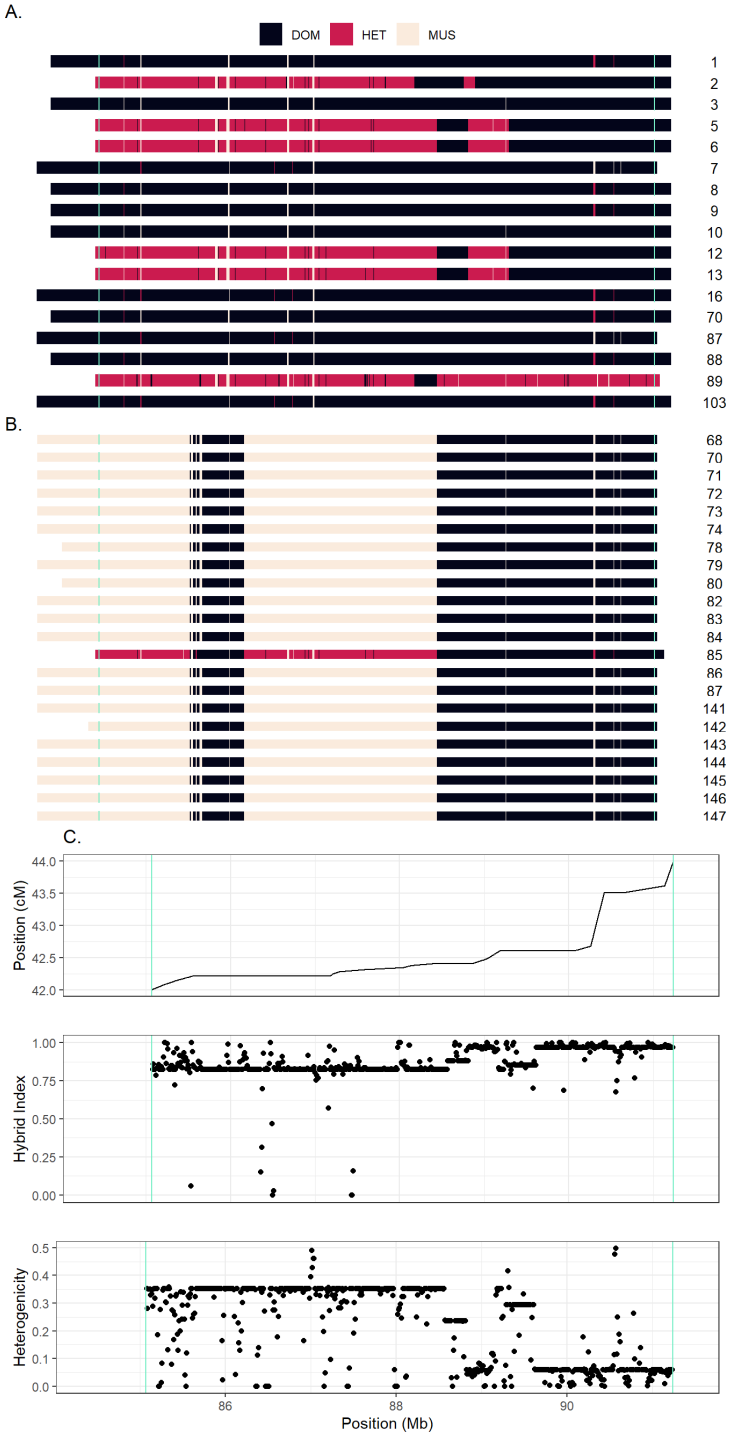
Supplemental Figure 4.3. HO ranks of FS outlier windows. The number of windows represented by each distribution are noted in the legend.



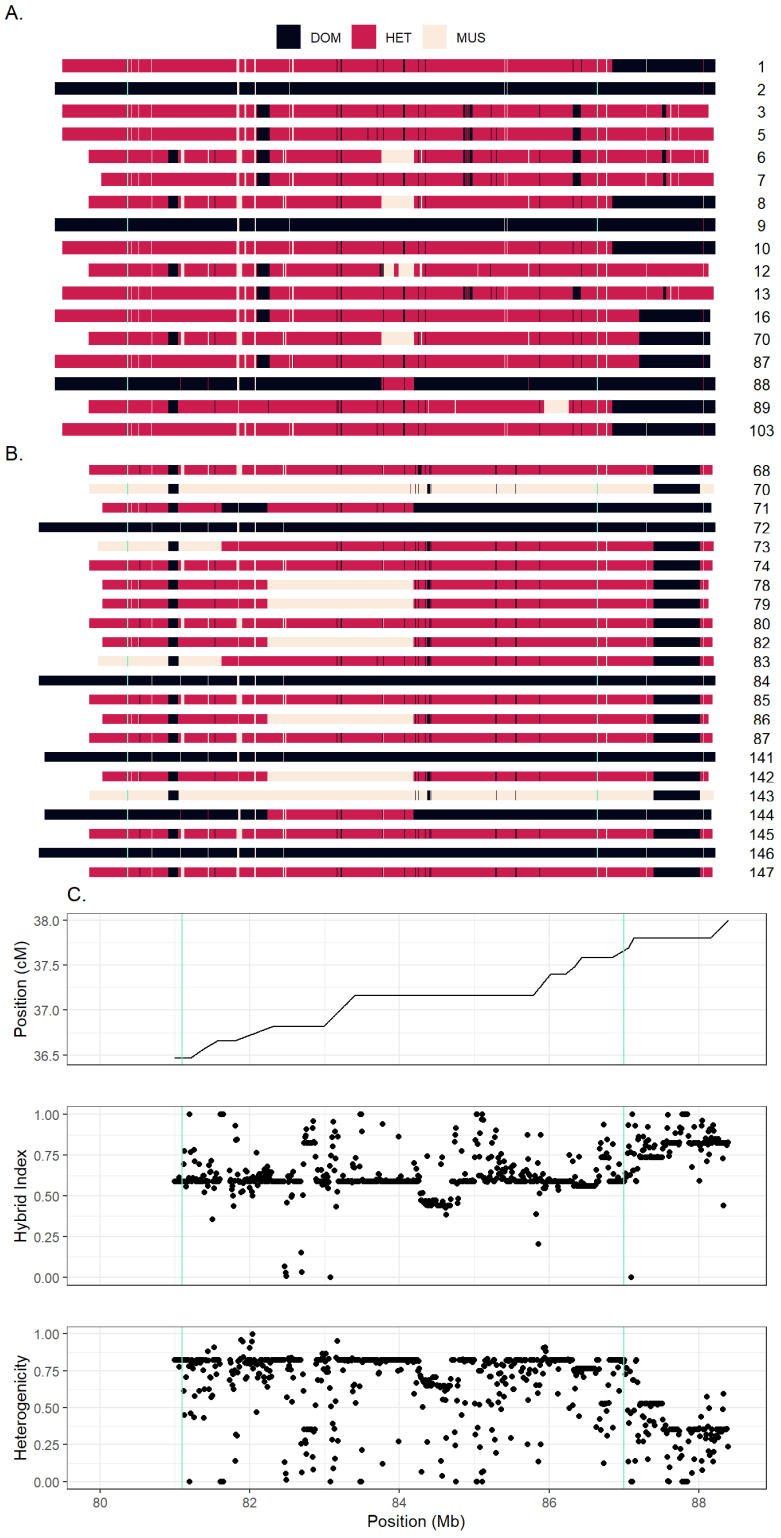
Supplemental Figure 4.4. Region of interest on chromosome 10. Green lines indicate the edges of the region. A) Ancestry across the region in the FS population. B) Ancestry across the region in the HO population. C) Recombination rate shown as physical versus genetic position, and hybrid index and heterogenicity shown in 10kb windows for the FS population (in which the window was identified).



Supplemental Figure 4.5. Region of interest on chromosome 13. Green lines indicate the edges of the region. A) Ancestry across the region in the FS population. B) Ancestry across the region in the HO population. C) Recombination rate shown as physical versus genetic position, and hybrid index and heterogenicity shown in 10kb windows for the FS population (in which the window was identified).



Supplemental Figure 4.6. Region of interest on chromosome 14. Green lines indicate the edges of the region. A) Ancestry across the region in the FS population. B) Ancestry across the region in the HO population. C) Recombination rate shown as physical versus genetic position, and hybrid index and heterogenicity shown in 10kb windows for the FS population (in which the window was identified).



Supplemental Table 4.1. Hybrid genome sequences.

Individual	Population	Batch	Read Count in Run 1	Read Count in Run 2	Average Coverage	New DNA Extraction?
LMT068	HO	1	491041180	0	52.61	N
LMT070	HO	1	290465565	0	31.12	Y
LMT071	HO	1	278321943	0	29.82	N
LMT072	HO	1	369295007	0	39.57	N
LMT073	HO	1	490212156	0	52.52	Y
LMT074	HO	1	309384382	0	33.15	N
LMT078	HO	1	519405318	0	55.65	Y
LMT079	HO	1	378011665	0	40.50	N
LMT080	HO	1	412321462	0	44.18	N
LMT082	HO	1	352783797	0	37.80	N
LMT083	HO	1	305778412	0	32.76	N
LMT084	HO	1	308785164	0	33.08	N
LMT085	HO	1	413593158	0	44.31	N
LMT086	HO	1	315175378	0	33.77	N
LMT087	HO	1	538210158	0	57.67	N
LMT141	HO	1	294932117	0	31.60	N
LMT142	HO	1	522253463	0	55.96	Y
LMT143	HO	1	493374124	0	52.86	N
LMT144	HO	1	426086290	0	45.65	Y
LMT145	HO	1	407931294	0	43.71	Y
LMT146	HO	2	164074365	263896170	45.85	N
LMT147	HO	2	102671848	366706409	50.29	N
NH001	FS	2	95240504	347684067	47.46	N
NH002	FS	2	138439466	317874185	48.89	N
NH003	FS	2	153862247	301196970	48.76	Y
NH005	FS	2	84660318	200987760	30.61	N
NH006	FS	2	176737401	233578140	43.96	N
NH007	FS	2	119936466	306892467	45.73	N
NH008	FS	2	141469642	265150451	43.57	N
NH009	FS	2	112201149	365338935	51.17	N
NH010	FS	2	110959082	339991916	48.32	N
NH012	FS	2	103658952	380771840	51.90	N
NH013	FS	2	144740425	308740422	48.59	N
NH016	FS	2	107083942	401526893	54.49	N
NH070	FS	2	168069551	219117541	41.48	N
NH087	FS	2	177256764	271218775	48.05	N
NH088	FS	2	170331969	235045894	43.43	Y
NH089	FS	2	145333207	288949825	46.53	Y
NH103	FS	2	134449057	290049861	45.48	Y

Appendix B

Ancestry Maps for Each Chromosome

Figure 1. Genomic ancestry patterns on chromosome 1 for each mouse.

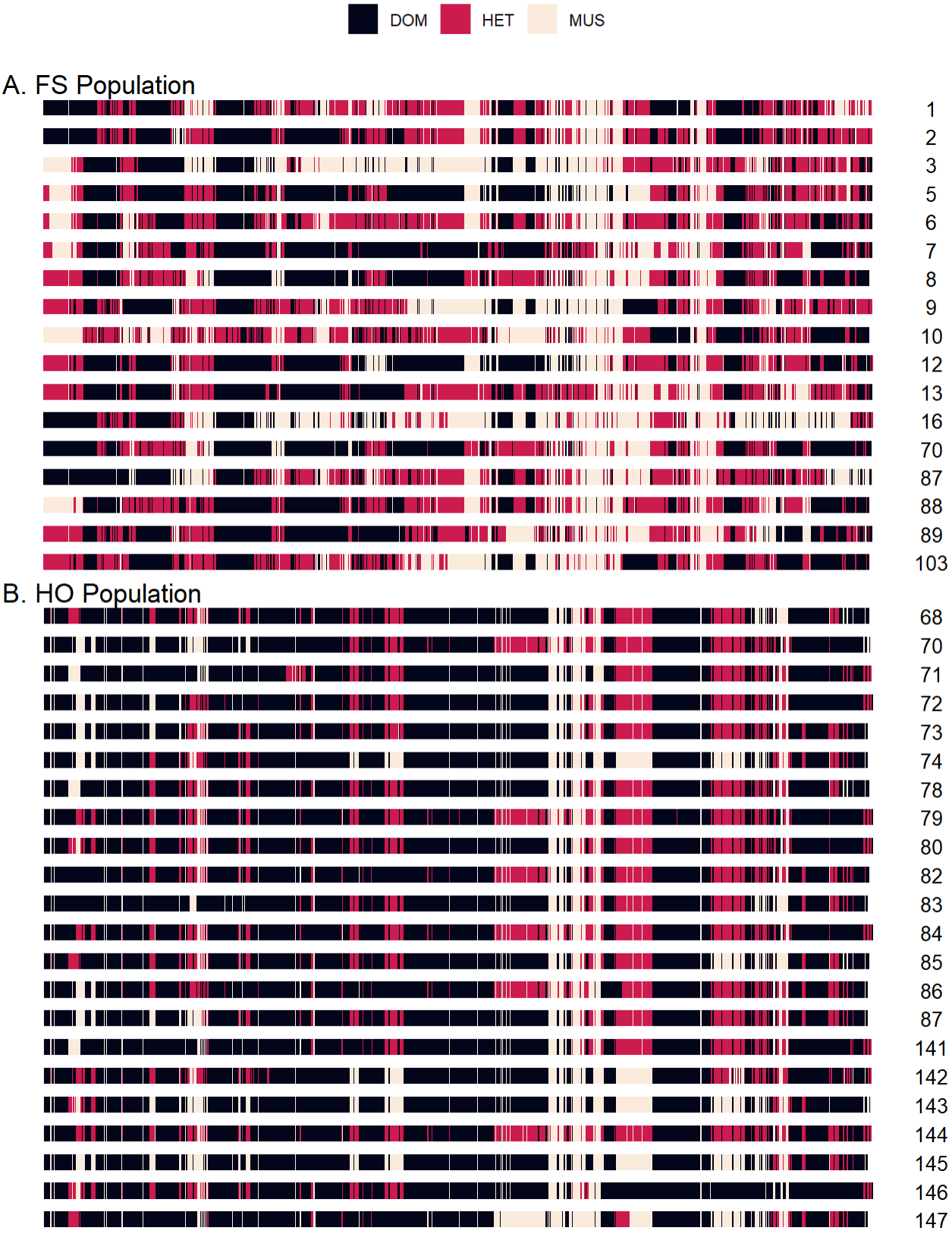


Figure 2. Genomic ancestry patterns on chromosome 2 for each mouse.

Figure 3. Genomic ancestry patterns on chromosome 3 for each mouse.

Figure 4. Genomic ancestry patterns on chromosome 4 for each mouse.

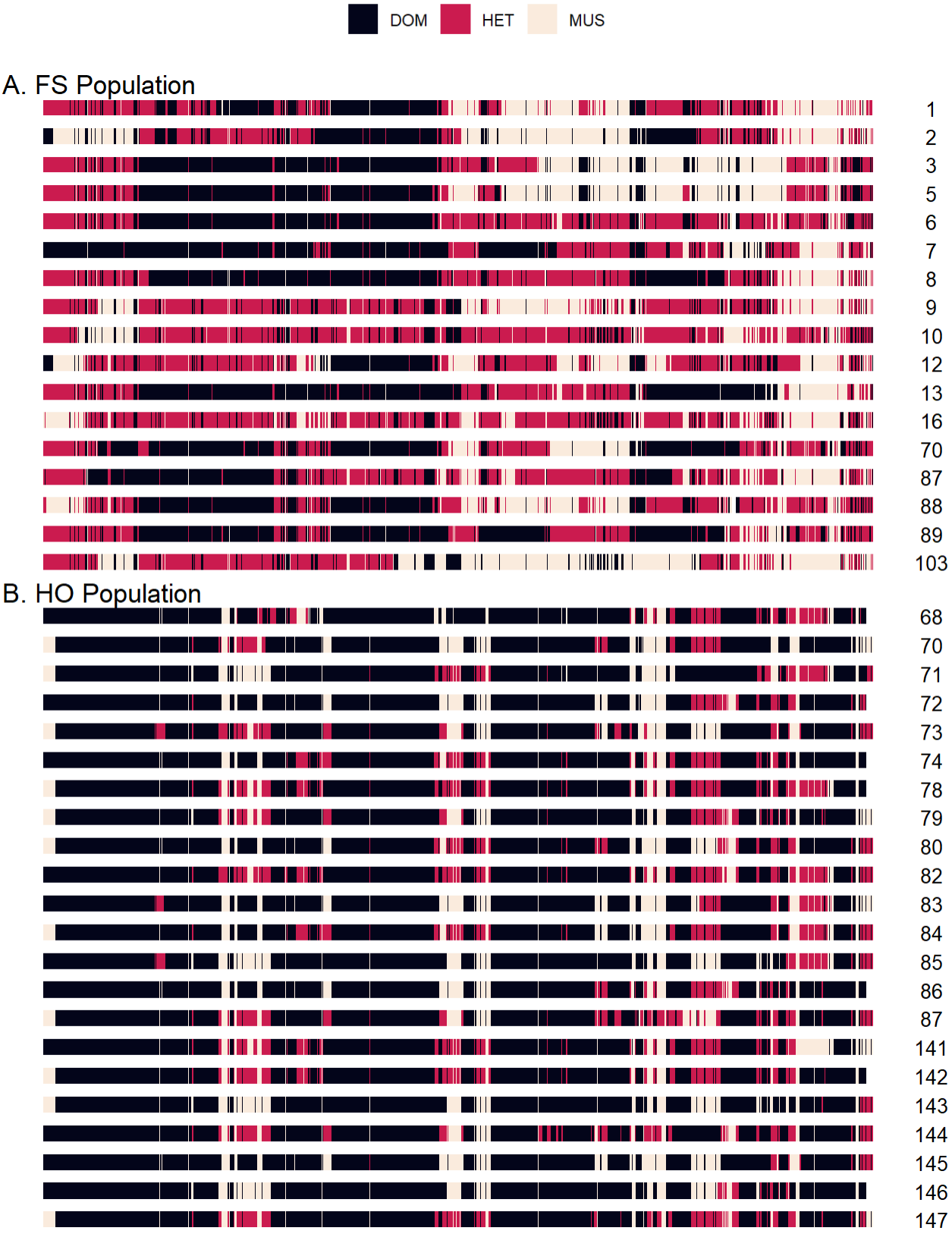


Figure 5. Genomic ancestry patterns on chromosome 5 for each mouse.

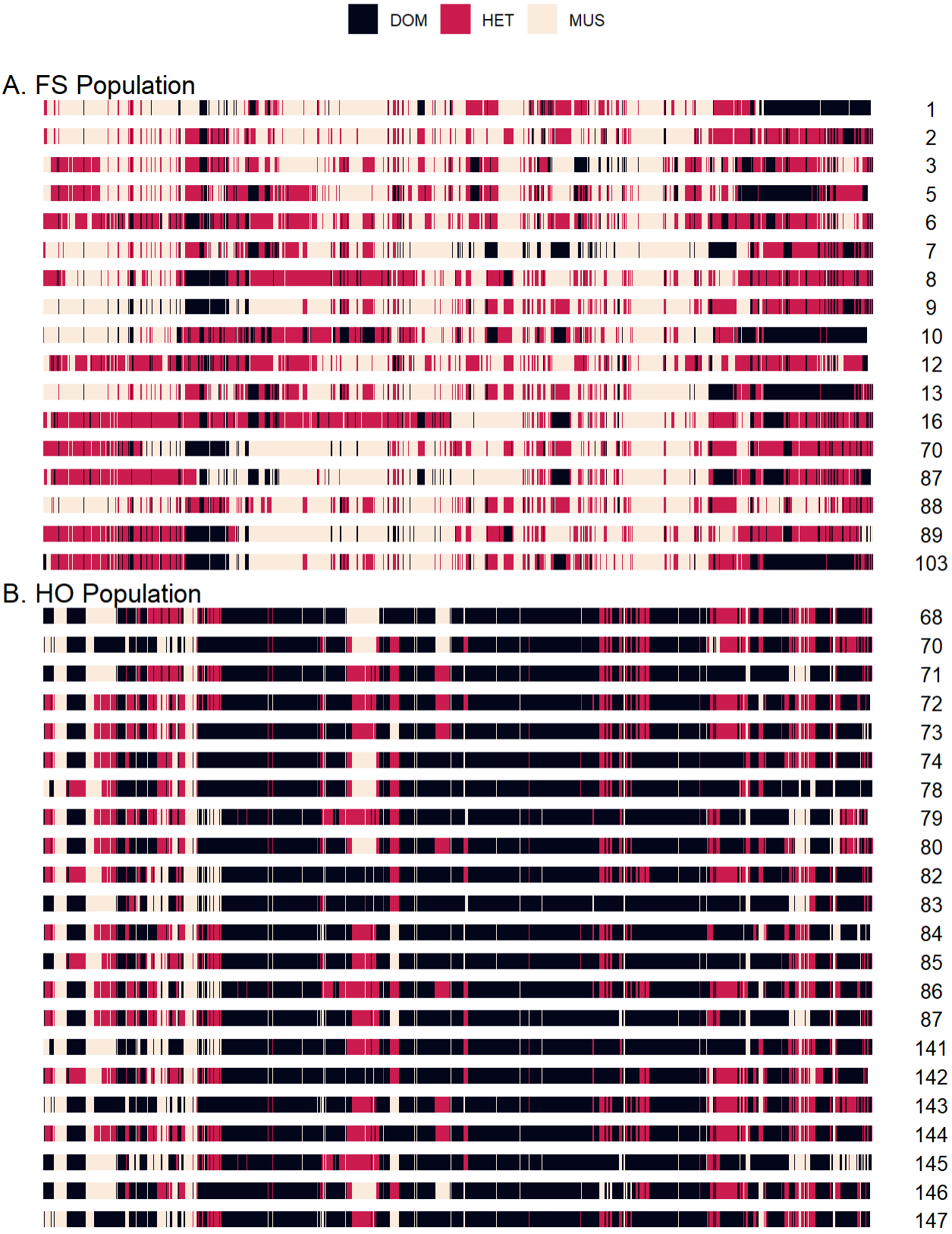


Figure 6. Genomic ancestry patterns on chromosome 6 for each mouse.

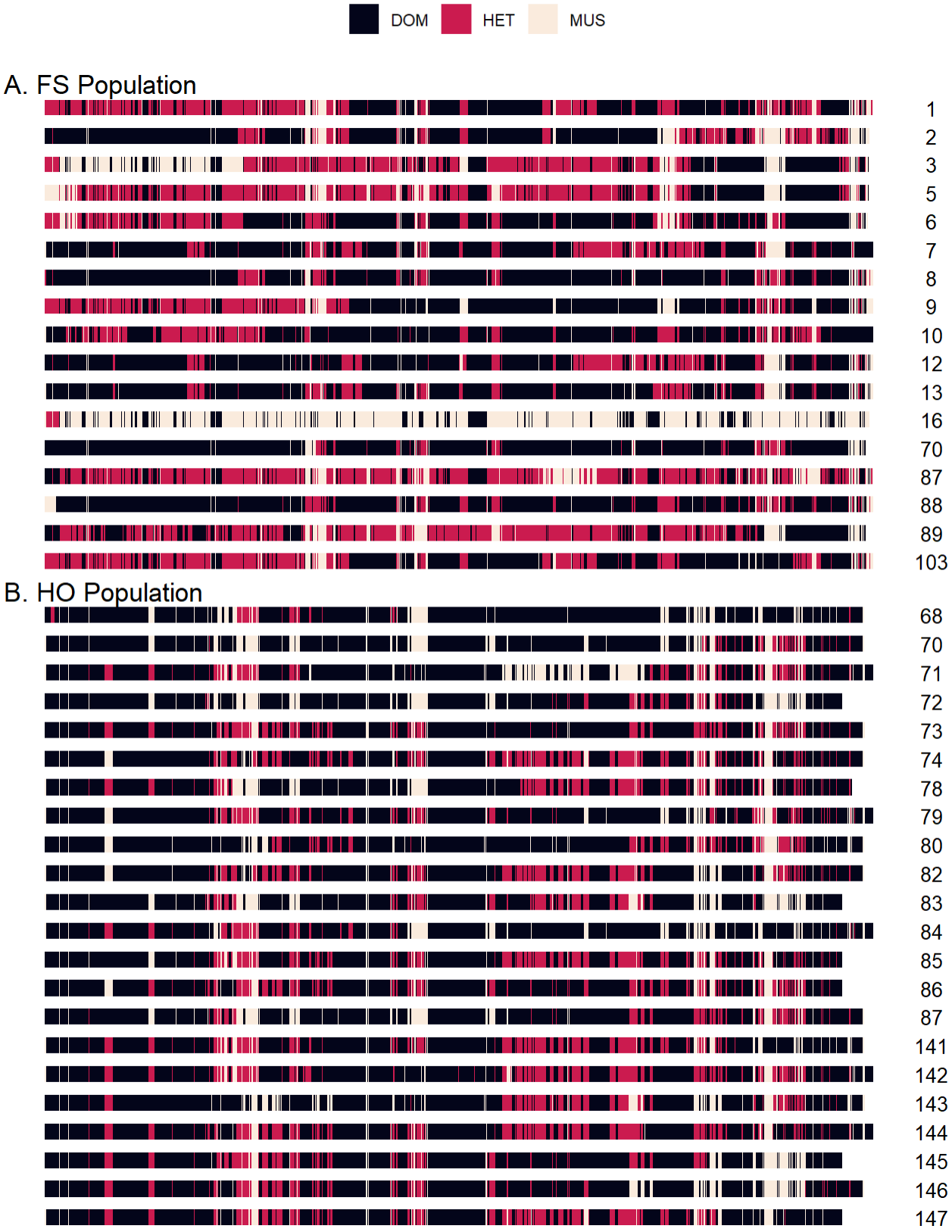


Figure 7. Genomic ancestry patterns on chromosome 7 for each mouse.

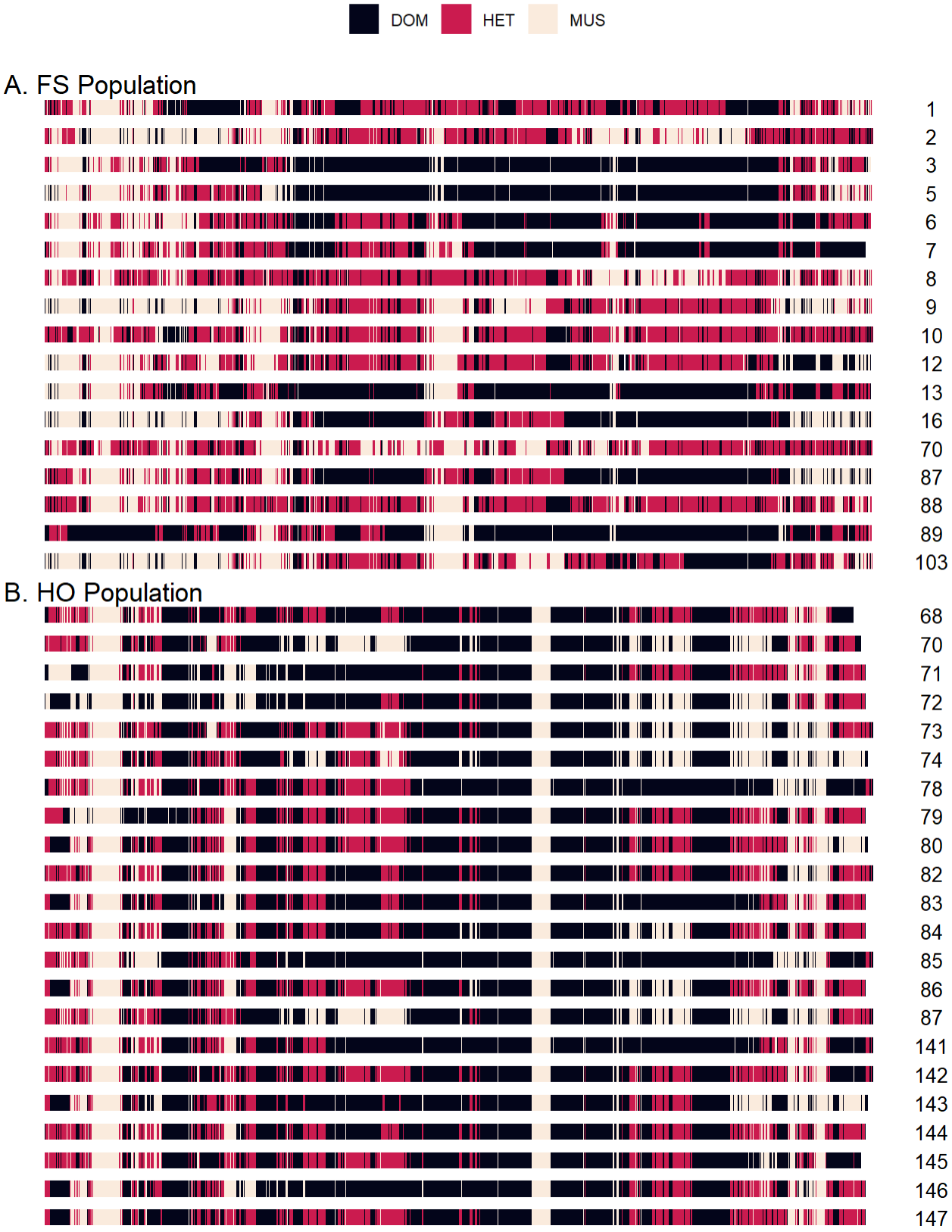


Figure 8. Genomic ancestry patterns on chromosome 8 for each mouse.

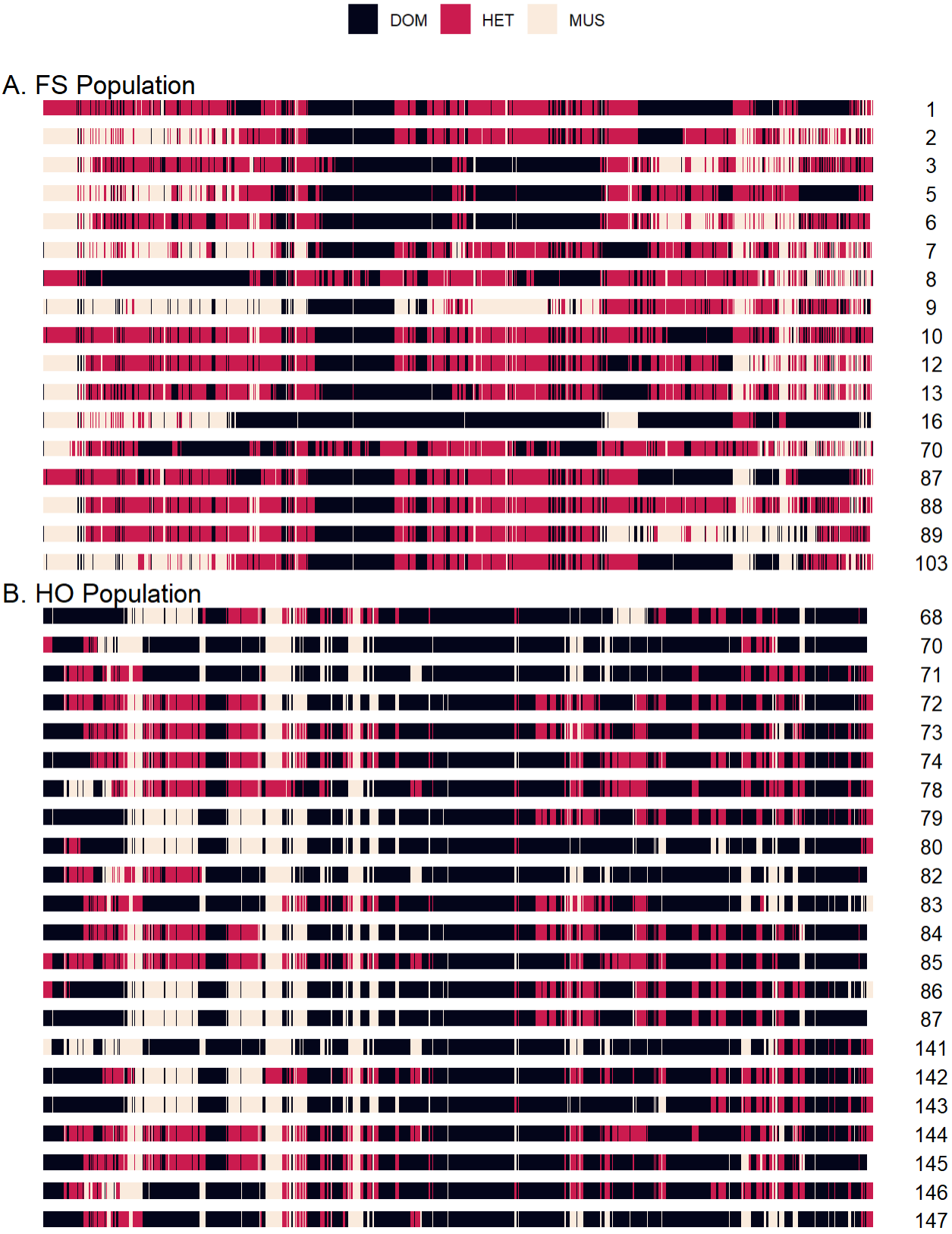


Figure 9. Genomic ancestry patterns on chromosome 9 for each mouse.

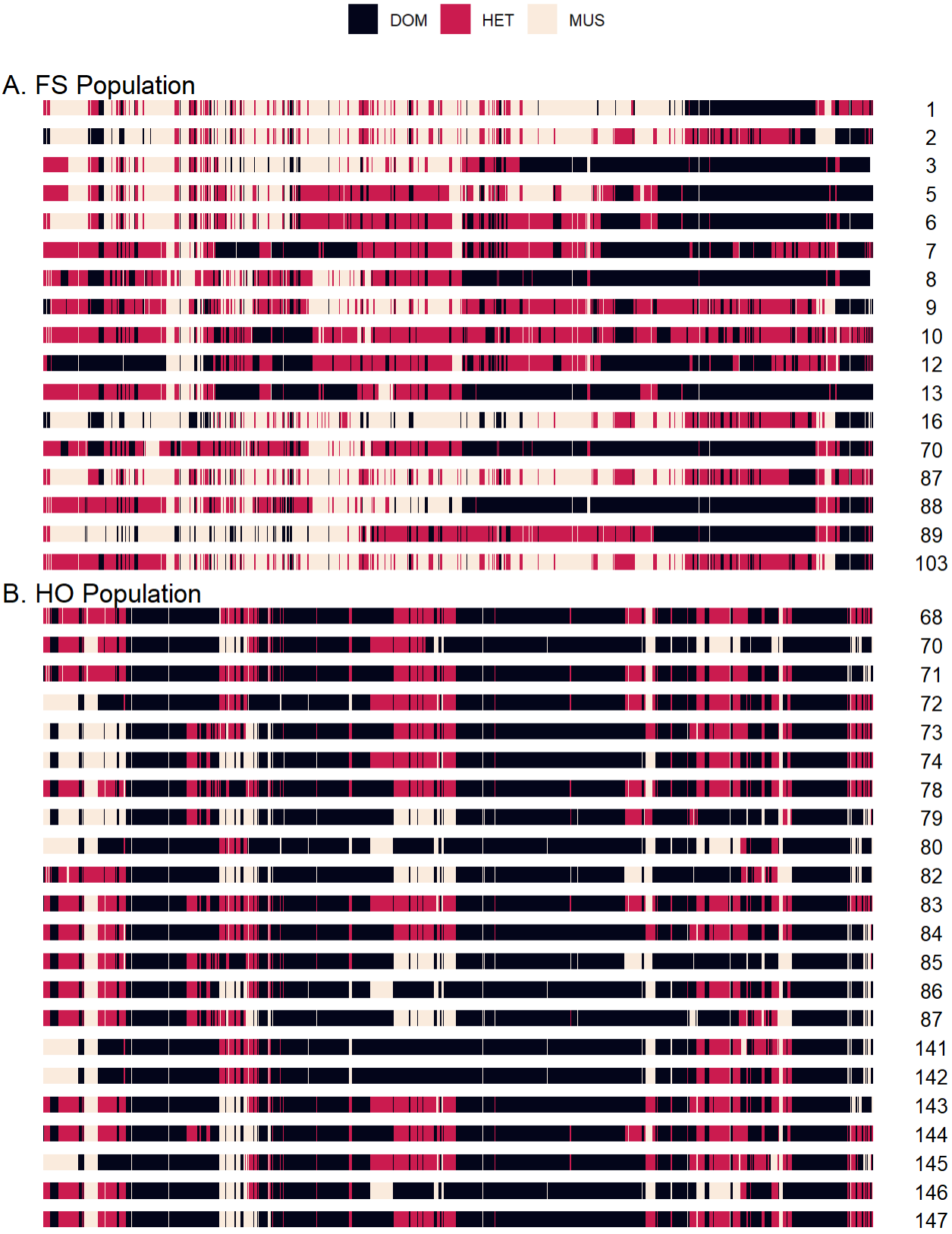


Figure 10. Genomic ancestry patterns on chromosome 10 for each mouse.

Figure 11. Genomic ancestry patterns on chromosome 11 for each mouse.

Figure 12. Genomic ancestry patterns on chromosome 12 for each mouse.

Figure 13. Genomic ancestry patterns on chromosome 13 for each mouse.

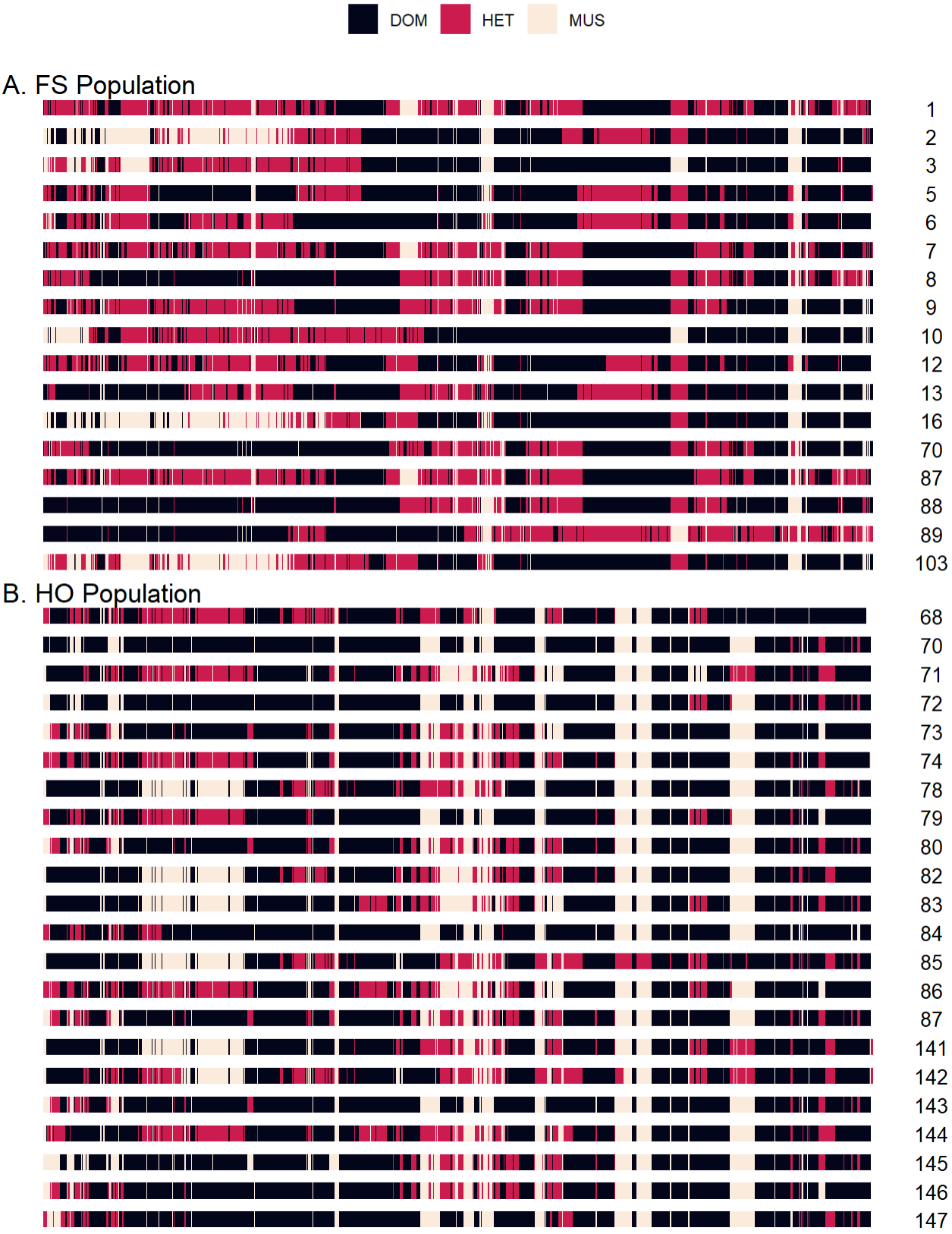


Figure 14. Genomic ancestry patterns on chromosome 14 for each mouse.

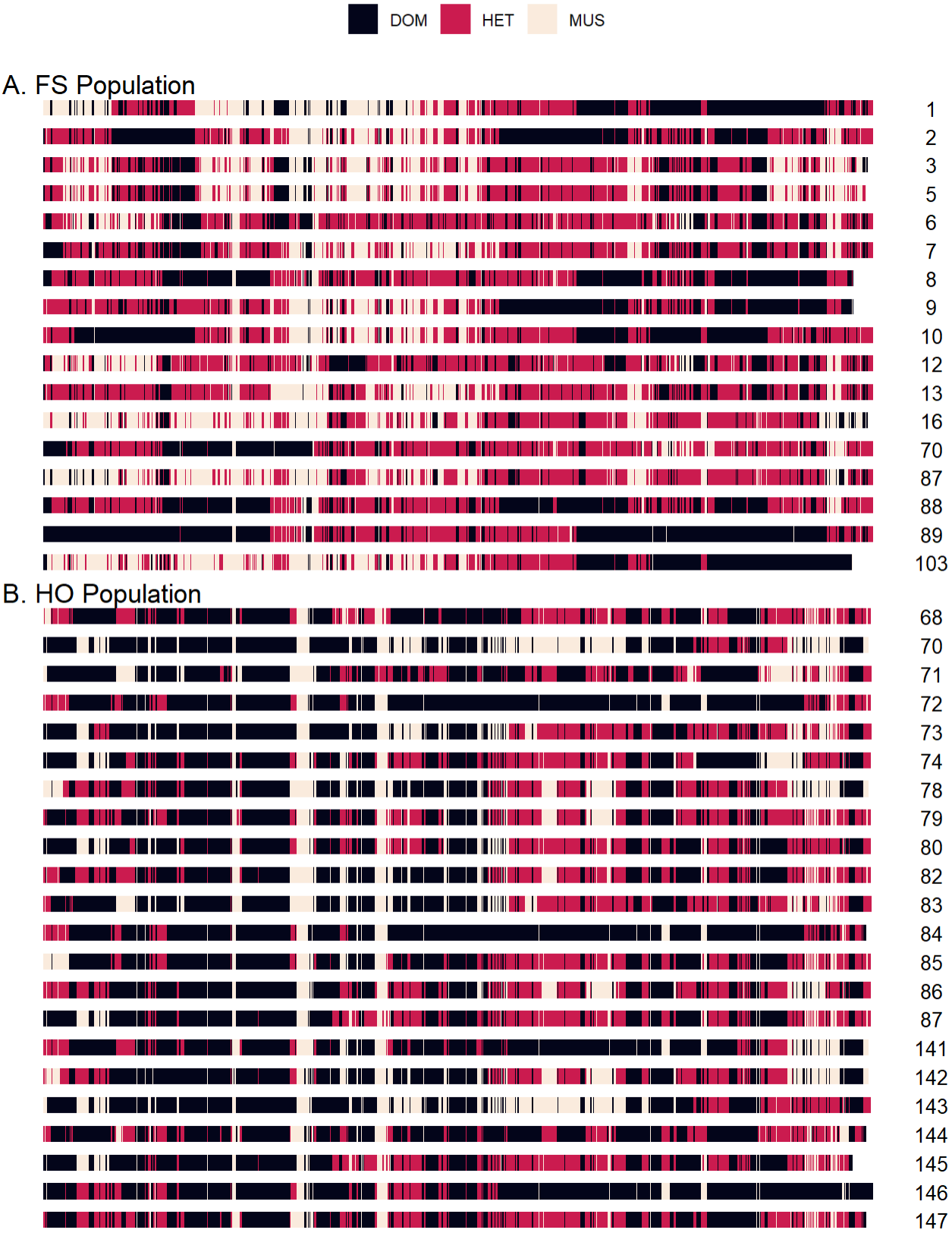


Figure 15. Genomic ancestry patterns on chromosome 15 for each mouse.

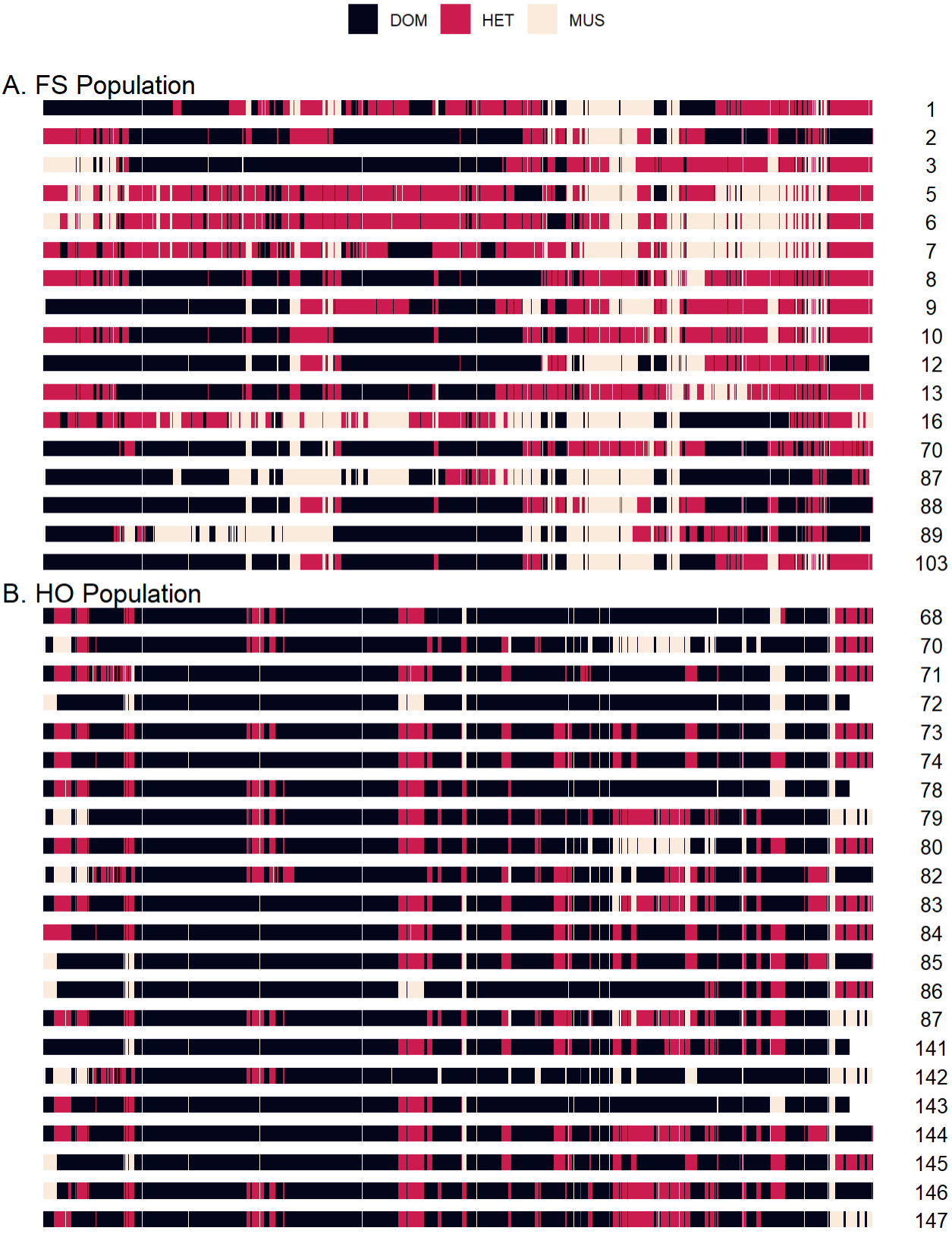


Figure 16. Genomic ancestry patterns on chromosome 16 for each mouse.

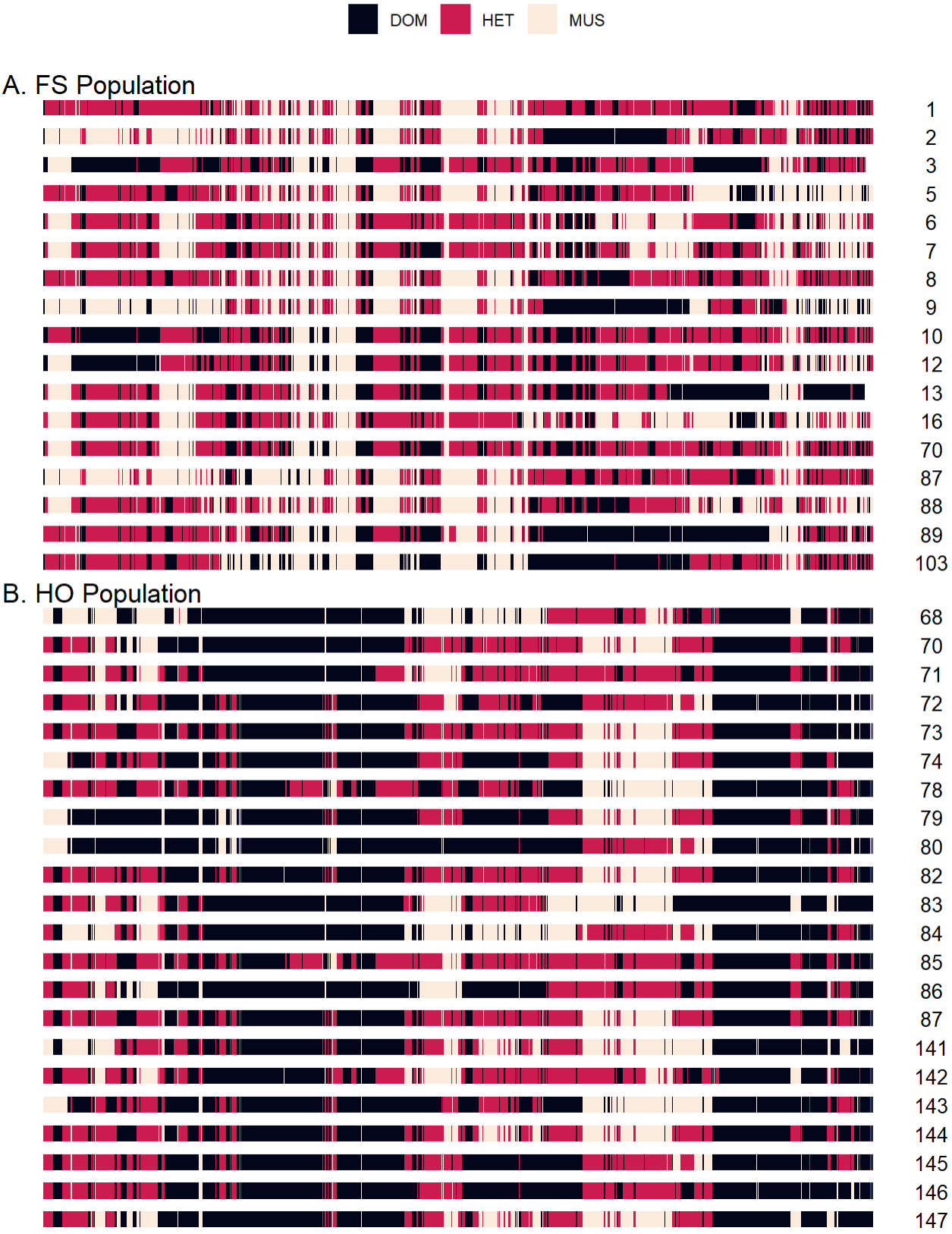


Figure 17. Genomic ancestry patterns on chromosome 17 for each mouse.

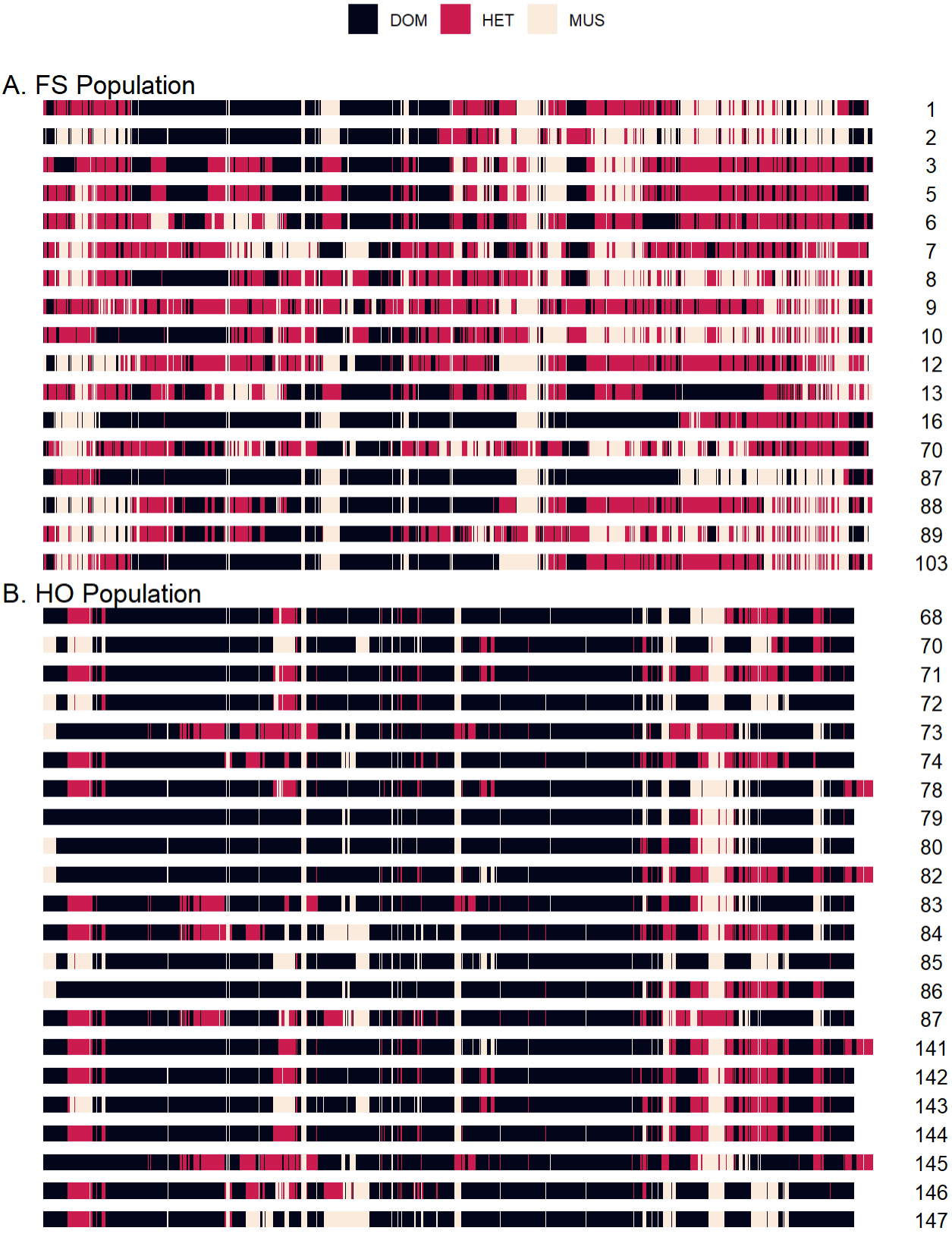


Figure 18. Genomic ancestry patterns on chromosome 18 for each mouse.

Figure 19. Genomic ancestry patterns on chromosome 19 for each mouse.

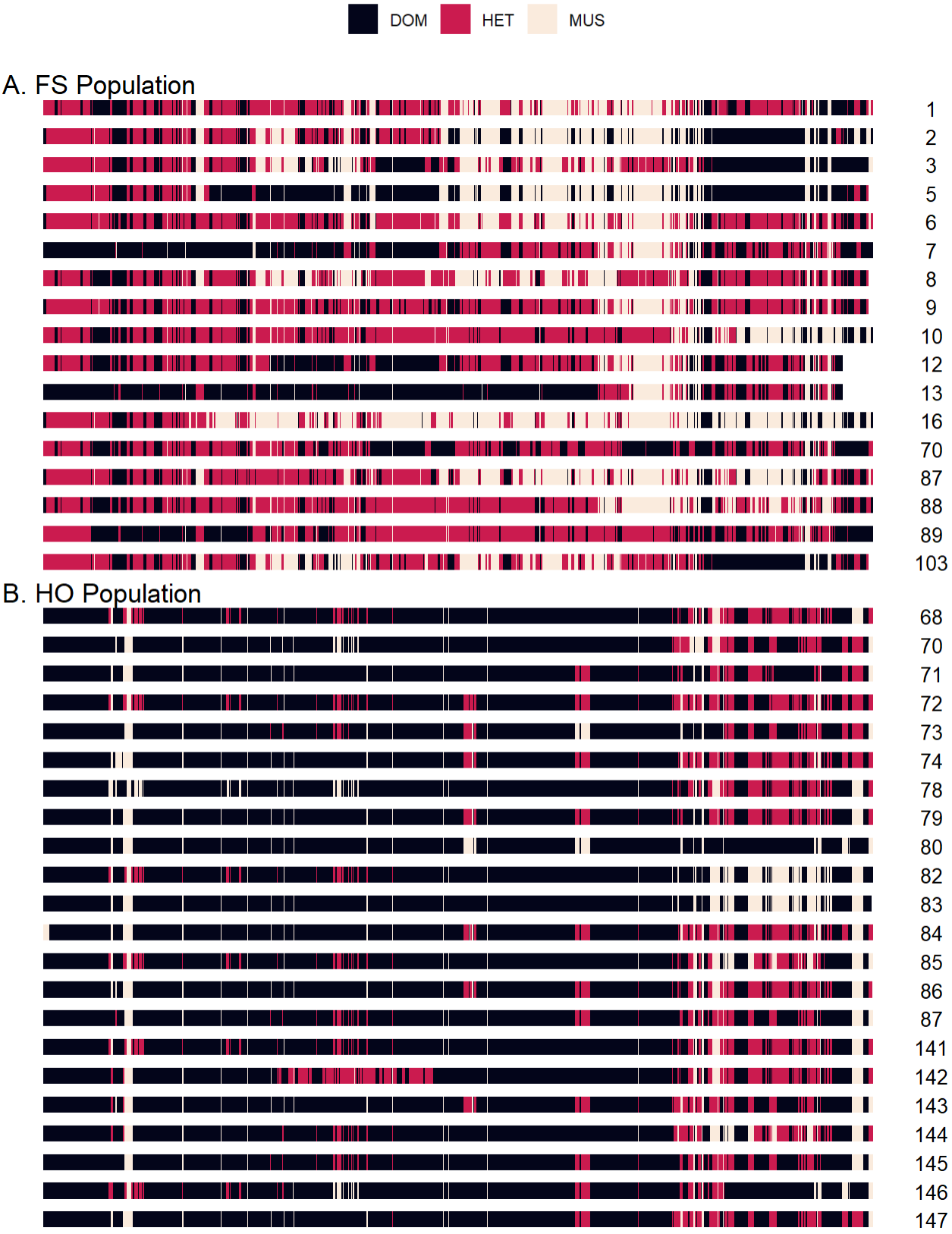


Figure 20. Genomic ancestry patterns on the X chromosome for each mouse.