Genome-Wide Patterns of Differentiation Among House Mouse Subspecies

Megan Phifer-Rixey,*[,1] Matthew Bomhoff,[†] and Michael W. Nachman*
*Museum of Vertebrate Zoology and Department of Integrative Biology, University of California, Berkeley, California 94720, and
[†]BIO5 Institute, University of Arizona, Tucson, Arizona 85721

ABSTRACT One approach to understanding the genetic basis of speciation is to scan the genomes of recently diverged taxa to identify highly differentiated regions. The house mouse, *Mus musculus*, provides a useful system for the study of speciation. Three subspecies (*M. m. castaneus*, *M. m. domesticus*, and *M. m. musculus*) diverged ~350 KYA, are distributed parapatrically, show varying degrees of reproductive isolation in laboratory crosses, and hybridize in nature. We sequenced the testes transcriptomes of multiple wild-derived inbred lines from each subspecies to identify highly differentiated regions of the genome, to identify genes showing high expression divergence, and to compare patterns of differentiation among subspecies that have different demographic histories and exhibit different levels of reproductive isolation. Using a sliding-window approach, we found many genomic regions with high levels of sequence differentiation in each of the pairwise comparisons among subspecies. In all comparisons, the X chromosome was more highly differentiated than the autosomes. Sequence differentiation and expression divergence were greater in the *M. m. domesticus*–*M. m. musculus* comparison than in either pairwise comparison with *M. m. castaneus*, which is consistent with laboratory crosses that show the greatest reproductive isolation between *M. m. domesticus* and *M. m. musculus*. Coalescent simulations suggest that differences in estimates of effective population size can account for many of the observed patterns. However, there was an excess of highly differentiated regions relative to simulated distributions under a wide range of demographic scenarios. Overlap of some highly differentiated regions with previous results from QTL mapping and hybrid zone studies points to promising candidate regions for reproductive isolation.

UNDERSTANDING the genetic basis of speciation is a fundamental goal of evolutionary biology. This problem has primarily been approached in two ways: through laboratory studies using crosses and through studies of genetic variation in natural populations. Laboratory studies control for genetic background and environment, and they make it possible to connect genotype and phenotype. These types of studies have produced some spectacular successes including the identification of individual genes underlying postzygotic isolation in *Drosophila* (*e.g.*, Ting *et al.* 1998; Presgraves *et al.* 2003; Brideau *et al.* 2006; Masly *et al.* 2006), *Arabidopsis* (Bomblies *et al.* 2007; Bikard *et al.* 2009), *Mus* (Mihola *et al.* 2009), and others (reviewed in Presgraves 2010 and Nosil and Schluter 2011).

Studies of natural populations rely on the idea that regions of the genome that are important in reproductive isolation may be more differentiated than other regions of the genome. Therefore, by studying patterns of differentiation, one may gain insight into the genomic regions that underlie isolation. The idea that the genomes of closely related species are mosaics of differentiated and less differentiated regions is not new and first emerged in the literature on hybrid zones (*e.g.*, Key 1968; Harrison 1986; Tucker *et al.* 1992; Rieseberg *et al.* 1999; reviewed in Harrison 2012). The advent of genomic methods has fueled a renewed interest in studying patterns of differentiation between closely related species, including work on mosquitoes (Turner *et al.* 2005; Lawniczak *et al.* 2010; Neafsey *et al.* 2010), mice (Harr 2006), *Drosophila* (Kulathinal *et al.* 2009), *Heliconius* butterflies (Nadeau *et al.* 2012), flycatchers (Ellegren *et al.* 2012), crickets (Andrés *et al.* 2013), sunflowers (Renaut *et al.* 2013), and others.

doi: 10.1534/genetics.114.166827
Manuscript received June 3, 2014; accepted for publication June 24, 2014; published Early Online July 3, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.166827/-/DC1.
[1] Corresponding author: The Museum of Vertebrate Zoology, University of California, 3101 Valley Life Sciences Bldg., Berkeley, CA 94720-3160. E-mail: mrixey@berkeley.edu

Despite their appeal, genome scans present a number of challenges. One is correctly identifying genomic regions that show unexpectedly high levels of differentiation. This has typically been done either by specifying an appropriate null demographic model against which an observed distribution can be compared or by simply identifying extreme values as potential candidate regions. Another challenge is interpreting the biological meaning of a genomic region showing a high level of differentiation. Shared polymorphism can result from retained ancestral variation or from gene flow; conversely, differentiation can result from sorted ancestral variation (due to drift or selection) or from absence of gene flow. Charlesworth (1998) pointed out that reduced variation within a population will inflate estimates of differentiation, such as F_{st}, that are based on both within- and between-population components of variation. As a result, background selection (Charlesworth 1993) and genetic hitchhiking (Maynard Smith and Haigh 1974) may lead to localized high values of F_{st} even for regions that are not involved in reproductive isolation (Cruickshank and Hahn 2014). Therefore, genomic "islands of differentiation" may reflect (1) stochastic variation in lineage sorting; (2) regions of reduced gene flow; (3) regions in which the effects of selection at linked sites are more pronounced, regardless of involvement with reproductive isolation; or (4) some combination of these processes.

House mice provide a valuable system for the study of speciation. *Mus musculus* consists of three subspecies that are distributed parapatrically: *M. m. domesticus* in western Europe, *M. m. musculus* in eastern Europe and northern Asia, and *M. m. castaneus* in southeast Asia. These subspecies are believed to have diverged in allopatry at roughly the same time—~350,000 years ago (Bonhomme *et al.* 2007; Geraldes *et al.* 2008, 2011; White *et al.* 2009)—and come into secondary contact much more recently (*e.g.*, Cucchi *et al.* 2005; Duvaux *et al.* 2011). Each subspecies meets and hybridizes with the other two species where their ranges come into contact (*e.g.*, Tucker *et al.* 1992; Boursot *et al.* 1993; Duvaux *et al.* 2011), although only the hybrid zone between *M. m. musculus* and *M. m. domesticus* is well-studied. Differences in estimated N_{e} among the subspecies provide an opportunity to investigate the effects of demography on patterns of differentiation. While estimates of effective population size (N_{e}) are large for *M. m. castaneus* (200,000–733,000), estimates are smaller for *M. m. domesticus* (58,000–200,000) and *M. m. musculus* (25,000–120,000; Salcedo *et al.* 2007; Geraldes *et al.* 2008, 2011; Halligan *et al.* 2010).

The degree of reproductive isolation differs in pairwise comparisons among house mouse subspecies. While significant reductions of F_1 male fertility are seen in crosses between *M. m. domesticus* and *M. m. musculus* (*e.g.*, Good *et al.* 2008; White *et al.* 2011), significant infertility is not observed until the F_2 in crosses between *M. m. castaneus* and *M. m. domesticus*, and the degree of infertility is not as severe (White *et al.* 2012). There are no published studies documenting reduced fertility in lab crosses between *M. m. castaneus* and *M. m. musculus*. In fact, a cross between *M. m. castaneus* and *M. m. musculus* was used for a recombination mapping study and infertility was not observed (Dumont and Payseur 2011).

In this study, we used short-read sequencing of the testis transcriptomes of wild-derived inbred lines of *M. m. castaneus*, *M. m. domesticus*, and *M. m. musculus* to characterize genome-wide patterns of sequence differentiation in pairwise comparisons between each subspecies. Although this study was primarily designed to investigate sequence differentiation, we also investigated patterns of differential gene expression among the subspecies.

Materials and Methods

Samples

All mice came from wild-derived inbred strains (Supporting Information, Table S1). We sequenced eight lines of *M. m. castaneus*, seven of *M. m. domesticus*, and eight of *M. m. musculus*. Wild-derived laboratory strains of *Mus spretus* and *Mus caroli* were included for use as outgroups (She *et al.* 1990; Suzuki *et al.* 2004; Tucker *et al.* 2005). Mice were killed and testes were dissected under RNAse-free conditions. Testes samples were kindly provided by François Bonhomme, Polly Campbell, Courtney Clayton, Matt Dean, and Annie Orth. Testes were placed in RNAlater at 4° overnight and then transferred to −80° for storage. RNA was extracted from frozen tissue using Quiagen's RNAeasy Plus Mini Kit.

Sequencing

Single-end 76-bp reads were sequenced from the mRNA of each individual on an Illumina GAIIx. For most lines, between 0.80 and 1.68 Gb of sequence was obtained (Table S2). One wild-derived inbred line of *M. musculus* and two wild-derived lines selected from outgroup taxa were sequenced at higher coverage (1.92–3.64 Gb). Reads containing <20 high-quality bases (phred ≥ 20) were removed prior to mapping. Sequence data can be accessed via National Center for Biotechnology Information BioProject PRJNA252743. TopHatv1.2 (Trapnell *et al.* 2009) was used with default settings to map reads to the C57BL/6 reference genome, and only reads that mapped uniquely were retained. Finally, sites with a depth <6× of high-quality sequence (phred ≥ 20) were removed from the analysis. These filters left between 12.01 and 22.70 Mb of sequence per line. Genomic sequence data (>20×) were available from the Wellcome Trust Mouse Genomes Project for two of the lines included in our study (SPRET/EiJ and PWK/PhJ; Yalcin *et al.* 2012), and genomic sequence data were used to augment transcriptome sequencing. We compared genotype calls between our data and the Wellcome Trust data in regions of overlap (Table S3). Although both data sets were obtained via shotgun sequencing, the higher coverage of the

Wellcome Trust data allows for an assessment of the possible risks of sequencing and mapping using our lower-coverage approach. Mismatches were rare, occurring at rates ranging from 1 in ~325,000 to 1 in ~420,000 coding sites (Table S3). In addition, we included data from two lines (CAST/EiJ and WSB/Eij) sequenced only by the Wellcome Trust Mouse Genomes Project (Yalcin *et al.* 2012).

Previous analyses have shown that wild-derived inbred lines can contain large introgressed segments from other subspecies (Yang *et al.* 2011). STRUCTURE analysis (Pritchard *et al.* 2000) was used to test for admixture in the wild-derived inbred lines sequenced in this study. We found that two lines of *M. m. castaneus* and one line of *M. m. musculus* were highly admixed, and we excluded them from the study (Table S1). After excluding these lines, the remaining subspecies formed three distinct groups corresponding to the three subspecies, and each line was assigned with most support to the expected subspecies (Table S4). After removing the admixed lines and including the lines sequenced by the Wellcome Trust, six *M. m. castaneus*, eight *M. m. domesticus*, and seven *M. m. musculus* were used in all analyses.

SAMtools was used with default settings to call bases and all SNPs within and among subspecies were identified using custom PERL scripts (Li *et al.* 2009; File S1). Inbred lines are expected to be homozygous. Observed heterozygosity may reflect true residual heterozygosity in inbred lines or errors in sequencing; distinguishing between these two possibilities is difficult with low-coverage data. When heterozygosity was inferred using SAMtools, the site was masked and not included in further analyses. In addition, indels and sites with more than two segregating alleles were excluded. After filtering, we identified >32,000 SNPs within and among subspecies of *M. m. musculus* from 4705 genes with an average of 6.12 SNPs per gene. This represents ~20% of the genes in the genome.

Measures of sequence differentiation

There are many statistics for measuring differentiation, and these capture different aspects of the data. Here, we calculated F_{st} (Hudson *et al.* 1992), D_{xy} (Nei 1987), and δ, the absolute value of the difference in allele frequencies (see Renaut *et al.* 2010; Gagnaire *et al.* 2012; equations given in File S2) for each SNP for each pairwise comparison: *M. m. castaneus–M. m. domesticus* (hereafter CD), *M. m. castaneus–M. m. musculus* (hereafter CM), and *M. m. domesticus–M. m. musculus* (hereafter DM). To account for unequal sample sizes of the subspecies, we subsampled five lines per subspecies 10,000 times at each site with sufficient data, and the average value of a given statistic was used for all subsequent analyses. Measures of differentiation were highly correlated in our data set (Table 1); thus we chose to use δ for subsequent analyses, although similar results were obtained using other measures. We defined SNPs as highly differentiated if average resampled values of δ per site were ≥ 0.8. In such cases, the two subspecies are one allele or fewer from fixation of alternate nucleotides. We defined highly differentiated as $\delta \geq 0.8$ rather than using an approach based on the distributions of statistics (*e.g.*, the upper 5% of values) because it allowed us to compare among the three pairwise analyses. We then asked how many sites were fixed in a single subspecies but polymorphic in both of the other two subspecies. Finally, we identified all fixed, derived sites in each subspecies using comparisons to the outgroup taxa, *M. spretus* and *M. caroli*.

Sliding windows

To identify genomic regions with groups of sites that are highly differentiated, we performed two kinds of sliding-window analyses. First, sliding-window analyses of δ (100-kb windows with a 25-kb step size) were used to identify regions of the genome that were highly differentiated among subspecies. We defined regions as highly differentiated if average values of δ were ≥ 0.8 across all sites in a window. All SNPs were included in these analyses, and windows were evaluated only when there were three or more SNPs in the window.

Private polymorphisms (*i.e.*, those segregating in just one species) can lower the average level of differentiation in a region as measured by F_{st}, D_{xy}, and δ. The presence of private polymorphisms does not mean that such regions are not potentially relevant to speciation, only that they are less likely, on average, to have experienced recent coalescent events. Analyses that do not distinguish between shared and private polymorphisms may fail to identify many regions that are fully sorted. To address this problem, we tracked the ratio of fixed differences to shared polymorphisms plus fixed differences using 100-kb windows with a 25-kb step size. This ratio can take on values between zero and one and is defined for all regions that contain an informative site. High values indicate reciprocally monophyletic gene genealogies while low values indicate populations that harbor ancestral polymorphism or are experiencing gene flow (*e.g.*, Carneiro *et al.* 2013). For this window analysis, we included only windows with at least three topologically informative SNPs.

In both sliding-window approaches, regions were delimited by joining overlapping or adjacent windows with the same classification, and overall levels of diversity and differentiation were estimated by averaging across all SNPs in a delimited region. The average number of SNPs in these regions is given in Table 2. We also adopted a third approach to defining regions of differentiation by delimiting runs of fixed differences uninterrupted by shared polymorphisms. The results of these analyses were very similar to the two sliding-window analyses and are given in File S2, Figure S1, Table S5, and Table S6.

Demographic simulations

We used coalescent simulations (Hudson 2002) to compare observed patterns of differentiation to those expected under different demographic scenarios (Table S7). Parameters in these models included divergence time, current and ancestral

Table 1 Summary statistics describing patterns of differentiation at all SNPs in pairwise comparisons of the subspecies of *M. musculus*

Subspecies	Subspecies	Chromosome	No. of SNPs	$\overline{F}_{st}$ (SD)	$\overline{D}_{xy}$ (SD)	$\overline{\delta}$ (SD)	$r_{F_{st},D_{xy}}$	$r_{F_{st},\delta}$	$r_{D_{xy},\delta}$	% fixed differences	% private polymorphisms	% shared polymorphisms
M. m. castaneus	*M. m. domesticus*	Autosomes	24,136	0.22 (0.32)	0.41 (0.28)	0.40 (0.29)	0.99	0.98	0.98	9.14	83.66	7.20
		X	226	0.32 (0.41)	0.53 (0.34)	0.53 (0.35)	0.99	0.99	1.00	23.89	73.91	2.21
M. m. castaneus	*M. m. musculus*	Autosomes	23,709	0.26 (0.38)	0.44 (0.30)	0.43 (0.31)	0.97	0.98	0.99	12.50	81.55	5.95
		X	237	0.35 (0.39)	0.54 (0.33)	0.53 (0.33)	0.97	0.98	0.99	18.14	73.00	8.86
M. m. domesticus	*M. m. musculus*	Autosomes	21,598	0.38 (0.41)	0.53 (0.35)	0.52 (0.35)	0.98	0.99	0.99	24.01	70.96	5.03
		X	246	0.46 (0.44)	0.57 (0.38)	0.57 (0.38)	0.99	0.99	1.00	30.08	68.29	1.63

N_e, and migration rates in each direction and were based on maximum-likelihood estimates obtained using the program Isolation with Migration (IM) (Nielsen and Wakeley 2001) in a previous study (Geraldes *et al.* 2011). We assumed no recombination within loci and free recombination among loci. This assumption is reasonable given that linkage disequilibrium decays over distances of 10–50 kb in house mice (Laurie *et al.* 2007). For each pairwise split, we simulated 100,000 gene genealogies, given five chromosomes from each subspecies, and assumed a scaled θ value of 1.33 based on estimates of the mutation rate (4×10^{-9}; Waterston *et al.* 2002), ancestral population size, and the approximate average number of sites surveyed per locus. Because the program ms (Hudson 2002) simulates individual loci, we then compared the distribution of δ from the simulations to the observed measures across individual genes in our data set. We also explored a wider range of demographic parameters to better match simulated distributions to observed values (see *Results*).

Recombination and inversions

Regions of low recombination are expected to be more highly differentiated than other regions of the genome (Noor *et al.* 2001; Rieseberg 2001; Nachman and Payseur 2012). For example, a recent study of sunflowers showed that regions of greater differentiation were strongly associated with reduced recombination (Renaut *et al.* 2013). We used the revised genetic map to estimate the recombination rate for 5-Mb intervals of the mouse genome (Shifman *et al.* 2006; Cox *et al.* 2009). We defined low-recombination regions as intervals with recombination rates falling in the bottom 10% of the genome and high-recombination regions as intervals falling in the top 10% of the genome. We then asked whether levels of differentiation differed between regions of low and high recombination. One limitation of this approach is that the genetic map derives from *M. m. domesticus* and there is some evidence that recombination rate varies among subspecies (Dumont and Payseur 2011; Dumont *et al.* 2011). This likely limits the power to detect differences if they exist. To compare these results with those from a previous study (Geraldes *et al.* 2011), we repeated the analyses estimating recombination rates over 10-Mb intervals. We also repeated the analyses defining high- and low-recombination regions as those falling in the upper or lower 5, 15, and 20% of the distribution of recombination rate.

Inversions may suppress recombination. Inversion data for *M. m. castaneus* and *M. m. musculus* relative to the reference mouse genome (C57BL/6) are available from the Wellcome Trust (Yalcin *et al.* 2012). C57BL/6 is primarily of *M. m. domesticus* origin but contains small introgressions from other subspecies. We used the Mouse Phylogeny Viewer (Wang *et al.* 2012) to eliminate regions not of *M. m. domesticus* origin from the inversion data for *M. m. castaneus* and *M. m. musculus*. The location of inversions between *M. m. castaneus* and *M. m. musculus* was determined

Table 2 Summary statistics describing regions identified as highly differentiated with a sliding-window analysis *vs.* all other regions in pairwise comparisons of the subspecies of M. musculus

Subspecies	Subspecies	Chromosome	Window type[a]	n^b	Average size of region (bp)[c] (SD)	Average no. of SNPs (SD)	$\overline{F}_{st}$ (SD)	$\overline{D}_{xy}$ (SD)	$\overline{\delta}$ (SD)	$\overline{\pi}_1^d$ (SD)	$\overline{\pi}_2^d$ (SD)
M. m. castaneus	*M. m. domesticus*	Autosomes	Highly differentiated	63	133,333* (37,567)	5.13* (2.85)	0.80* (0.10)	0.87* (0.06)	0.87* (0.06)	0.11* (0.12)	0.05* (0.07)
		Autosomes	All others	1651	235,933* (122,354)	13.74* (13.67)	0.21* (0.14)	0.40* (0.12)	0.39* (0.13)	0.26* (0.11)	0.16* (0.09)
		X	Highly differentiated	5	160,000 (33,541)	4.60 (2.30)	0.84 (0.15)	0.90 (0.09)	0.90 (0.09)	0.11 (0.15)	0.03 (0.04)
		X	All others	27	169,444 (37,553)	5.11 (2.87)	0.31 (0.23)	0.45 (0.20)	0.45 (0.20)	0.19 (0.17)	0.13 (0.11)
		All	All	1746	230,985 (121,119)	13.27 (13.45)	0.24 (0.18)	0.42 (0.15)	0.41 (0.16)	0.25 (0.12)	0.15 (0.10)
M. m. castaneus	*M. m. musculus*	Autosomes	Highly differentiated	105	129,762* (35,878)	5.31* (2.84)	0.80* (0.09)	0.87* (0.06)	0.87* (0.06)	0.11* (0.10)	0.05* (0.07)
		Autosomes	All others	1625	229,708* (114,947)	13.55* (13.46)	0.24* (0.15)	0.43* (0.12)	0.41* (0.13)	0.27* (0.11)	0.14* (0.11)
		X	Highly differentiated	6	154,167 (33,229)	3.67 (1.03)	0.86 (0.17)	0.92 (0.09)	0.92 (0.09)	0.03 (0.08)	0.12 (0.14)
		X	All others	26	168,269 (43,335)	5.38 (3.54)	0.32 (0.18)	0.49 (0.14)	0.48 (0.15)	0.21 (0.17)	0.16 (0.12)
		All	All	1762	222,588 (113,626)	12.90 (13.14)	0.28 (0.20)	0.46 (0.16)	0.45 (0.17)	0.25 (0.11)	0.13 (0.11)
M. m. domesticus	*M. m. musculus*	Autosomes	Highly differentiated	287	135,279* (38,250)	5.98* (3.62)	0.81* (0.09)	0.87* (0.06)	0.87* (0.06)	0.07* (0.07)	0.07* (0.09)
		Autosomes	All others	1561	216,944* (111,486)	11.99* (11.47)	0.34* (0.17)	0.49* (0.14)	0.48* (0.14)	0.18* (0.10)	0.15* (0.11)
		X	Highly differentiated	4	162,500 (25,000)	4.75 (0.50)	0.90 (0.09)	0.93 (0.07)	0.93 (0.07)	0.04 (0.04)	0.05 (0.09)
		X	All others	33	170,455 (36,150)	4.97 (2.58)	0.38 (0.22)	0.50 (0.19)	0.50 (0.19)	0.11 (0.09)	0.15 (0.11)
		All	All	1885	203,581 (106,856)	10.94 (10.79)	0.41 (0.23)	0.55 (0.19)	0.54 (0.19)	0.16 (0.10)	0.14 (0.11)

* $P < 10^{-10}$ in two-sided t-tests comparing highly differentiated autosomal regions and all other autosomal regions in each pairwise comparison.

[a] Highly differentiated regions defined as average $\delta \geq 0.8$.

[b] Number of delimited regions.

[c] Regions were delimited by joining overlapping or contiguous windows with the same classification. The resolution of individual windows is limited to the sliding-window increment of 25,000.

[d] $\overline{\pi}_1$ and $\overline{\pi}_2$ refer to nucleotide diversity (Nei and Li 1979) in the first and second subspecies, respectively.

by identifying and removing inversions in both lines that overlap and therefore represent inversions relative to *M. m. domesticus*. Many inversions were identified between each pair of subspecies, but most were small (CD: $\bar{x} = 1762$, range = 99–19,005, $n = 398$; CM: $\bar{x} = 1907$, range = 63–19,752, $n = 620$; DM: $\bar{x} = 1749$, range = 63–23,239, $n = 479$). Therefore, variant SNPs in inversions were rare in our data set. However, many runs of fixed differences spanned inversions. To investigate the relationship between inversions and differentiation, we asked whether runs of fixed differences were more likely to overlap with inversions than expected by chance. To determine the expected overlap, we randomly generated the same number of regions across the genome sampled with replacement from the same size distribution as the runs data and determined the overlap with inversions. We did this 10,000 times and determined the percentile rank of the observed data.

Gene expression differences

The primary motivation for generating transcriptome data was to provide a set of common loci at which patterns of sequence differentiation could be analyzed. Nonetheless, these data also provide an opportunity to study gene expression differences and to compare expression divergence with sequence differentiation.

All mice were unmated, reproductively mature males, but they differed in age. In addition, *M. m. domesticus* individuals were reared in one facility while *M. m. castaneus* and *M. m. musculus* individuals were reared in another. We used two approaches to assess whether differences in rearing conditions might bias expression analysis. First, we calculated the average count of transcripts mapped for each gene in each subspecies correcting for differences in sequencing effort. Pairwise comparisons between subspecies showed that expression patterns were highly correlated (Pearson's correlation, $r_{CD} > 0.99$, d.f. = 15,123, $P < 10^{-15}$; $r_{CM} > 0.99$, d.f. = 15,123, $P < 10^{-15}$; $r_{DM} > 0.99$, d.f. = 15,123, $P < 10^{-15}$). Second, we compared our results to those of another study on gene expression in *M. m. domesticus* and *M. m. musculus* (M. Nachman, unpublished data). In that study, testis transcriptomes were sequenced for three individuals of one inbred line of *M. m. domesticus* (LEWES) and three individuals of one inbred line of *M. m. musculus* (PWK). All mice were unmated, reproductively mature males of the same age, and all were housed in the same room of a single animal care facility. Average mean counts of transcripts mapped per gene for each subspecies after normalization were highly correlated in the two data sets ($r_{DOM\ Base\ Mean} = 0.98$, d.f. = 11,671, $P < 10^{-15}$; $r_{MUS\ Base\ Mean} = 0.98$, d.f. = 11,671, $P < 10^{-15}$). In addition, the $\log_2$fold change in expression for each gene between the two subspecies was significantly correlated between the two studies ($r_{\log2\ fold\ change} = 0.71$, d.f. = 11,671, $P < 10^{-15}$). Although the power of the two studies differs due to design, the majority of genes (~80%) identified in the smaller study as having significantly different expression between the two subspecies after correction for multiple testing ($\alpha = 0.01$) were identified as significantly differently expressed in this study after correction for multiple testing given a less conservative cutoff ($\alpha = 0.05$). These analyses suggest that expression patterns in this study were not strongly biased by differences in rearing conditions.

We identified genes that were differentially expressed in each pairwise comparison of subspecies. Given results from TopHat (see above), HTseq (Anders *et al.* 2014) was used to create tables of counts of reads mapped for all represented genes. All genes with an average read count of $\leq$10 in more than one subspecies were removed from the analysis. The DESeq package in R (Anders and Huber 2010) was used to further filter the data and identify genes with significant differential expression using a binomial test. We first normalized counts to account for differences in sequencing effort among individuals. We then filtered out the bottom 20% of the data based on sums of the counts across all subspecies. This left 12,098 genes with sufficient data for analysis. We estimated dispersions for each subspecies and then used a binomial test to identify differentially expressed genes between each pair of subspecies. *P*-values were adjusted via a Benjamini–Hochberg correction with a false discovery rate of 0.01 (Benjamini and Hochberg 1995). Sites were filtered in each pairwise test if the average normalized read count was $\leq$10 across both subspecies.

We estimated the correlation between measures of sequence differentiation on a gene-by-gene basis and the absolute value of the $\log_2$fold change in expression for each pair of subspecies. Sequence differentiation and expression differentiation might be correlated, particularly if differences in expression are due to sequence changes at or near the gene itself (*i.e.*, *cis*-regulatory changes). In another study, patterns of allele-specific expression in the testes of F_1 hybrids of *M. m. domesticus* and *M. m. musculus* were used to identify genes in which differences in expression between the two subspecies were due to *cis*-regulatory changes (M. Nachman, unpublished data) as in Wittkopp *et al.* 2004. We used those data and repeated the correlation analysis in the DM comparison, restricting it to genes identified as having *cis*-regulatory changes. Similar data were not available for the other two pairwise comparisons.

Testis-specific expression

Genes involved in reproduction are known to evolve quickly (*e.g.*, Begun *et al.* 2000; Wyckoff *et al.* 2000; Good and Nachman 2005), and genes that are tissue-specific have higher rates of evolution than others in mammals (Duret and Mouchiroud 2000). We asked whether regions containing testis-specific genes were more highly differentiated than other regions in the window analysis based on all SNPs. Genes with testis-specific expression were identified using data from Su *et al.* (2004) available via BioGPS (Wu *et al.* 2009). Expression data were reduced to those tissues with support for independent expression, and expression values were averaged over the available measurements for a given tissue (Winter *et al.* 2004). Genes were considered testis-specific if the proportion of total expression in the testis compared to overall expression was $\geq$ 0.1 (Winter *et al.* 2004; File S3). Measures of differentiation for all regions containing testis-specific genes were then compared to measures for all other regions using one-sided *t*-tests.

Identifying candidate regions for reproductive isolation

One approach to identifying candidate genes for reproductive incompatibilities is to look for overlap between the results of genomic scans and other methods such as QTL mapping studies and cline analyses in hybrid zones. There are many reasons to expect that the results of such studies will not overlap: QTL analyses focus on specific traits, hybrid zone data may track more recent processes, and genomic scans of differentiation will identify many regions that do not contribute to reproductive isolation. Nevertheless, intervals that are identified consistently across different methods are good candidates for additional study.

There are no published QTL mapping data of traits relevant to reproductive isolation for crosses between *M. m. castaneus* and *M. m. musculus*, nor are there any detailed hybrid zone studies between these taxa. However, there are published QTL mapping results for the other two pairwise comparisons (White *et al.* 2011, 2012) and many studies of the hybrid zone between *M. m. domesticus* and *M. m. musculus* (*e.g.*, Vanlerberghe *et al.* 1986, 1988; Tucker *et al.* 1992; Prager *et al.* 1993; Munclinger *et al.* 2002; Macholan *et al.* 2007; Teeter *et al.* 2008, 2010; Janoušek *et al.* 2012). For QTL, we identified overlap between 1.5-LOD intervals associated with sterility phenotypes and highly differentiated regions in the window analysis based on all SNPs (White *et al.* 2011, 2012). We combined QTL into a single region for analysis if the QTL 1.5-LOD intervals were overlapping. For comparison with patterns in the *musculus–domesticus* hybrid zone, we used the study by Janoušek *et al.* (2012) in which candidate Bateson–Dobzhansky–Muller incompatibility (BDMI) loci were identified from patterns of introgression and epistasis in two different transects. We identified overlap between a 2-MB window centered on the candidate BDMI loci of Janoušek *et al.* (2012) and highly differentiated regions in the window analysis based on all SNPs. For both types of comparisons, we compared the observed overlap to a distribution created using 10,000 simulated data sets of genomic regions from the same size distribution as those identified as highly differentiated in our study. We repeated all analyses identifying overlap between genes that were differentially expressed in our data and the results of previous studies. For these analyses, we compared the observed overlap to a distribution created using 10,000 simulated data sets. Simulations were conducted by sampling with replacement from among those genes for which there were expression data.

We used the coordinates of all SNPs in the Ensembl mouse genome assembly GRCm38 to identify genes found in

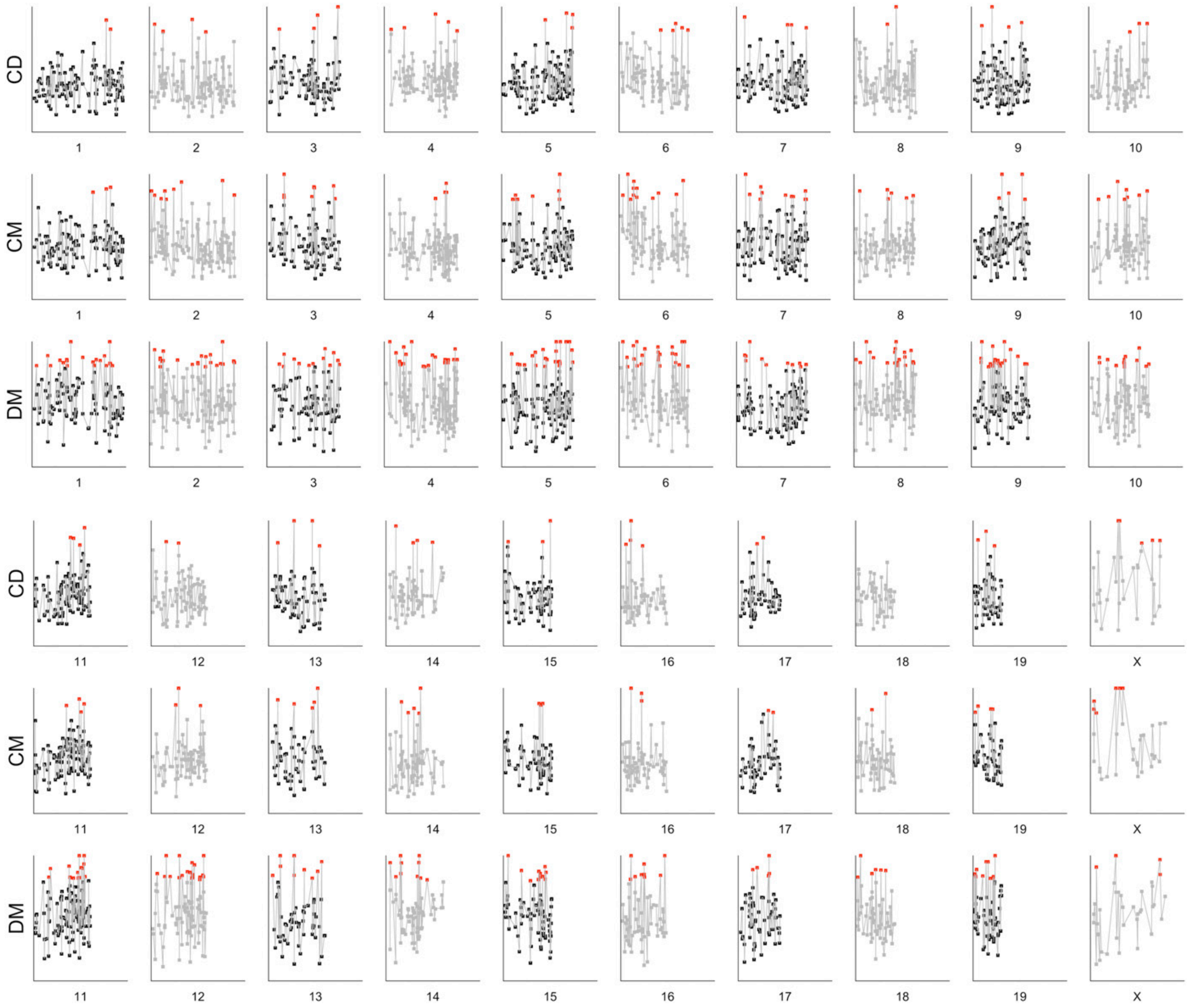


Figure 1 Sliding-window analysis showing average values of δ throughout all chromosomes for each pairwise subspecies comparison (CD, *M. m. castaneus vs. M. m. domesticus*; CM, *M. m. castaneus vs. M. m. musculus*; DM, *M. m. domesticus vs. M. m. musculus*). Each dot marks the start of a delimited region, and red dots represent regions for which the average value of δ is ≥ 0.8.

overlapping regions, and we used the Mouse Genome Database to identify phenotypes in laboratory lines associated with mutations in these genes (Eppig *et al.* 2012). When highly differentiated regions were flanked by regions with fewer than three SNPs, we expanded the query regions to the next region with data or 2 MB, whichever was smaller. We did not include regions from the X chromosome as the QTL associated with male infertility in both crosses encompassed most of the chromosome.

Results

Measures of sequence differentiation

Over 20,000 SNPs were segregating in each pairwise comparison, but many were private, segregating at low-to-moderate frequency within a single subspecies in a given comparison (Table 1). Different measures of differentiation (F_{st}, D_{xy}, and δ) were highly correlated (Table 1). Average levels of differentiation varied among pairwise comparisons, and all measures of differentiation were consistently higher in DM than in either of the other two pairwise comparisons. In addition, all measures of differentiation were higher on the X chromosome than on the autosomes (Table 1). Nonsynonymous sites showed slightly lower levels of differentiation than synonymous sites in each pairwise comparison (Table S8).

There were 9529 individual SNPs that were highly differentiated ($\delta \geq 0.80$) in at least one of the three pairwise comparisons (CD: 4223 from 1970 genes; CM: 5338 from 2343 genes; DM: 7423 from 2783 genes). Fewer SNPs were fixed in *M. m. castaneus* but segregating in both of the other lines (744 from 437 genes) than in either *M. m. domesticus*

(1084 from 651 genes) or *M. m. musculus* (1396 from 881 genes). In addition, many fewer sites represented derived states relative to the other subspecies and both outgroups in *M. m. castaneus* (372 from 263 genes) than in either *M. m. domesticus* (770 from 554 genes) or *M. m. musculus* (1570 from 1099 genes).

Sliding-window analyses

The first sliding-window approach (based on average values of δ) included ~385–403 Mb in each pairwise comparison after filtering (~15% of the genome). We identified many windows with an average value of $\delta \geq 0.80$ in each pairwise comparison between subspecies (Figure 1 and Table 2). Highly differentiated regions were characterized by higher measures of F_{st} and D_{xy} and lower measures of within-subspecies variation than other regions (Table 2). Strikingly, regions of high differentiation represent a much larger part of the total surveyed transcriptome in the DM comparison than in either of the other two comparisons (Table 2). In each pairwise comparison, >65% of genes sampled in highly differentiated regions contained at least one fixed difference. Approximately half of those genes contained at least one nonsynonymous fixed difference (CD: 57.4% of genes; CM: 46.6% of genes; DM: 42.4% of genes). Although this implies that approximately half of these genes have no nonsynonymous fixed differences, it is important to bear in mind that coverage was incomplete for many genes and thus some nonsynonymous changes may have been missed.

The second sliding-window approach (based on the ratio of fixed differences to shared polymorphisms plus fixed differences), after filtering, included ~89–146 Mb in each pairwise comparison. This analysis required at least three topologically informative SNPs per window and thus covered much less of the genome than the first sliding-window approach. We identified many fully sorted windows in each pairwise comparison (Table 3; Figure S2, A–C). Even when including private polymorphisms, autosomal regions that were fully sorted were characterized by higher measures of F_{st}, D_{xy}, and δ and by lower measures of within-subspecies variation than other regions (Table 3). Notably, all regions on the X chromosome were fully sorted in all pairwise comparisons. Fully sorted regions represent a much larger part of the total surveyed transcriptome in the DM comparison than in either of the other two comparisons (Table 3). As expected, a much higher proportion of the surveyed genome was identified as fully sorted in this analysis than was identified as highly differentiated in the analysis averaging across all variable sites in a window.

Demographic simulations

We conducted coalescent simulations based on demographic parameters estimated in a previous study (Geraldes *et al.* 2011; Table S7). The simulations predicted the greatest overall levels of differentiation in the DM comparison and the lowest overall levels in the CD comparison (Figure 2), a pattern consistent with the observed data. However, for all

Table 3 Summary statistics describing patterns of differentiation in fully sorted regions *vs.* all other regions in pairwise comparisons of the subspecies of *M. musculus*

Subspecies	Subspecies	Chromosome type	Window type[a]	n^b	Average size of region (bp) (SD)	Average no. of SNPs (SD)	$\overline{F}_{st}$ (SD)	$\overline{D}_{xy}$ (SD)	$\overline{\delta}$ (SD)	$\overline{\pi}_1^c$ (SD)	$\overline{\pi}_2^c$ (SD)
M. m. castaneus	*M. m. domesticus*	Autosomes	Fully sorted	223	159,193 (38,840)	15.05** (9.17)	0.48*** (0.16)	0.59*** (0.13)	0.59*** (0.13)	0.15*** (0.10)	0.11*** (0.07)
		Autosomes	All others	321	157,243 (46,855)	19.36** (12.56)	0.20*** (0.13)	0.43 *** (0.10)	0.39*** (0.11)	0.29*** (0.09)	0.20*** (0.09)
		X	Fully sorted	5	175,000 (0)	5.20 (1.92)	0.80 (0.12)	0.85 (0.09)	0.85 (0.09)	0.06 (0.06)	0.05 (0.04)
M. m. castaneus	*M. m. musculus*	Autosomes	Fully sorted	317	172,003** (52,573)	16.54 (10.86)	0.47*** (0.17)	0.60*** (0.13)	0.60*** (0.13)	0.20*** (0.10)	0.07*** (0.06)
		Autosomes	All others	287	154,355** (52,334)	18.20 (12.42)	0.19*** (0.14)	0.42*** (0.11)	0.38*** (0.13)	0.28*** (0.09)	0.21*** (0.10)
		X	Fully sorted	3	175,000 (0)	3.67 (1.15)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)
M. m. domesticus	*M. m. musculus*	Autosomes	Fully sorted	624	178,766 *** (55,386)	13.93 (9.50)	0.57*** (0.17)	0.66*** (0.13)	0.66*** (0.13)	0.12*** (0.07)	0.08*** (0.06)
		Autosomes	All others	234	140,171*** (51,005)	14.50 (10.45)	0.27*** (0.16)	0.48*** (0.12)	0.44*** (0.14)	0.22*** (0.09)	0.23*** (0.10)
		X	Fully sorted	10	16,000 (24,152)	5.50 (1.27)	0.70 (0.18)	0.76 (0.15)	0.76 (0.15)	0.07 (0.06)	0.05 (0.06)

* $P < 0.05$, ** $P < 10^{-3}$, *** $P < 10^{-7}$ in two-sided *t*-tests comparing highly differentiated autosomal regions and all other autosomal regions in each pairwise comparison.

[a] Fully sorted regions are defined as those in which: $\frac{\#fixed\ differences}{\#fixed\ differences + \#shared\ polymorphisms} = 1$ (see *Materials and Methods*).

[b] Number of delimited regions.

[c] $\overline{\pi}_1$ and $\overline{\pi}_2$ refer to nucleotide diversity (Nei and Li 1979) in the first and second subspecies, respectively.

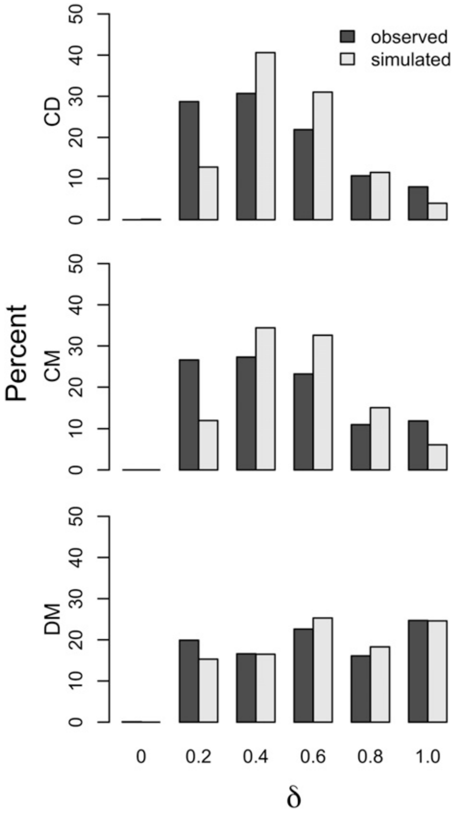


Figure 2 The distribution of δ values in the observed data and in the simulated data based on demographic parameters from Geraldes *et al.* (2011).

three pairwise comparisons, the observed distribution of δ was flatter than the simulated distribution, resulting from a larger-than-predicted proportion of genes with extreme values. Differences between the observed and the simulated distributions were significant in all pairwise comparisons (Kolmogorov–Smirnov test; CD: $D_{\text{2-sided}} = 0.23$, $P < 1 \times 10^{-10}$; CM: $D_{\text{2-sided}} = 0.21$, $P < 1 \times 10^{-10}$; DM: $D_{\text{2-sided}} = 0.19$, $P < 1 \times 10^{-10}$). This pattern was most pronounced in the CD and CM comparisons. An excess of genes with low values of differentiation is consistent with higher-than-simulated rates of gene flow, whereas an excess of genes with high levels of differentiation is consistent with longer-than-simulated divergence times, lineage-specific positive selection, and/or barriers to gene flow. However, on average, the simulated loci had more variable sites than were surveyed in the observed data (Table S7). We repeated the simulations choosing values of current and ancestral N_e and divergence time to try to match more closely the number of SNPs in the simulated and observed data (Table S9). Overall patterns were similar under both demographic scenarios (Figure S3 and Figure S4) with more loci falling in the extremes of the distribution than expected based on the simulations.

We further explored the simulation parameter space to determine if increasing gene flow and divergence times could generate patterns similar to those observed in the data. We started with the original parameter values (Table S7) but with a common divergence time of 325 KYA. We then explored different values of gene flow for each pairwise comparison until the proportion of genes with low values of differentiation ($0 < \delta \leq 0.2$) in the simulations matched the proportion observed in the data. The levels of gene flow required were high, ranging from 7 to 15 times the values originally simulated (Table S10). In all cases, increasing gene flow to the level required resulted in an even larger excess of highly differentiated genes in the observed data relative to the simulations than under the original demographic scenario (Figure S4). Next, we tested whether increasing divergence times in tandem with gene flow could result in a distribution more similar to the one observed. We increased the divergence time first to 425 KYA and then to 825 KYA. Increasing divergence time had little effect on the proportion of genes falling in the extreme tails of the distribution for both the CD and CM comparison (Figure S5 and Figure S6). Increasing divergence times given such high levels of gene flow tended to increase the proportion of simulations for which the average value of δ was low with little effect on the proportion of highly differentiated loci. In the DM comparison, increasing divergence time did increase the proportion of the distribution that was highly differentiated to levels close to or exceeding those observed (Figure S5 and Figure S6). However, in both simulations with older divergence times, the proportion of simulated loci with low values of differentiation was much smaller than observed. We did not exhaustively explore the effects of uncertainty in estimates of effective population size, recombination rate, or mutation rate, any of which might affect the expected distribution of differentiation. Nonetheless, taken together, the simulations suggest that differences in demography can account for some of the observed patterns, such as increased differentiation in the DM comparison, but also that some regions of high differentiation result from either lineage-specific positive selection or barriers to gene flow.

Recombination and inversions

Levels of differentiation were generally higher in regions of low recombination than in regions of high recombination, but the difference was significant only in the CM comparison. Results were similar among different measures of differentiation; we report results for δ (Table 4). Repeating the analysis with 10-Mb windows or with different cutoff values for high- and low-recombination regions yielded qualitatively similar results (data not shown). We found no evidence of greater-than-expected overlap between inversions and runs of fixed differences in the CM and CD comparisons, but we did find significant overlap in the DM comparison, with the observed overlap falling in the extreme tail of the simulated distribution ($P = 0.015$; Table S11)

Gene expression differences

We identified many more significantly differentially expressed genes in the DM comparison than in either of the other two

Table 4 Average sequence differentiation in regions of low and high recombination

Subspecies	Subspecies	$\bar{\delta}_{low\ \ recombination}$ (SD)	$\bar{\delta}_{high\ \ recombination}$ (SD)	t	$P_{1\text{-tailed}}$	n
M. m. castaneus	*M. m. domesticus*	0.29 (0.12)	0.28 (0.06)	0.22	0.83	88
M. m. castaneus	*M. m. musculus*	0.36 (0.10)	0.31 (0.05)	2.72	<0.01	88
M. m. domesticus	*M. m. musculus*	0.38 (0.13)	0.35 (0.07)	1.17	0.12	88

comparisons (CD: 594 of 12,078; CM: 685 of 12,078; DM: 1049 of 12,081; File S4). Average δ per gene and $\log_2$fold change in expression were significantly positively correlated in all pairwise comparisons of subspecies, although the correlation coefficients were small (CD: $r_{\delta,\ abs(\log2\ \ fold\ \ change)} = 0.05$, d.f. = 2483, $P = 0.01$; CM: $r_{\delta,\ abs(\log2\ \ fold\ \ change)} = 0.06$, d.f. = 2471 $P = 0.001$; DM: $r_{\delta,\ abs(\log2\ \ fold\ \ change)} = 0.05$, d.f. = 2356, $P = 0.02$). Results were very similar for other measures of differentiation (data not shown). Restricting the data to genes identified in another study as being significantly differentially transcribed due to *cis*-regulatory changes (M. Nachman, unpublished data) strengthened the correlation in the DM comparison although the significance was reduced given less power ($r_{\delta,\ abs(\log2\ \ fold\ \ change)} = 0.09$, d.f. = 439, $P = 0.05$).

Testis-specific expression

Approximately 30% of regions surveyed in the analyses including all SNPs contained at least one testis-specific gene (Table S12). Overall, testis-specific genes were significantly more common in highly differentiated regions than in other regions ($\chi^2_1 = 3.74$, $n = 9822$, $P_{1\text{-tailed}} = 0.03$). Regions containing testis-specific genes were more differentiated than other regions in the CM and DM comparisons, but these differences were consistently significant only in the DM comparison (Table S12).

Identifying candidate regions for reproductive isolation between M. m. castaneus and M. m. domesticus

In comparisons between *M. m. castaneus* and *M. m. domesticus*, we did not observe significant overlap between differentially expressed genes and QTL associated with hybrid male infertility. However, we did observe significant overlap between highly differentiated regions and QTL (White *et al.* 2012; Figure 3A). Of the nine QTL intervals, seven contained peaks of high differentiation, and the observed overlap ranked in the 97th percentile of the simulated distribution.

Regions of overlap on the autosomes contained 221 protein-coding regions (Table S13). Across all 221 autosomal genes in regions of overlap, 20 genes were testis-specific (*BC049635*, *Bps9*, *Catsper2*, *Ccdc53*, *Ccl27a*, *Ccl27b*, *Eif3j1*, *Faf1*, *Gm13306*, *Lin7a*, *Lrrc57*, *Nup37*, *Parpbp*, *Psmc3*, *Sord*, *Sycp3*, *Tex26*, *Trim69*, *Tsc22d4*, *Ttbk2*), 11 genes had functional annotations and/or phenotypes associated with male fertility (*Arhgap1*, *Bps9*, *Cdkn2c*, *Celf1*, *Duox2*, *Ehd4*, *Igf1*, *Illra1*, *Nr1h3*, *Pmch*, and *Pparg*), and 3 genes were both testis-specific and had mutational variants associated with male infertility (*Catsper2*, *Sord*, *Sycp3*). Nine genes in regions of overlap showed significant expression differences (p_{adj} <0.05; *2700089E24Rik*, *Atg7*, *B2m*, *Capn3*, *Cdkn2c*, *Igf1*, *Nup37*, *Ppip5k1*, *Shf*). Two of those, *Cdkn2c* and *Igf1*, are associated with male fertility, and one, *Nup37*, is testis-specific.

In general QTL are large, while regions of high differentiation are relatively small, potentially helping to narrow QTL intervals. For example, one QTL on chromosome 9 associated with amorphous sperm heads encompasses 26.7 Mb and contains ~220 protein-coding genes (White *et al.* 2012). It overlaps with only one highly differentiated region that contains only one protein-coding gene, *Bbs9*. *Bbs9* has gene ontology (GO) terms relating to cilia and is testis-specific. In humans, *Bbs9* mutations are associated with Bardet–Biedl syndrome, a disease with multiple phenotypic effects including reduced testis size. Expression levels at *Bbs9* were different in *M. m. castaneus* and *M. m. domesticus*, but this difference was not significant after correction for multiple testing ($P < 0.025$, $P_{adj} = 0.16$). We observed three silent and no replacement fixed differences between *M. m. castaneus* and *M. m. domesticus* at *Bbs9*. Not all sites were surveyed, and thus observed patterns of differentiation may reflect linkage to functionally important unsurveyed sites. It is also important to bear in mind that not all genes in QTL intervals were surveyed.

Identifying candidate regions for reproductive isolation between M. m. domesticus and M. m. musculus

In comparisons between *M. m. domesticus* and *M. m. musculus*, the overlap between highly differentiated regions or differentially expressed genes and QTL associated with male infertility (White *et al.* 2011) was not more than expected by chance. The overlap between differentially expressed genes and candidate BDMIs loci from the hyrbid zone study of Janoušek *et al.* (2012) was also not more than expected by chance. However, the overlap between highly differentiated regions and the candidate BDMI loci ranks in the 92nd percentile of simulated data (Janoušek *et al.* 2012). Regions of overlap between all three kinds of studies (QTL, candidate BDMI loci from the hybrid zone, and regions of high differentiation) are particularly promising candidates for reproductive isolation. Importantly, six autosomal regions were identified in which candidate BDMIs and regions of high differentiation overlap precisely or are contiguous and fall within QTL intervals (Figure 3B). These regions collectively contain 242 genes, and the number of genes found in each specific region ranges from 17 to 97 (Table S14).

Two regions fall in relatively small QTL intervals. The first is on chromosome 4. This QTL is associated with relative testis weight (White *et al.* 2011) and contains only

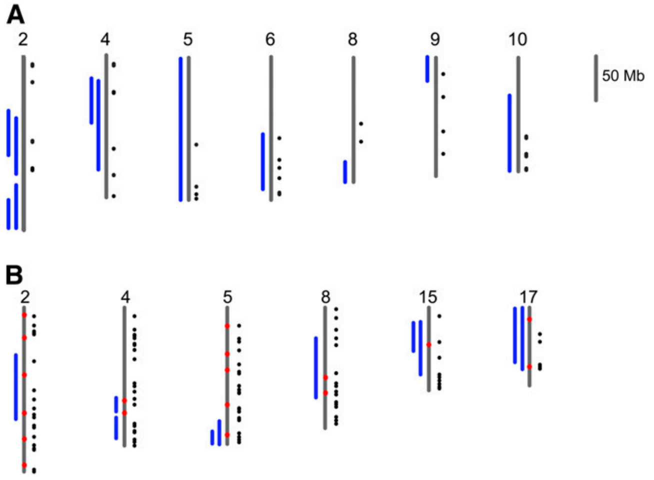


Figure 3 Candidate regions for reproductive incompatibilities. (A) Overlap between autosomal regions identified as QTL associated with male sterility in a cross between *M. m. castaneus* and *M. m. domesticus* (blue bars) (White *et al.* 2012) and regions identified as highly differentiated in our scan based on all SNPs (black dots). (B) Overlap between QTL associated with male infertility (blue bars) (White *et al.* 2011), regions identified as contributing to BDMIs between *M. m. domesticus* and *M. m. musculus* in two-hybrid zones in central Europe (red dots) (Janoušek *et al.* 2012) and regions identified as highly differentiated in our scan based all SNPs (black dots).

16 genes (Figure 4). Of these 16 genes, 4 are testis-specific (*4921539E11Rik*, *Mier1*, *Tctex1d1*, and *Wdr78*) and two (*Insl5* and *Dab1*) are associated with male infertility. We found that three genes (*C8b*, *Dab1*, and *Prkaa2*) in this interval are differentially expressed between *M. m. domesticus* and *M. m. musculus* after correction for multiple testing ($\alpha = 0.05$).

The second small QTL interval with precise overlap is found on chromosome 5. This QTL is associated with both total abnormal sperm and distal bent-tail phenotypes. This interval contains 97 genes and is relatively well sampled in our study. There are 14 testis-specific genes in this interval (*Fbxo24*, *Mcm7*, *Mepce*, *Muc3*, *Myl10*, *Ppp1r35*, *Rabl5*, *Srrt*, *Stag3*, *Taf6*, *Tmem184a*, *Tsc22d4*, *Znhit1*, and *Zscan21*). Two of these genes (*Stag3* and *Zscan21*) have GO annotations and/or phenotypes relating to male fertility. *Myl10* and *Rabl5* were differentially expressed between *M. m. domesticus* and *M. m. musculus* after correction for multiple testing ($P_{adj} < 0.05$). There are 10 additional genes in the region with known functional annotations and/or phenotypes relating to male fertility (*Ache*, *Cnpy4*, *Cux1*, *Fam20c*, *Pdgfa*, *Smok3a*, *Smok3b*, *Sun1*, *Vgf*, and *Zan*).

Discussion

Genome-wide patterns of sequence differentiation

We used a transcriptomic approach to characterize genome-wide patterns of differentiation between the three subspecies of house mice and discovered many highly differentiated regions in all pairwise comparisons. By comparing three subspecies that split from one another at approximately the same time but that have different estimated effective population sizes, we were able to study the influence of demography on the early stages of speciation and divergence. In this case, we found higher levels of sequence differentiation between *M. m. domesticus* and *M. m. musculus* than between the other pairs of subspecies. This result is consistent with estimates of the demographic history of these species; both *M. m. domesticus* and *M. m. musculus* are believed to have undergone significant bottlenecks resulting in a current N_e of ~1/10 to 1/2 of the ancestral N_e. *M. m. castaneus*, on the other hand, is estimated to have a population size very similar to the ancestral population (Geraldes *et al.* 2011). Lineage-specific changes observed in this study support those expectations. The fewest lineage-specific changes were assigned to *M. m. castaneus*, the subspecies with the highest N_e, and the most were assigned to *M. m. musculus*, the subspecies with the smallest N_e. More generally, the coalescent simulations performed here recapitulated the broad patterns of differentiation seen among the three subspecies, suggesting that many of the observed patterns can be explained by differences in N_e and levels of gene flow (Figure 2).

At the same time, greater reproductive isolation is seen in laboratory crosses between *M. m. domesticus* and *M. m. musculus* than between *M. m. castaneus* and *M. m. domesticus* or *M. m. castaneus* and *M. m. musculus* (Dumont and Payseur 2011; White *et al.* 2011, 2012). This observation, by itself, leads to the prediction of greater differentiation in the DM comparison than in the other two comparisons. Because this pattern is also predicted by demographic differences among the subspecies, it is difficult to disentangle the relative contribution of differences in demography and differences in reproductive isolation to the observed patterns. It is also unclear whether differences in demography are the cause of the differing levels of reproductive isolation. For example, if most BDMI alleles were neutral on their own genetic background, then subspecies with smaller N_e would be expected to accumulate more BDMI differences due to drift and would therefore show greater reproductive isolation. However, most BDMI genes in other systems seem to show evidence of positive selection, suggesting that drift is not the predominant process fixing alleles involved in BDMIs (Coyne and Orr 2004; Presgraves 2010).

Differentiation on the X chromosome

The X chromosome was significantly more differentiated than the autosomes in all pairwise comparisons (Table 1). In principle, this pattern is expected for two reasons. First, the smaller estimated N_e of the X chromosome should result in faster lineage sorting. Second, this pattern is consistent with the large X effect, that is, the disproportionate accumulation of reproductive incompatibilities on the X chromosome (*e.g.*, Coyne and Orr 1989). In this case, the greater level of differentiation appears to be more than can be explained by differences in the X to autosome ratio of N_e. At migration-drift equilibrium, assuming constant bidirectional migration,

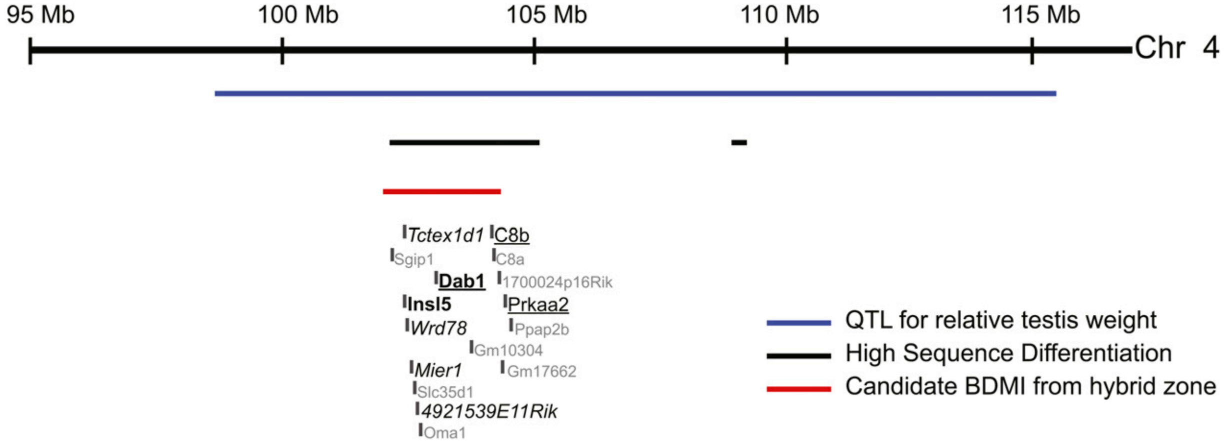


Figure 4 A region on chromosome 4 in which a QTL for relative testis weight (White *et al.* 2011), a candidate BDMI (Janoušek *et al.* 2012), and a highly differentiated region identified in this study overlap in the DM comparison. Testis-specific genes are given in italics, differentially expressed genes are underlined, and genes known to be related to male fertility are given in boldface type. All other genes are shown in grey.

a sex ratio of 1, and equal migration of males and females, $F_{st} = 1/(4Nm + 1)$ for the autosomes and $F_{st} = 1/(3Nm + 1)$ for the X chromosome. If Nm is ~0.1 (Table S7), then the expected X:autosome ratio of F_{st} is 1.08 and the observed ratios are 1.45 (CD), 1.35 (CM), and 1.21 (DM) (Table 1). This model is clearly overly simplistic. For example, there is some evidence of male-biased dispersal in this system (Pocock *et al.* 2005). However, these rough calculations suggest that differences in N_e alone cannot account for the greater differentiation seen on the X chromosome.

On the other hand, our observations are consistent with previous work suggesting a large X effect. For example, hybrid zone studies of *M. m. domesticus* and *M. m. musculus* indicate reduced introgression on the X (Tucker *et al.* 1992; Dod *et al.* 1993), and IM analysis of a limited number of loci in all three subspecies suggests that gene flow on the X chromosome is lower than that of autosomes (Geraldes *et al.* 2008, 2011). In laboratory crosses, hybrid male sterility phenotypes map to the X chromosome in crosses between *M. m. domesticus* and *M. m. musculus* (Storchová *et al.* 2004; Good *et al.* 2008; White *et al.* 2011) and *M. m. castaneus* and *M. m. domesticus* (White *et al.* 2012). Our findings here demonstrate that elevated differentiation on the X chromosome is a general pattern, is observed in allopatric populations, and extends to all pairs of subspecies.

Recombination

Several recent models suggest that chromosomal rearrangements may lead to reduced gene flow via their effect on suppressing recombination (Noor *et al.* 2001; Rieseberg 2001; Navarro and Barton 2003). Recombination can also influence differentiation by amplifying the effects of genetic hitchhiking and background selection within lineages (Maynard Smith and Haigh 1974; Charlesworth 1993), reducing variation within lineages and thus leading to increased differentiation between lineages. We found weak support for a negative relationship between differentiation and recombination, consistent with these models. However, the power of this approach may be limited by the absence of data on recombination rate variation across the genome in all three subspecies.

Testis-specific expression

Regions of high differentiation contained a significantly higher proportion of testis-specific genes than other regions. This result is unexpected if highly differentiated regions simply reflect stochastic variation in lineage sorting. In contrast, this result is consistent with (1) reduced gene flow due to BDMIs associated with testis-specific genes, (2) hitchhiking effects associated with positive selection at testis-specific genes, (3) or both. Importantly, this observation suggests that some proportion of highly differentiated regions is associated with functional differences within or between nascent species.

Candidate regions for reproductive incompatibilities

It would be incorrect to claim that all regions of high differentiation contribute to reproductive isolation when many such regions are expected simply as a consequence of drift in small populations. In addition, some regions of high differentiation are likely the result of lineage-specific selection that may not contribute to reproductive isolation. Nonetheless, the observed data differed from demographic simulations in one major way: the distribution of differentiation statistics was flatter, with more values in the extremes of the distribution. This is consistent with more gene flow and more differentiation than expected. Even when exploring a broad range of values for gene flow and divergence time, we were unable to find a demographic scenario that recapitulated observed patterns. Therefore, some highly differentiated regions likely result either from lineage-specific positive selection and/or from barriers to gene flow at loci underlying incompatibilities.

One approach to prioritizing candidate reproductive isolation loci is to identify overlap between the results of genome scans, laboratory crosses, and hybrid zone analyses.

We identified several areas of overlap between QTL associated with male sterility in a cross of *M. m. castaneus* and *M. m. domesticus* (White *et al.* 2012) and highly differentiated regions identified in our study. The overlap was more than expected by chance, but in most cases QTL were large, making the overlap difficult to interpret (*e.g.*, chromosome 5, Figure 3A). However, in other cases, the QTL intervals were narrower, there was reasonable coverage in our data set, and relatively few genes were found in the overlap. For example, on chromosome 9, there is just one gene in a region of high differentiation that falls in a moderately sized QTL. Even though >200 genes fall in all of the areas of overlap, this number is considerably smaller than the total number of genes that fall in QTL intervals (White *et al.* 2012). Moreover, fewer than 3 dozen of those genes have known phenotypes or GO terms that relate to male fertility and/or are testis-specific as might be expected if they affect male sterility phenotypes measured in the QTL analyses. Of course, GO annotation and documentation of phenotypes associated with mutations or knockouts in mice are far from complete, and additional genes in these regions may be related to fertility.

In the DM comparison, overlap between our results and QTL analyses was considerable, but not more than expected by chance. More promisingly, there was meaningful overlap between candidate BMDIs (Janoušek *et al.* 2012) and our results. In particular, there were six cases in which regions identified as highly differentiated in our study were directly overlapping or contiguous with a candidate BDMI and fell in a QTL interval. In two of those cases, the overlap was relatively precise, and the region of overlap contains a short list of genes that are testis-specific, differentially expressed, and/or related to male infertility. While there is still much work to be done, the intersection of results from multiple studies is encouraging and highlights the promise of this approach for narrowing QTL intervals.

Acknowledgments

We thank F. Bonhomme, P. Campbell, C. Clayton, M. Dean, and A. Orth for providing tissue samples; J. MacDonald for providing access to the source code for the program DNA Slider; and members of the Nachman Lab, D. Begun, D. Matute, and one anonymous reviewer for their thoughtful comments on the manuscript. This research was funded by National Institutes of Health grant R01 GM074245 to M.W.N. This work used the Extreme Science and Engineering Discovery Environment, which is supported by National Science Foundation grant ACI-1053575.

Literature Cited

Anders, S., and W. Huber, 2010 Differential expression analysis for sequence count data. Genome Biol. 11: R106.

Anders S., P. T. Pyl, and W. Huber, 2014 HTSeq: a Python framework to work with high-throughput sequencing data. bioRxiv DOI: 10.1101/002824.

Andrés, J. A., E. L. Larson, S. M. Bogdanowicz, and R. G. Harrison, 2013 Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. Genetics 193: 501–513.

Begun, D. J., P. Whitley, B. L. Todd, H. M. Waldrip-Dail, and A. G. Clark, 2000 Molecular population genetics of male accessory gland proteins in Drosophila. Genetics 156: 1879–1888.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., B 57: 289–300.

Bikard, D., D. Patel, C. L. Metté, V. Giorgi, C. Camilleri *et al.*, 2009 Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. Science 323: 623–626.

Bomblies, K., J. Lempe, P. Epple, N. Warthmann, C. Lanz *et al.*, 2007 Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. PLoS Biol. 5: e236.

Bonhomme, F., E. Rivals, A. Orth, G. R. Grant, A. J. Jeffreys *et al.*, 2007 Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among house mouse subspecies. Genome Biol. 8: R80.

Boursot, P., J. C. Auffray, J. Britton-Davidian, and F. Bonhomme, 1993 The evolution of house mice. Annu. Rev. Ecol. Syst. 24: 119–152.

Brideau, N. J., H. A. Flores, J. Wang, S. Maheshwari, X. Wang *et al.*, 2006 Two Dobzhansky-Muller genes interact to cause hybrid lethality in Drosophila. Science 314: 1292–1295.

Carneiro, M., S. J. E. Baird, S. Afonso, E. Ramirez, P. Tarroso *et al.*, 2013 Steep clines within a highly permeable genome across a hybrid zone between two subspecies of the European rabbit. Mol. Ecol. 22: 2511–2525.

Charlesworth, B., 1993 The evolution of sex and recombination in a varying environment. J. Hered. 84: 345–350.

Charlesworth, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15: 538–543.

Cox, A., C. L. Ackert-Bicknell, B. L. Dumont, Y. Ding, J. T. Bell *et al.*, 2009 A new standard genetic map for the laboratory mouse. Genetics 182: 1335–1344.

Coyne, J. A., and H. A. Orr, 1989 Patterns of speciation in Drosophila. Evolution 43: 362–381.

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. Sinauer Associates, Sunderland, MA.

Cruickshank, T. C., and M. W. Hahn, 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23: 3133–3157.

Cucchi, T., J.-D. Vigne, and J.-C. Auffray, 2005 First occurrence of the house mouse (*Mus musculus domesticus* SCHWARTZ & SCHWARTZ, 1943) in Western Mediterranean: a revision of sub-fossil house mice occurrences using a zooarchaeological critical grid. Biol. J. Linn. Soc. Lond. 84: 429–445.

Dod, B., L. S. Jermiin, P. Boursot, V. H. Chapman, J. T. Nielsen *et al.*, 1993 Counterselection on sex chromosomes in the *Mus musculus* European hybrid zone. J. Evol. Biol. 6: 529–546.

Dumont, B. L., and B. A. Payseur, 2011 Genetic analysis of genome-scale recombination rate evolution in house mice. PLoS Genet. 7: e1002116.

Dumont, B. L., M. A. White, B. Steffy, T. Wiltshire, and B. A. Payseur, 2011 Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. Genome Res. 21: 114–125.

Duret, L., and D. Mouchiroud, 2000 Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17: 68–070.

Duvaux, L., K. Belkhir, M. Boulesteix, and P. Boursot, 2011 Isolation and gene flow: inferring the speciation history of European house mice. Mol. Ecol. 20: 5248–5264.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backström *et al.*, 2012 The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756–760.

Eppig, J. T., J. A. Blake, C. J. Bult, J. A. Kadin, and J. E. Richardson, 2012 The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res. 40: D881–D886.

Gagnaire, P.-A., E. Normandeau, and L. Bernatchez, 2012 Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American eels. Mol. Biol. Evol. 29: 2909–2919.

Geraldes, A., P. Basset, B. Gibson, K. L. Smith, B. Harr *et al.*, 2008 Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol. Ecol. 17: 5349–5363.

Geraldes, A., P. Basset, K. L. Smith, and M. W. Nachman, 2011 Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. Mol. Ecol. 20: 4722–4736.

Good, J. M., and M. W. Nachman, 2005 Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. Mol. Biol. Evol. 22: 1044–1052.

Good, J. M., M. D. Dean, and M. W. Nachman, 2008 A complex genetic basis to X-linked hybrid male sterility between two species of house mice. Genetics 179: 2213–2228.

Halligan, D. L., F. Oliver, A. Eyre-Walker, B. Harr, and P. D. Keightley, 2010 Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 6: e1000825.

Harr, B., 2006 Genomic islands of differentiation between house mouse subspecies. Genome Res. 16: 730–737.

Harrison, R. G., 1986 Pattern and process in a narrow hybrid zone. Heredity 56: 347–359.

Harrison, R. G., 2012 The language of speciation. Evolution 66: 3643–3657.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Hudson, R. R., M. Slatkin, and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.

Janoušek, V., L. Wang, K. Luzynski, P. Dufková, M. M. Vyskočilová *et al.*, 2012 Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between *Mus musculus musculus* and *M. m. domesticus*. Mol. Ecol. 21: 3032–3047.

Key, K. H. L., 1968 The concept of stasipatric speciation. Syst. Zool. 17: 14–22.

Kulathinal, R. J., L. S. Stevison, and M. A. F. Noor, 2009 The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet. 5: e1000550.

Laurie, C. C., D. A. Nickerson, A. D. Anderson, B. S. Weir, R. J. Livingston *et al.*, 2007 Linkage disequilibrium in wild mice. PLoS Genet. 3: e144.

Lawniczak, M. K. N., S. J. Emrich, A. K. Holloway, A. P. Regier, M. Olson *et al.*, 2010 Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. Science 330: 512–514.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*; 1000 Genome Project Data Processing Subgroup, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Macholán, M., P. Munclinger, M. Sugerková, P. Dufková, B. Bímová *et al.*, 2007 Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. Evolution 61: 746–771.

Masly, J. P., C. D. Jones, M. A. F. Noor, J. Locke, and H. A. Orr, 2006 Gene transposition as a cause of hybrid sterility in Drosophila. Science 313: 1448–1450.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.

Mihola, O., Z. Trachtulec, C. Vlcek, J. C. Schimenti, and J. Forejt, 2009 A mouse speciation gene encodes a meiotic histone H3 methyltransferase. Science 323: 373–375.

Munclinger, P., E. Boziková, M. Sugerková, J. Piálek, and M. Macholán, 2002 Genetic variation in house mice (*Mus*, Muridae, Rodentia) from the Czech and Slovak Republics. Folia Zool. (Brno) 51: 81–92.

Nachman, M. W., and B. A. Payseur, 2012 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367: 409–421.

Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra *et al.*, 2012 Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367: 343–353.

Navarro, A., and N. H. Barton, 2003 Chromosomal speciation and molecular divergence: accelerated evolution in rearranged chromosomes. Science 300: 321–324.

Neafsey, D. E., M. K. N. Lawniczak, D. J. Park, S. N. Redmond, M. B. Coulibaly *et al.*, 2010 SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. Science 330: 514–517.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., and W. H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 5269–5273.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov Chain Monte Carlo approach. Genetics 158: 885–896.

Noor, M. A., K. L. Grams, L. A. Bertucci, and J. Reiland, 2001 Chromosomal inversions and the reproductive isolation of species. Proc. Natl. Acad. Sci. USA 98: 12084–12088.

Nosil, P., and D. Schluter, 2011 The genes underlying the process of speciation. Trends Ecol. Evol. 26: 160–167.

Pocock, M. J. O., H. C. Hauffe, and J. B. Searle, 2005 Dispersal in house mice. Biol. J. Linn. Soc. Lond. 84(3): 565–583.

Prager, E. M., R. D. Sage, U. Gyllensten, W. K. Thomas, R. Hübner *et al.*, 1993 Mitochondrial DNA sequence diversity and the colonization of Scandinavia by house mice from East Holstein. Biol. J. Linn. Soc. Lond. 50: 85–122.

Presgraves, D. C., 2010 The molecular evolutionary basis of species formation. Nat. Rev. Genet. 11: 175–180.

Presgraves, D. C., L. Balagopalan, S. M. Abmayr, and H. A. Orr, 2003 Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. Nature 423: 715–719.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Renaut, S., A. W. Nolte, and L. Bernatchez, 2010 Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae). Mol. Ecol. 19: 115–131.

Renaut, S., C. J. Grassa, S. Yeaman, B. T. Moyers, Z. Lai *et al.*, 2013 Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat. Commun. 4: 1827.

Rieseberg, L. H., 2001 Chromosomal rearrangements and speciation. Trends Ecol. Evol. 16: 351–358.

Rieseberg, L. H., J. Whitton, and K. Gardner, 1999 Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. Genetics 152: 713–727.

Salcedo, T., A. Geraldes, and M. W. Nachman, 2007 Nucleotide variation in wild and inbred mice. Genetics 177: 2277–2291.

She, J. X., F. Bonhomme, P. Boursot, L. Thaler, and F. Catzeflis, 1990 Molecular phylogenies in the genus Mus: comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. Biol. J. Linn. Soc. 41: 83–103.

Shifman, S., J. T. Bell, R. R. Copley, M. S. Taylor, R. W. Williams *et al.*, 2006 A high-resolution single nucleotide polymorphism genetic map of the mouse genome. PLoS Biol. 4: e395.

Storchová, R., S. Gregorová, D. Buckiová, V. Kyselová, P. Divina *et al.*, 2004 Genetic analysis of X-linked hybrid sterility in the house mouse. Mamm. Genome 15: 515–524.

Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching *et al.*, 2004 A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA 101: 6062–6067.

Suzuki, H., T. Shimada, M. Terashima, K. Tsuchiya, and K. Aplin, 2004 Temporal, spatial, and ecological modes of evolution of Eurasian Mus based on mitochondrial and nuclear gene sequences. Mol. Phylogenet. Evol. 33: 626–646.

Teeter, K. C., B. A. Payseur, L. W. Harris, M. A. Bakewell, L. M. Thibodeau *et al.*, 2008 Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. 18: 67–76.

Teeter, K. C., L. M. Thibodeau, Z. Gompert, C. A. Buerkle, M. W. Nachman *et al.*, 2010 The variable genomic architecture of isolation between hybridizing species of house mice. Evolution 64: 472–485.

Ting, C.-T., S.-C. Tsaur, M.-L. Wu, and C.-I. Wu, 1998 A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282: 1501–1504.

Trapnell, C., L. Pachter, and S. L. Salzberg, 2009 TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.

Tucker, P. K., R. D. Sage, J. Warner, A. C. Wilson, and E. M. Eicher, 1992 Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. Evolution 46: 1146–1163.

Tucker, P. K., S. A. Sandstedt, and B. L. Lundrigan, 2005 Phylogenetic relationships in the subgenus Mus (genus Mus, family Muridae, subfamily Murinae): examining gene trees and species trees. Biol. J. Linn. Soc. Lond. 84: 653–662.

Turner, T. L., M. W. Hahn, and S. V. Nuzhdin, 2005 Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. 3: e285.

Vanlerberghe, F., B. Dod, P. Boursot, M. Bellis, and F. Bonhomme, 1986 Absence of Y-chromosome introgression across the hybrid zone between *Mus musculus domesticus* and *Mus musculus musculus*. Genet. Res. 48: 191–197.

Vanlerberghe, F., P. Boursot, J. T. Nielsen, and F. Bonhomme, 1988 A steep cline for mitochondrial DNA in Danish mice. Genet. Res. 52: 185–193.

Wang, J. R., F. P.-M. de Villena, H. A. Lawson, J. M. Cheverud, G. A. Churchill *et al.*, 2012 Imputation of single-nucleotide polymorphisms in inbred mice using local phylogeny. Genetics 190: 449–458.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, and J. F. Abril, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

White, M. A., C. Ané, C. N. Dewey, B. R. Larget, and B. A. Payseur, 2009 Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5: e1000729.

White, M. A., B. Steffy, T. Wiltshire, and B. A. Payseur, 2011 Genetic dissection of a key reproductive barrier between nascent subspecies of house mice, *Mus musculus domesticus* and *Mus musculus musculus*. Genetics 169: 289–304.

White, M. A., M. Stubbings, B. L. Dumont, and B. A. Payseur, 2012 Genetics and evolution of hybrid male sterility in house mice. Genetics 191: 917–934.

Winter, E. E., L. Goodstadt, and C. P. Ponting, 2004 Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res. 14: 54–61.

Wittkopp, P. J., B. K. Haerum, and A. G. Clark, 2004 Evolutionary changes in *cis* and *trans* gene regulation. Nature 430: 85–88.

Wu, C., C. Orozco, J. Boyer, M. Leglise, J. Goodale *et al.*, 2009 BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 10: R130.

Wyckoff, G. J., W. Wang, and C. I. Wu, 2000 Rapid evolution of male reproductive genes in the descent of man. Nature 403: 304–309.

Yalcin, B., D. J. Adams, J. Flint, and T. M. Keane, 2012 Next-generation sequencing of experimental mouse strains. Mamm. Genome 23: 490–498.

Yang, H., J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell *et al.*, 2011 Subspecific origin and haplotype diversity in the laboratory mouse. Nat. Genet. 43: 648–655.

Communicating editor: D. Begun

GENETICS

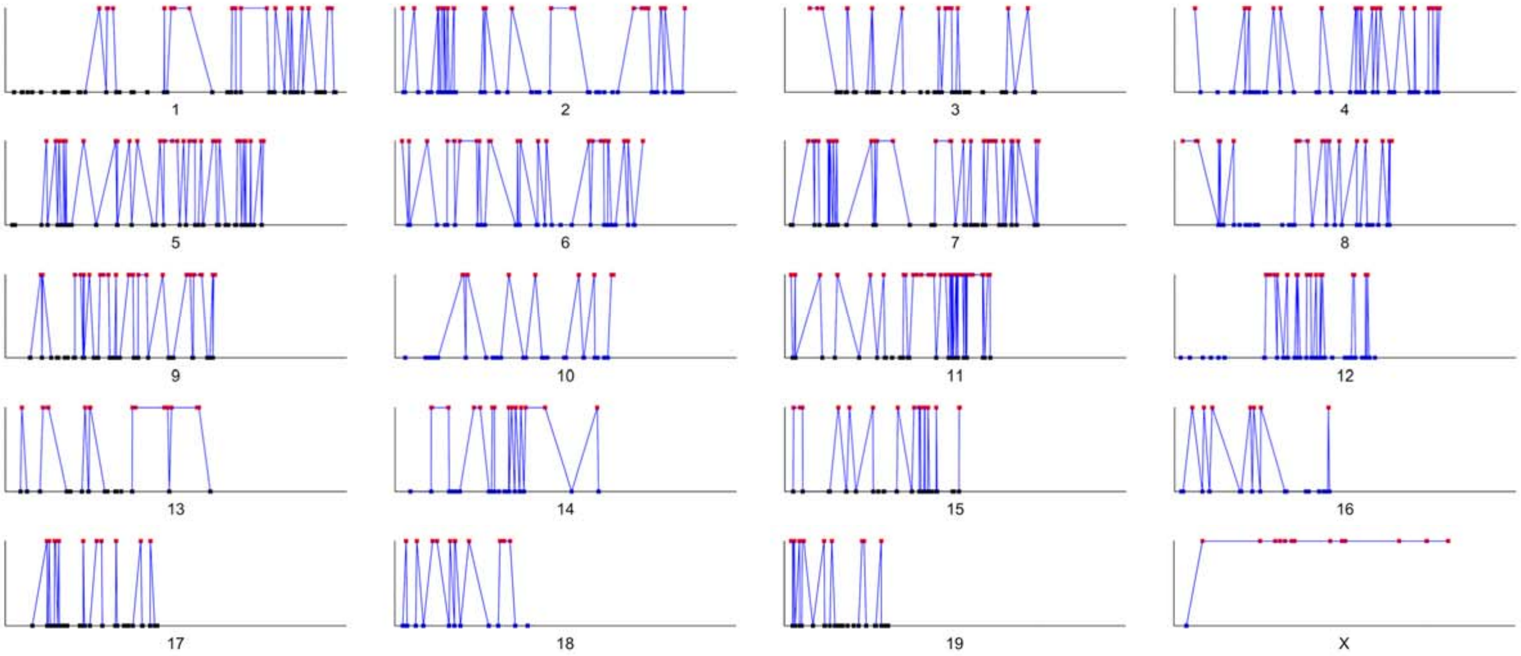
Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.166827/-/DC1

Genome-Wide Patterns of Differentiation Among House Mouse Subspecies

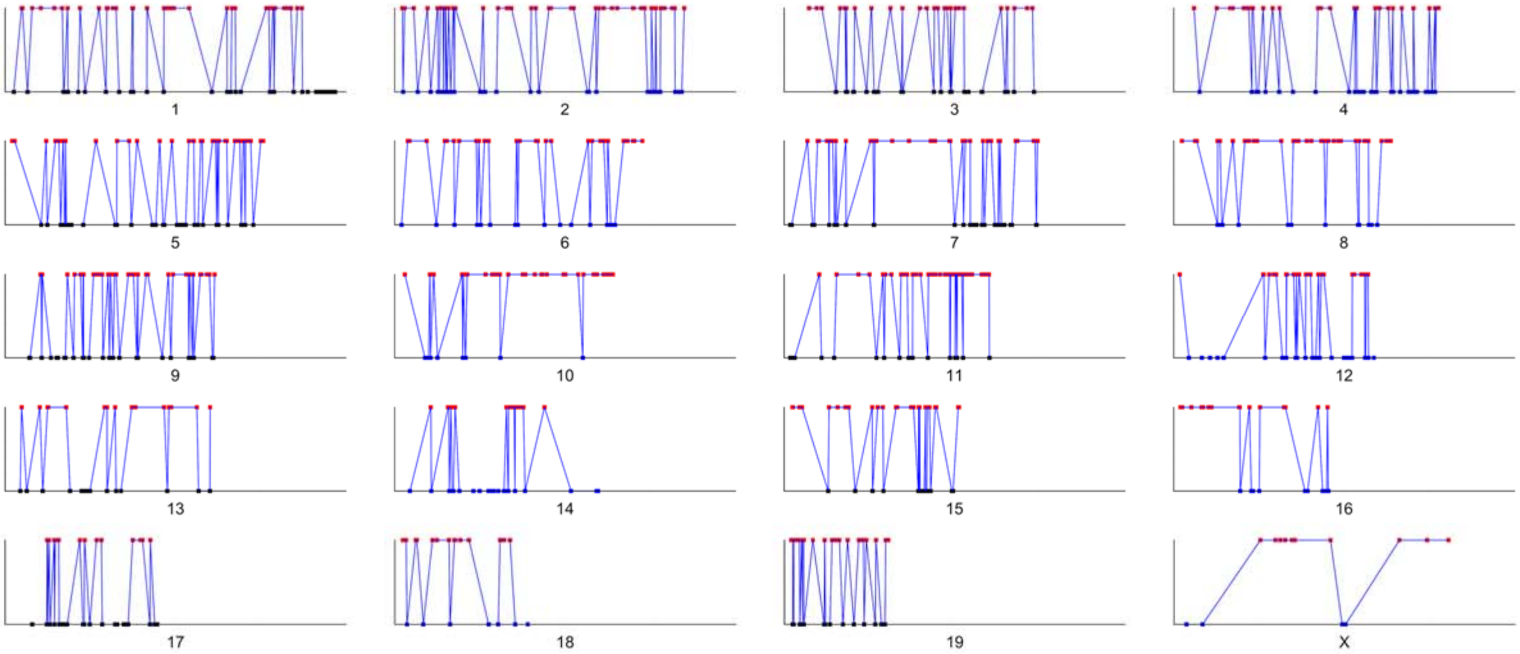
Megan Phifer-Rixey, Matthew Bomhoff, and Michael W. Nachman

DOI: 10.1534/genetics.114.166827

A



B



C

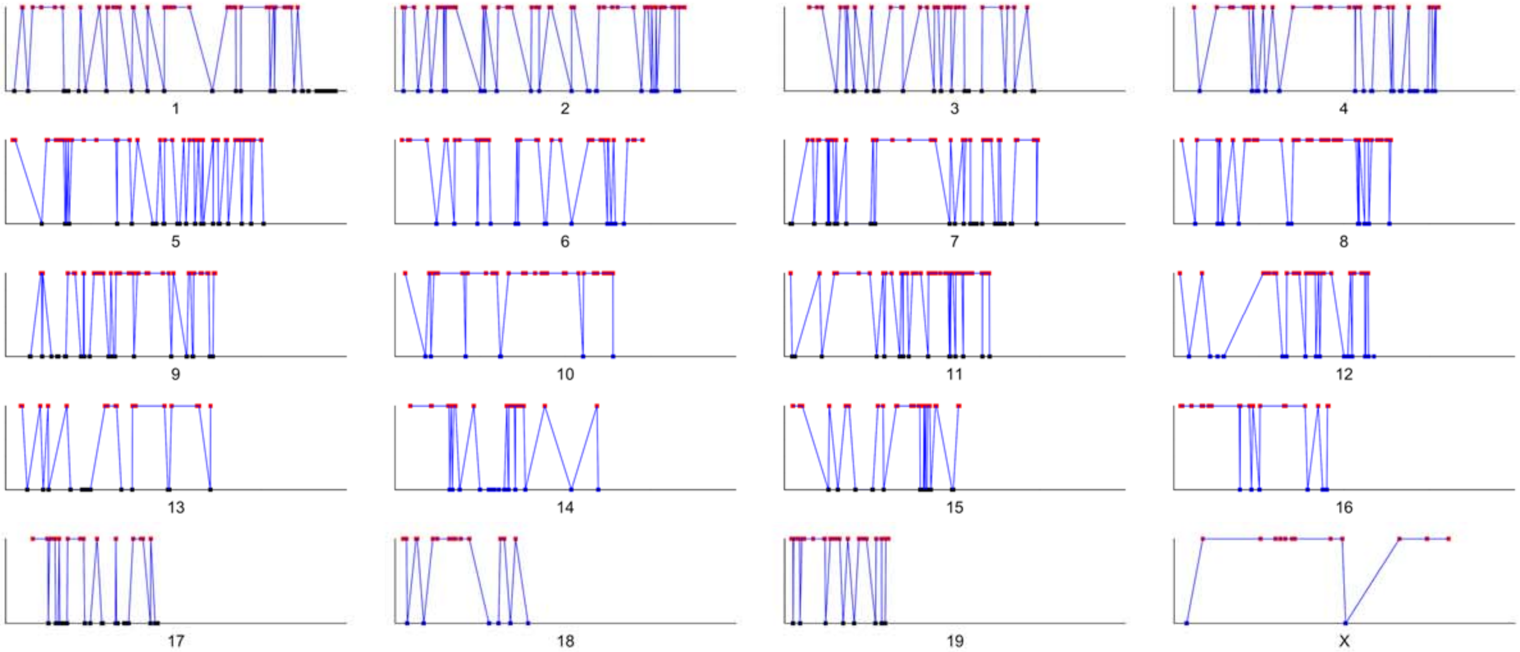
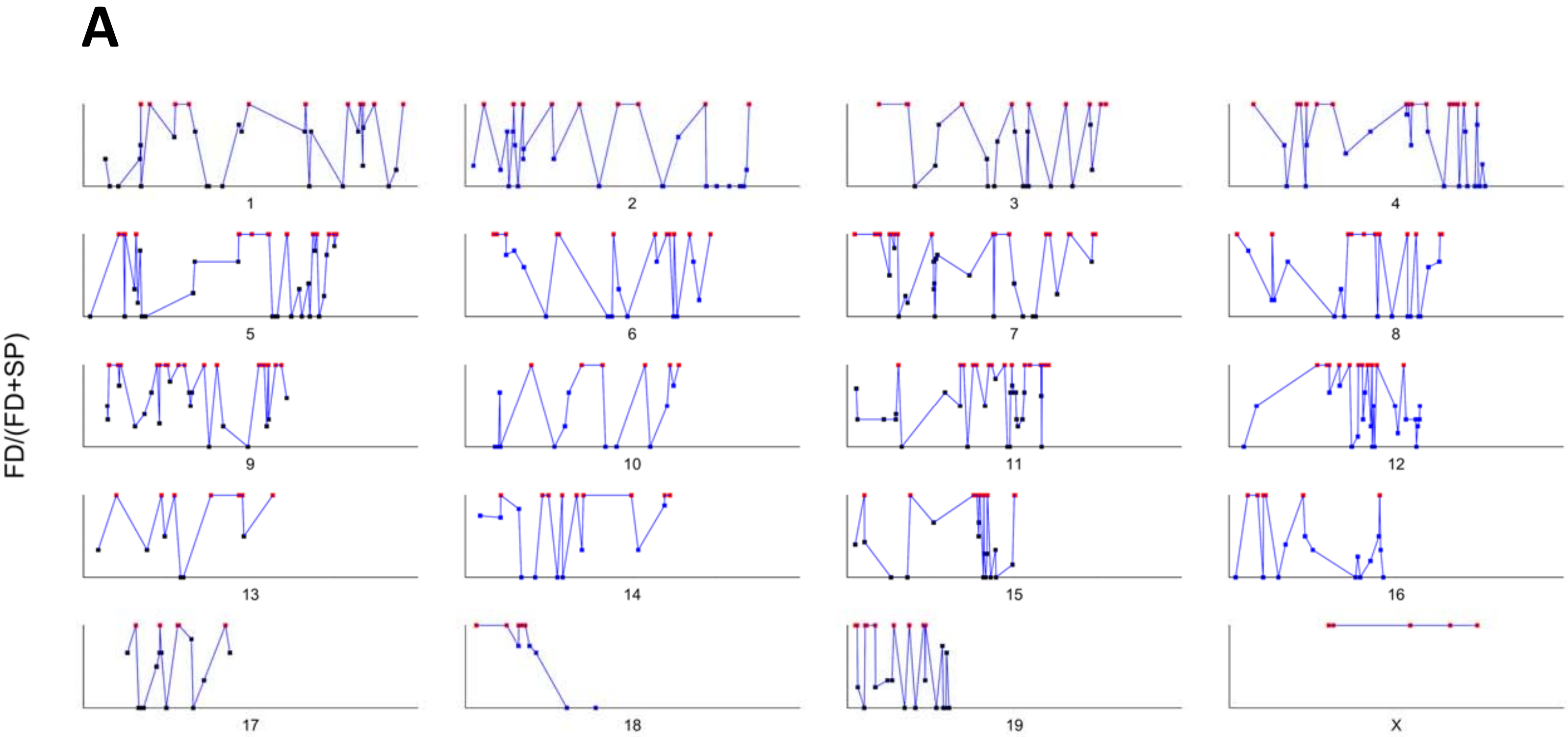
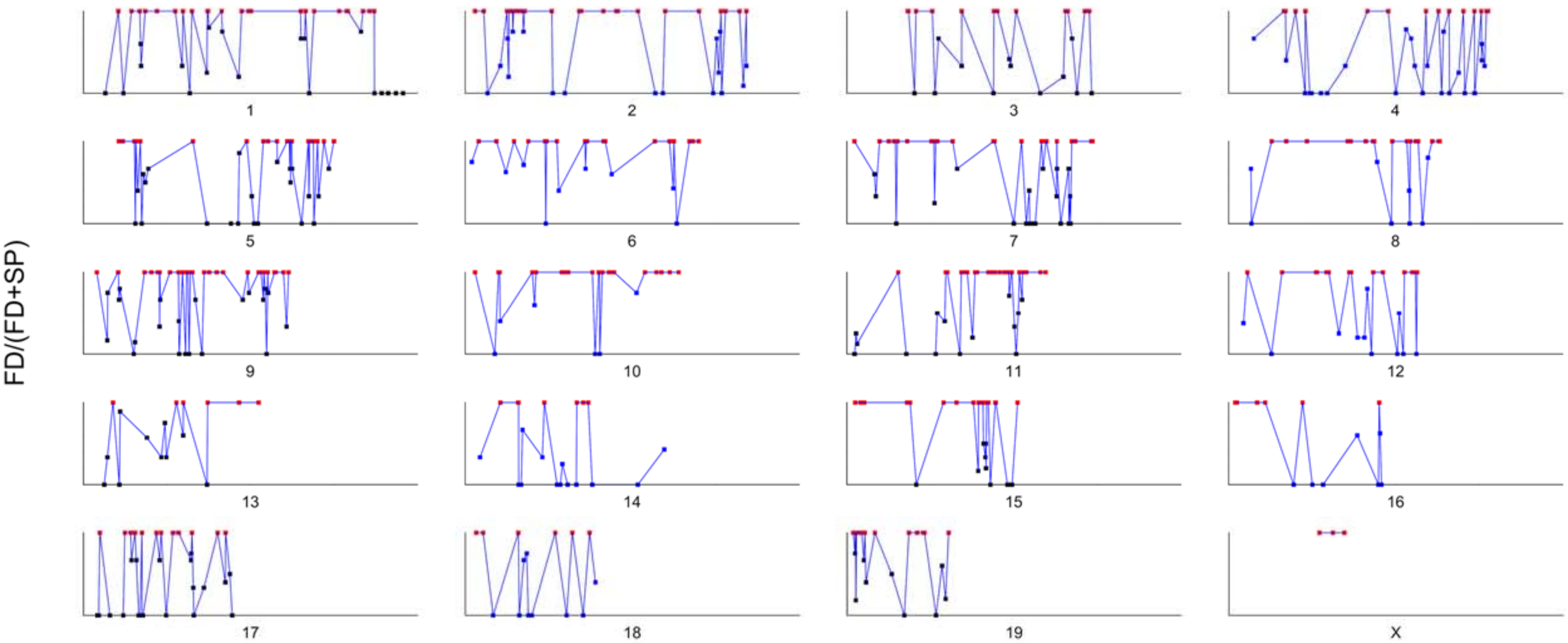


Figure S1 Fixed differences and shared polymorphisms across the genome for all pairwise comparisons of subspecies of *Mus*. Fixed differences are shown as red dots above the axis while shared polymorphisms are shown as dots on the x axis. (A) *M. m. castaneus* and *M. m. domesticus*. (B) *M. m. castaneus* and *M. m. musculus*. (C) *M. m. domesticus* and *M. m. musculus*.


A
FD/(FD+SP)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X

B



C

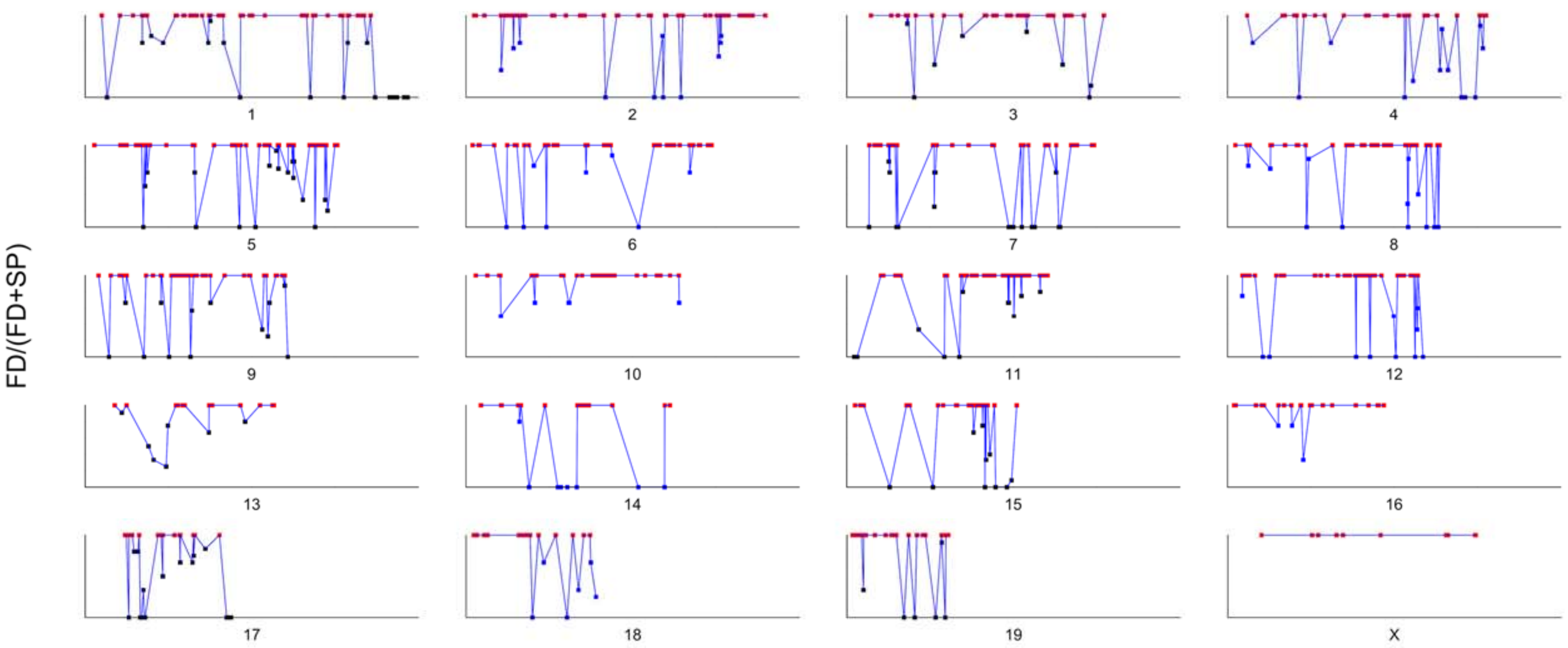


Figure S2 The ratio of fixed differences (FDs) to topologically informative sites, fixed differences and shared polymorphisms (SPs), across the genome for all pairwise comparisons of *Mus musculus* subspecies. Dots indicate the start of each region and red dots indicate fully sorted regions. (A) *M. m. castaneus* and *M. m. domesticus*. (B) *M. m. castaneus* and *M. m. musculus*. (C) *M. m. domesticus* and *M. m. musculus*.

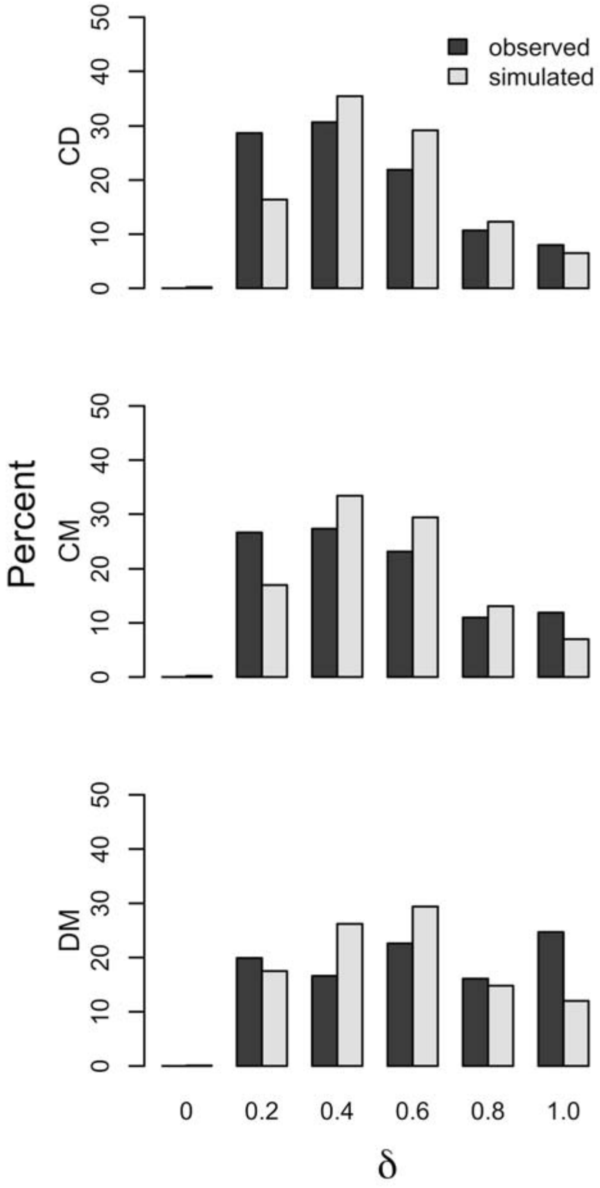


Figure S3 The distribution of values of δ in the observed data and in simulations based on demographic parameters from Supporting Information Table 7.

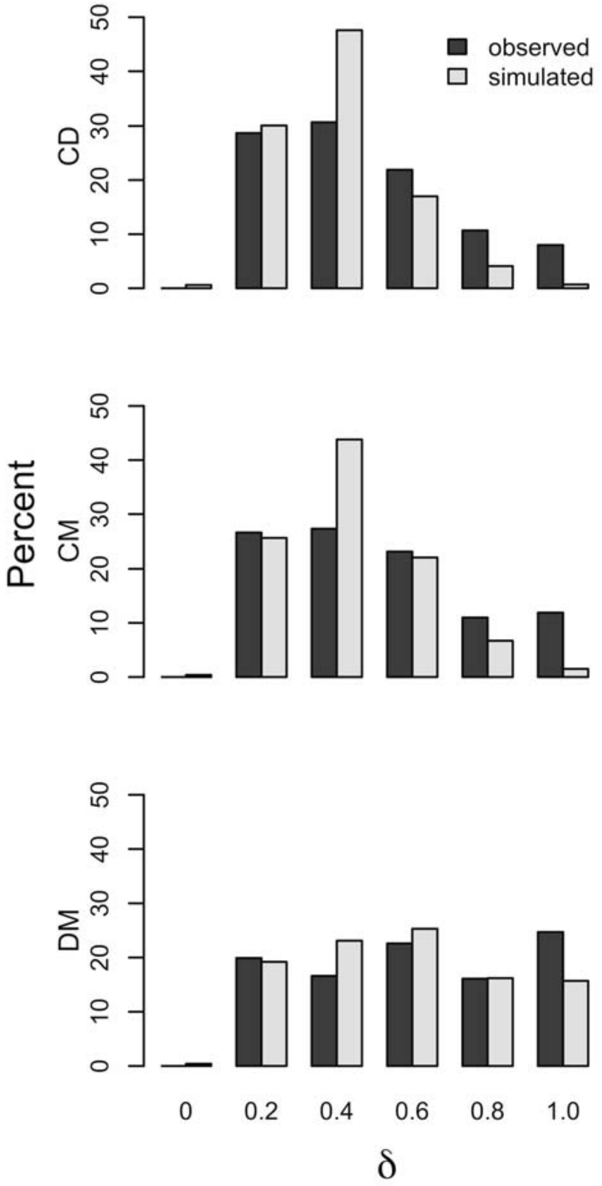


Figure S4 The distribution of values of δ in the observed data and in simulations based on demographic parameters from Supporting Information Table 9.

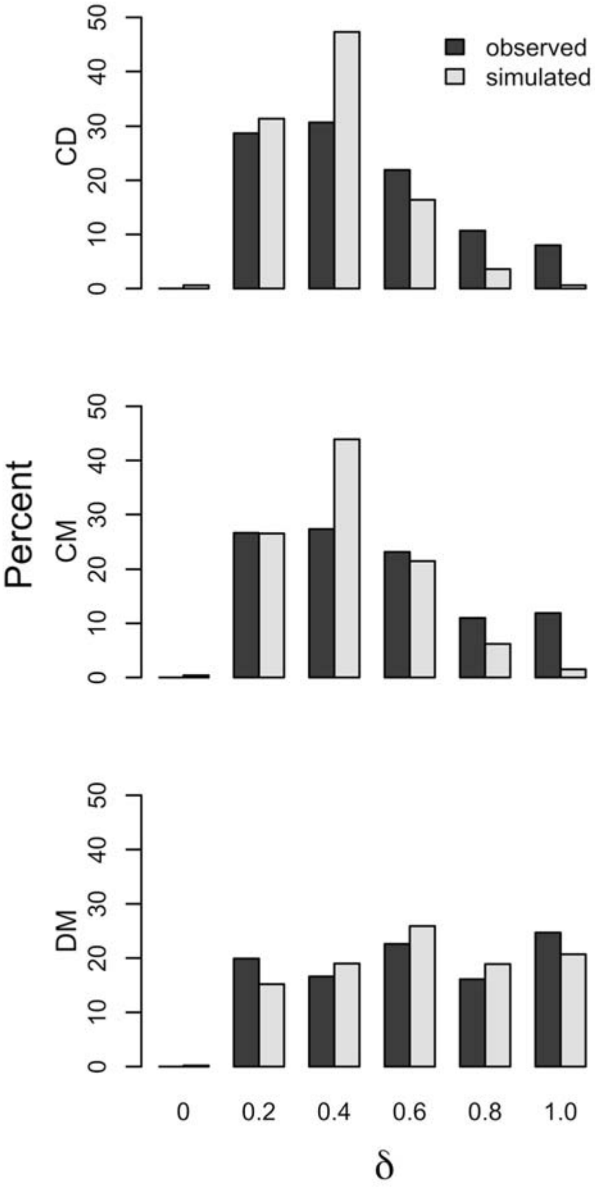


Figure S5 The distribution of values of δ in the observed data and in simulations based on demographic parameters from Supporting Information Table 7, but with a divergence time of 425 Kya.

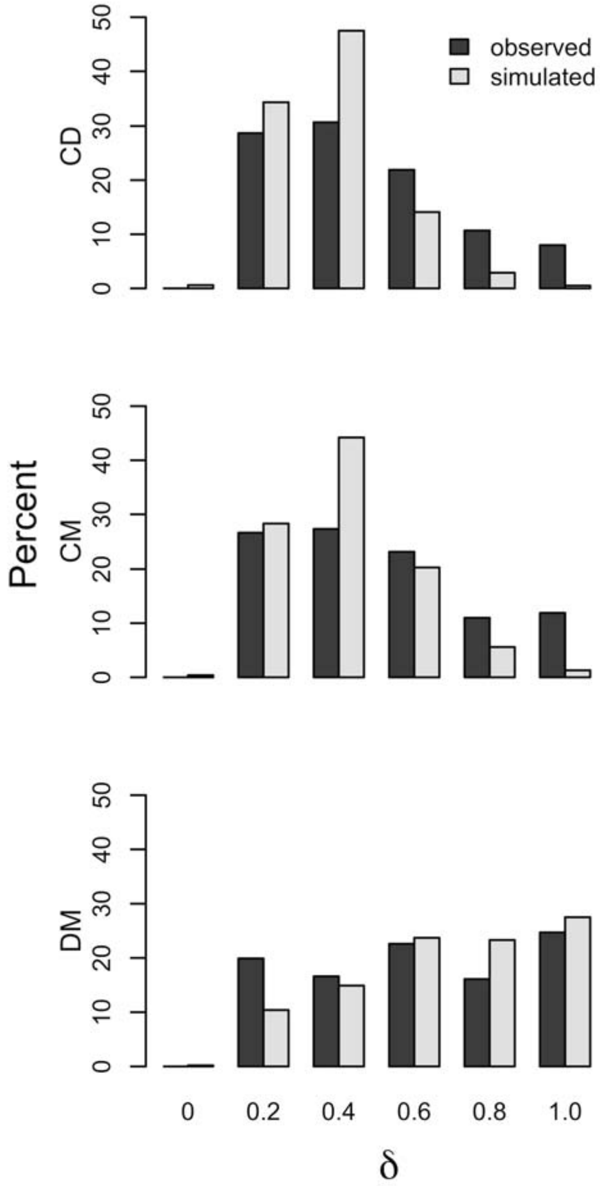


Figure S6 The distribution of values of δ in the observed data and in simulations based on demographic parameters from Supporting Information Table 7, but with a divergence time of 825 Kya.

File S1

SNP Table

Available for download as a .txt file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.166827/-/DC1

File S2

Supplementary Methods and Results

Measures of differentiation measured on a per SNP basis

δ is the absolute value of the difference in minor allele frequency among populations.

$$\delta = |\text{ Minor Allele Frequency}_{\text{pop1}} - \text{Minor Allele Frequency}_{\text{pop2}}|$$

D_{xy} can be thought of as the number of mismatches between two sets divided by the total number of comparisons between two sets.

$$D_{XY} = \frac{(\text{Minor Allele Count}_{\text{pop1}} * \text{Major Allele Count}_{\text{pop2}}) + (\text{Major Allele Count}_{\text{pop1}} * \text{Minor Allele Count}_{\text{pop2}})}{\text{Number of Alleles}_{\text{pop1}} * \text{Number of Alleles}_{\text{pop2}}}$$

F_{st} is the portion of the variance in the data that lies between two populations.

$$F_{st} = \frac{Pi_{total} - \overline{Pi}_{within}}{Pi_{total}}$$

$$Pi_{total} = \frac{\text{Minor Allele Count}_{\text{total}} * \text{Major Allele Count}_{\text{total}}}{\binom{\text{Total number of Alleles}}{2}}$$

$$Pi_{within\ for\ popk} = \frac{\text{Minor Allele Count}_{\text{popk}} * \text{Major Allele Count}_{\text{popk}}}{\binom{\text{Number of Alleles in popk}}{2}}$$

Runs of fixed differences Another approach to evaluating differentiation across the genome is to consider runs of fixed differences. When sampling is adequate, runs of fixed differences uninterrupted by shared polymorphisms, can also identify fully sorted gene genealogies. For this analysis, we only included genes that contained at least one fixed difference or shared polymorphism from each pairwise comparison. We sampled a single SNP from each gene included in the analysis. Because we were interested in identifying highly differentiated regions, to be conservative, if a gene contained fixed differences and shared polymorphisms, the SNP included in the analysis was selected from among the shared polymorphisms. On average, “pruned” SNPs included in these analyses were ~2.19 Mbs apart. Using publicly available source code, we amended the program SLIDER (McDonald 1996) to generate a distribution of runs of fixed differences based on 10,000 Monte Carlo simulations of coalescence and recombination for each pairwise comparison. In each simulation, the observed number of polymorphisms and fixed differences were distributed randomly among sites such that the number of polymorphisms and fixed differences matched the observed data. These simulations assumed a constant N_e, uniform recombination rates among adjacent sites, random union of gametes, point mutation, and silent site neutrality. We used data from chromosome two for these simulations as it had, on average, the largest number of topologically informative markers and is the second largest autosome (~182 Mb). We replicated 10,000 simulations over ten recombination parameters ranging from one to ten.

We identified many runs of fixed differences in all pairwise comparisons (Supporting Information Figures 1a, b, c). Consistent with the window analyses, we found that there were more runs of fixed differences in the DM comparison and that those runs were, on average, larger both in terms of number of SNPs and distance covered (Supporting Information Table 5). However, SLIDER analysis failed to reject the null model. Regardless of recombination rate, summary statistics for the distribution of runs did not fall in the extreme tails of results from simulations of coalescence and recombination (Supporting Information Table 6). The X chromosome was characterized by long runs of fixed differences in all three pairwise comparisons (Supporting Information Figures 1a, b, c).

References

McDonald J. H., 1996 Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol Biol Evol 13: 253–260.

Files S3-S4

Available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.166827/-/DC1

File S3 Testis Specific Expression Table

File S4 Genes identified as significantly differentially expressed in each pairwise comparison among subspecies of *M. musculus*

Table S1 Sampling localities for all wild-derived inbred laboratory strains used in this study.

Subspecies	Wild Derived Inbred Line ID	Country	Locality
M. m. castaneus	CAST/EiJ[a]	Thailand	Thonburi
	CIM/MPL	India	Masinagudi
	CKN/MPL	Kenya	Nairobi
	CKS/MPL	Kenya	Shanzu
	CTP/MPL[b]	Thailand	Pathumthani
	DKN/MPL	Kenya	Nairobi
	MDG/MPL	Madagascar	Manakasina
	MPR/MPL[b]	Pakistan	Rawalpindi
M. m. domesticus	BIK/MPL	Israel	Kefar Galim
	BZ0/MPL	Algeria	Oran
	DCP/MPL	Cyprus	Paphos
	DJO/MPL	Italy	Orcetto
	DMZ/MPL	Morrocco	Azemmour
	LEWES/EiJ	USA	Delaware
	WLA/MPL	France	Toulouse
	WSB/EiJ[a]	USA	Maryland
M. m. musculus	BID/MPL[b]	Iran	Birdjand
	CZECHII/EiJ	Czechoslovakia	
	MBK/MPL	Bulgaria	Kranevo
	MBT/MPL	Bulgaria	Général Toshevo
	MCZ/MPL	Czech Republic	Bialowieza
	MDH/MPL	Denmark	Hov
	MPB/MPL	Poland	Prague
	PWK/PhJ[c]	Czech Republic	Lhotka
M. caroli	CAROLI/EiJ	Thailand	
M. spretus	SPRET/EiJ[c]	Spain	Cadiz

[a] Data were taken from the Wellcome Trust Mouse Genomes Project.
[b] Data were excluded from further analyses due to admixture.
[c] Data from transcriptome sequencing was combined with data from the Wellcome Trust Mouse Genomes Project.

Table S2 Short read transcriptome sequencing yields in megabases for all wild-derived inbred lines included in the study.

Subspecies	Line	Sequenced	Mapped	Mapped Uniquely	6X high quality sequence
M. m. castaneus	CIM	1,330.84	712	377	16.63
	CKN	1,122.00	631	324	14.13
	CKS	869.97	453	273	12.01
	CTP	968.65	533	302	13.76
	DKN	1,189.87	671	341	14.85
	MDG	1,113.24	596	319	14.03
	MPR	1,190.66	621	334	15.41
M. m. domesticus	BIK/MPL	998.97	543	296	13.42
	BZ0/MPL	1,014.07	588	333	15.02
	DCP/MPL	1,489.47	799	380	16.55
	DJO/MPL	1,169.41	631	324	14.39
	DMZ/MPL	1,397.35	776	376	16.85
	LEWES/EiJ	3,640.53	1,585	573	22.69
	WLA/MPL	1,241.30	652	324	14.25
M. m. musculus	BID/MPL	863.31	486	288	13.3
	CZECHII/EiJ	1,237.15	716	375	16.93
	MBK/MPL	1,005.21	546	295	13.01
	MBT/MPL	1,651.32	931	444	19.16
	MCZ/MPL	1,574.79	977	474	19.86
	MDH/MPL	1,127.44	663	355	15.53
	MPB/MPL	801.8	488	287	12.85
	PWK/PhJ	1,675.66	1,048	496	21.25
M. caroli	CAROLI/EiJ	3,063.49	1,442	596	21.68
M. spretus	SPRET/EiJ	1,921.25	1,092	513	21.85

Table S3 Summary data from comparisons of genotype data in coding regions collected by this study and data collected by the Wellcome Trust.

Inbred Line	Bases in common	Mismatches	% mismatch
PWK	13,019,770	40	0.0003
SPRET	13,547,788	32	0.0002

Table S4 The results of a STRUCTURE analysis to determine the probability of different numbers of populations (K) within wild-derived inbred lines sampled from the three subspecies of *M. musculus* after the removal of lines found to be highly admixed in previous runs of STRUCTURE.

Model	K	Average ln Pr(X|K)	Pr(K)
Admixture	1	-15937.8	<0.001
Admixture	2	-12492.3	<0.001
Admixture	3[a]	-8918.2	>0.999
Admixture	4	-9648.40	<0.001
Admixture	5	-9603.07	<0.001
No admixture	1	-15932.5	<0.001
No admixture	2	-11890.1	<0.001
No admixture	3[a]	-8879.2	>0.999
No admixture	4	-10641.9	<0.001
No admixture	5	-9059.4	<0.001

[a] In these runs, the lines assigned to the three clusters were consistent with our subspecies assignment as shown in Supporting Information Table 1.

Table S5 Summary statistics describing runs of fixed differences in pairwise comparisons among subspecies of *Mus musculus*.

Subspecies 1	Subspecies 2	Chr Type	n	Avg. # SNPs/run (SD)	Max # of SNPs/run	Avg. Mb covered (SD)	Max Mb covered (SD)
M. m. castaneus	*M. m. domesticus*	Autosomes	98	3.07 (1.82)	17	3.22 (4.14)	20.71
		X	1	14 (-)	(-)	144.19	(-)
M. m. castaneus	*M. m. musculus*	Autosomes	138	4.09 (2.59)	19	5.87 (7.59)	44.23
		X	2	5.5 (2.12)	41.5	34.96 (8.75)	41.5
M. m. domesticus	*M. m. musculus*	Autosomes	144	4.55 (2.75)	19	7.17 (8.29)	40.99
		X	2	6.5 (2.54)	9	55.30 (37.51)	81.83

Table S6 Summary statistics describing runs of fixed differences on chromosome 2 in all pairwise comparisons of the subspecies of *Mus musculus* as well as the percentile rank of those statistics in 10,000 coalescent and recombination simulations.

		# of		Avg. # of SNPs/run		Max # of	
Subspecies	Subspecies	runs	Perc. Rank	(SD)	Perc. Rank	SNPs/run	Perc. Rank
M. m. castaneus	*M. m. domesticus*	9	75%	2.89 (1.54)	67%	6	73%
M. m. castaneus	*M. m. musculus*	11	30%	3.91 (2.07)	66%	8	39%
M. m. domesticus	*M. m. musculus*	12	49%	4.67 (3.31)	50%	12	52%

Table S7 Demographic parameters used in ms (Hudson 2002) simulations. All values are based on averages of estimates from Geraldes *et al.* (2011) and assume a generation length of 1 year.

Subspecies 1	Subspecies 2	$N_{e\ species\ 1}$	$N_{e\ species\ 2}$	$N_{e\ Ancestral}$	t	2Nm (species1)[a]	2Nm (species2)[b]	Avg. # SNPs surveyed in observed loci (SD)	Avg. # SNPs in simulated loci (SD)
M. m. castaneus	*M. m. domesticus*	366,700	82,600	277,800	313,800	0.193	0.000	2.61 (2.39)	5.28 (3.05)
M. m. castaneus	*M. m. musculus*	366,700	36,600	277,800	345,800	0.190	0.058	2.59 (2.36)	4.90 (2.93)
M. m. domesticus	*M. m. musculus*	82,600	36,600	277,800	320,800	0.003	0.057	2.38 (2.08)	2.24 (1.29)

[a] The effective rate at which genes enter subspecies 1 from subspecies 2.

[b] The effective rate at which genes enter subspecies 2 from subspecies 1.

Table S8 Average values of δ for different classes of sites in all pairwise comparisons between subspecies of *M. musculus*.

Subspecies 1	Subspecies 2	$\overline{\delta}_{non-synonymous}$ (SD)	$n_{\text{non-synonymous}}$	$\overline{\delta}_{synonymous}$ SD	$n_{\text{synonymous}}$	P^{a}
M. m. castaneus	*M. m. domesticus*	0.39 (0.30)	7,118	0.40 (0.29)	16,772	<0.001
M. m. castaneus	*M. m. musculus*	0.42 (0.31)	6,965	0.43 (0.30)	16,503	<0.0001
M. m. domesticus	*M. m. musculus*	0.48 (0.35)	6,740	0.54 (0.35)	14,687	<0.0001

[a] Results of *t*-tests comparing average measures of differentiation for non-synonymous and synonymous sites

Table S9 Demographic parameters used in ms (Hudson 2002) simulations intended to more closely match the number of SNPs surveyed in the observed data.

Subspecies 1	Subspecies 2	$N_{e\ species\ 1}$	$N_{e\ species\ 2}$	$N_{e\ Ancestral}$	*t*	2Nm (species1)[a]	2Nm (species2)[b]	Avg. # SNPs surveyed in observed loci (SD)	Avg. # SNPs in simulated loci (SD)
M. m. castaneus	*M. m. domesticus*	167,000	101,000	280,000	325,000	0.193	0.000	2.61 (2.39)	3.34 (1.93)
M. m. castaneus	*M. m. musculus*	167,000	89,000	280,000	325,000	0.190	0.058	2.59 (2.36)	3.28 (1.90)
M. m. domesticus	*M. m. musculus*	101,000	89,000	280,000	325,000	0.003	0.057	2.38 (2.08)	2.65 (1.52)

[a] The effective rate at which genes enter subspecies 1 from subspecies 2.

[b] The effective rate at which genes enter subspecies 2 from subspecies 1.

Table S10 Demographic parameters used in ms (Hudson 2002) simulations. Population size estimates are based on averages of estimates from Geraldes *et al.* (2011) and assume a generation length of 1 year. Gene flow was increased until the proportion of simulated loci with low average values of differentiation matched observed proportions.

						2Nm	2Nm
Subspecies 1	Subspecies 2	$N_{e\ species\ 1}$	$N_{e\ species\ 2}$	$N_{e\ Ancestral}$	*t*	(species1)[a]	(species2)[b]
M. m. castaneus	*M. m. domesticus*	366,700	82,600	277,800	325,000	1.930	0.000
M. m. castaneus	*M. m. musculus*	366,700	36,600	277,800	325,000	1.330	0.406
M. m. domesticus	*M. m. musculus*	82,600	36,600	277,800	325,000	0.045	0.855

[a] The effective rate at which genes enter subspecies 1 from subspecies 2

[b] The effective rate at which genes enter subspecies 2 from subspecies 1

Table S11 Overlap between inversions and runs of fixed differences identified between each pair of subspecies of *Mus musculus.*

Subspecies	Subspecies	# of runs of fixed differences	Observed overlap with inversions	Perc. Rank
M. m. castaneus	*M. m. domesticus*	99	36	35%
M. m. castaneus	*M. m. musculus*	140	70	54%
M. m. domesticus	*M. m. musculus*	146	80	98.5%

Table S12 Average measures of differentiation in regions containing testis specific genes and all other regions for all pairwise comparisons of *Mus musculus* subspecies.

Subspecies	Subspecies	Regions	n	$\overline{F}_{st}$ (SD)	*t*	$\overline{D}_{xy}$ (SD)	*t*	$\overline{\delta}$ (SD)	*t*
M. m. castaneus	*M. m. domesticus*	Contain testis specific genes	520	0.24 (0.16)	0.41	0.42 (0.14)	0.29	0.41 (0.14)	0.16
		All Others	1226	0.24 (0.19)		0.42 (0.16)		0.41 (0.16)	
M. m. castaneus	*M. m. musculus*	Contain testis specific genes	515	0.29 (0.19)	1.73*	0.47 (0.15)	1.47	0.46 (0.16)	1.64*
		All Others	1247	0.27 (0.20)		0.45 (0.16)		0.44 (0.17)	
M. m. domesticus	*M. m. musculus*	Contain testis specific genes	518	0.43 (0.22)	2.15*	0.56 (0.18)	1.81*	0.56 (0.18)	1.93*
		All Others	1364	0.41 (0.24)		0.55 (0.19)		0.54 (0.20)	

*P<=0.05 in 1-sided *t*-tests comparing measures from regions containing testis specific regions and all others.

Table S13 Genes identified in regions of overlap between the results of QTL mapping and our study in comparisons between *M. m. castaneus* and *M. m. domesticus.*

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000005510	2	90734791	90744984	*Ndufs3*
ENSMUSG00000005505	2	90744897	90751783	*Kbtbd4*
ENSMUSG00000005506	2	90780539	90859654	*Celf1*
ENSMUSG00000002104	2	90875777	90885886	*Rapsn*
ENSMUSG00000002102 [a]	2	90894166	90906526	*Psmc3*
ENSMUSG00000002105	2	90901948	90910574	*Slc39a13*
ENSMUSG00000002111	2	90922547	90955913	*Spi1*
ENSMUSG00000002100	2	90958301	90976673	*Mybpc3*
ENSMUSG00000040687	2	90977517	91023994	*Madd*
ENSMUSG00000002108	2	91024218	91042991	*Nr1h3*
ENSMUSG00000002103	2	91043042	91054255	*Acp2*
ENSMUSG00000002109	2	91051729	91077139	*Ddb2*
ENSMUSG00000027257	2	91096111	91104836	*Pacsin3*
ENSMUSG00000027255	2	91105131	91117088	*Arfgap2*
ENSMUSG00000027253	2	91297668	91354058	*Lrp4*
ENSMUSG00000040549	2	91366919	91460821	*Ckap5*
ENSMUSG00000027249	2	91465477	91476571	*F2*
ENSMUSG00000075040	2	91483826	91489948	*Zfp408*
ENSMUSG00000027247	2	91490017	91512483	*Arhgap1*
ENSMUSG00000040591	2	91275068	91444704	*1110051M20Rik*
ENSMUSG00000027244	2	91514775	91550733	*Atg13*
ENSMUSG00000027243	2	91551009	91561702	*Harbi1*
ENSMUSG00000040506	2	91570291	91759006	*Ambra1*
ENSMUSG00000040495	2	91762346	91769986	*Chrm4*
ENSMUSG00000027239	2	91769962	91772454	*Mdk*
ENSMUSG00000040479	2	91772981	91816021	*Dgkz*
ENSMUSG00000095332	2	91785862	91786173	*Gm9821*
ENSMUSG00000027230	2	91815044	91864659	*Creb3l1*
ENSMUSG00000058318	2	91933274	92204823	*Phf21a*
ENSMUSG00000027293	2	119914911	119980342	*Ehd4*
ENSMUSG00000050211	2	119992148	120071071	*Pla2g4e*
ENSMUSG00000070719	2	120091331	120114933	*Pla2g4d*
ENSMUSG00000046971	2	120125693	120139901	*Pla2g4f*
ENSMUSG00000027291	2	120142197	120178873	*Vps39*
ENSMUSG00000033808	2	120181045	120229852	*Tmem87a*
ENSMUSG00000062646	2	120229632	120287436	*Ganc*
ENSMUSG00000079110	2	120281755	120330649	*Capn3*
ENSMUSG00000027288	2	120332556	120389579	*Zfp106*
ENSMUSG00000027287	2	120393407	120426991	*Snap23*
ENSMUSG00000027286 [a]	2	120429974	120435256	*Lrrc57*
ENSMUSG00000027285	2	120435119	120447296	*Haus2*
ENSMUSG00000033705	2	120454862	120557633	*Stard9*
ENSMUSG00000027284	2	120541890	120675864	*Cdan1*
ENSMUSG00000090100 [a]	2	120558552	120676340	*Ttbk2*
ENSMUSG00000027272	2	120686005	120796451	*Ubr1*
ENSMUSG00000054484	2	120802753	120833588	*Tmem62*
ENSMUSG00000023572	2	120834139	120842640	*Ccndbp1*
ENSMUSG00000023216	2	120843627	120862808	*Epb4.2*
ENSMUSG00000053675	2	120871847	120911577	*Tgm5*
ENSMUSG00000079103	2	120919301	120935531	*Tgm7*

[a] indicates genes that are testis-specific.

Table S13. cont'd.

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000074890	2	120954043	120966434	*Lcmt2*
ENSMUSG00000027259	2	120966164	120982416	*Adal*
ENSMUSG00000050619	2	120984009	120996861	*Zscan29*
ENSMUSG00000027263	2	120996390	121024506	*Tubgcp4*
ENSMUSG00000043909	2	121019017	121097143	*Trp53bp1*
ENSMUSG00000027254	2	121115336	121136568	*Map1a*
ENSMUSG00000033526	2	121136297	121181132	*Ppip5k1*
ENSMUSG00000000308	2	121183450	121189473	*Ckmt1*
ENSMUSG00000033498	2	121189464	121212904	*Strc*
ENSMUSG00000033486 [a]	2	121218367	121240317	*Catsper2*
ENSMUSG00000027248	2	121239511	121264423	*Pdia3*
ENSMUSG00000027246	2	121264746	121270014	*Ell3*
ENSMUSG00000046110	2	121264795	121282517	*Serinc4*
ENSMUSG00000074884	2	121274931	121284049	*Serf2*
ENSMUSG00000027245	2	121279026	121284408	*Hypk*
ENSMUSG00000048222	2	121285971	121299803	*Mfap1b*
ENSMUSG00000068479	2	121317647	121332401	*Mfap1a*
ENSMUSG00000027242	2	121332459	121370596	*Wdr76*
ENSMUSG00000027238	2	121371265	121632823	*Frmd5*
ENSMUSG00000060227	2	121692706	121761956	*Casc4*
ENSMUSG00000074881	2	121779488	121781096	*Mageb3*
ENSMUSG00000033411	2	121781737	121839378	*Ctdspl2*
ENSMUSG00000027236 [a]	2	121854282	121882334	*Eif3j1*
ENSMUSG00000033396	2	121879256	121944122	*Spg11*
ENSMUSG00000027233	2	121945844	122011925	*Patl2*
ENSMUSG00000060802	2	121973422	121978819	*B2m*
ENSMUSG00000033368 [a]	2	121986436	122004763	*Trim69*
ENSMUSG00000027229	2	122012008	122032133	*4933406J08Rik*
ENSMUSG00000027227 [a]	2	122060485	122091076	*Sord*
ENSMUSG00000068452	2	122104983	122124185	*Duox2*
ENSMUSG00000027225	2	122124636	122128621	*Duoxa2*
ENSMUSG00000027224	2	122127927	122139466	*Duoxa1*
ENSMUSG00000033268	2	122141408	122173708	*Duox1*
ENSMUSG00000033256	2	122174628	122194898	*Shf*
ENSMUSG00000027219	2	122251126	122286873	*Slc28a2*
ENSMUSG00000079071	2	122310677	122353776	*Gm14085*
ENSMUSG00000073889	4	41647021	41716347	*Il11ra1*
ENSMUSG00000028447	4	41661830	41670202	*Dctn3*
ENSMUSG00000066224	4	41670868	41678174	*Arid3c*
ENSMUSG00000036078	4	41685366	41703030	*Sigmar1*
ENSMUSG00000036073	4	41702101	41705998	*Galt*
ENSMUSG00000073888 [a]	4	41716340	41721120	*Ccl27a*
ENSMUSG00000073884	4	41774204	41775337	*Ccl21b*
ENSMUSG00000096543	4	41870187	41870612	*Gm21966*
ENSMUSG00000094065	4	41903610	41904743	*Gm21541*
ENSMUSG00000078747	4	41941572	41943124	*Gm20878*
ENSMUSG00000078746	4	41966058	41971856	*Gm20938*
ENSMUSG00000096256	4	42033017	42034726	*Gm21093*
ENSMUSG00000095611	4	42035113	42035538	*Gm10597*
ENSMUSG00000095881	4	42083899	42084291	*Gm21968*
ENSMUSG00000094293	4	42091207	42092287	*Gm3893*
ENSMUSG00000073878	4	42114817	42115917	*Gm13304*
ENSMUSG00000073877 [a]	4	42153436	42158839	*Gm13306*

[a] indicates genes that are testis-specific.

Table S13. cont'd.

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000073876	4	42158842	42168603	*Gm13305*
ENSMUSG00000096609	4	42170845	42171335	*1700045I11Rik*
ENSMUSG00000094984	4	42219428	42219853	*Gm10595*
ENSMUSG00000083929	4	42240639	42242685	*Gm10600*
ENSMUSG00000095675	4	42255767	42256432	*Ccl21b*
ENSMUSG00000094695	4	42294267	42294855	*Gm21953*
ENSMUSG00000093996	4	42318334	42323929	*Gm21598*
ENSMUSG00000095234	4	42439378	42439966	*Gm21586*
ENSMUSG00000096892	4	42458751	42459176	*Gm10597*
ENSMUSG00000093909	4	42459563	42461272	*Gm3883*
ENSMUSG00000095779	4	42466752	42589938	*Gm2163*
ENSMUSG00000094066	4	42522580	42528175	*Gm13298*
ENSMUSG00000096260	4	42581229	42581621	*Gm10592*
ENSMUSG00000096596	4	42612195	42612860	*Gm10591*
ENSMUSG00000091938	4	42629719	42631714	*Gm2564*
ENSMUSG00000096826 [a]	4	42655251	42656005	*Ccl27b*
ENSMUSG00000078735	4	42656355	42661893	*Il11ra2*
ENSMUSG00000094731	4	42668043	42668438	*Gm9969*
ENSMUSG00000095375	4	42714926	42719893	*Gm21955*
ENSMUSG00000054885	4	42735545	42846248	*4930578G10Rik*
ENSMUSG00000071005	4	42754525	42756577	*Ccl19*
ENSMUSG00000094686	4	42772860	42773993	*Ccl21a*
ENSMUSG00000078722	4	42781928	42856771	*Gm12394*
ENSMUSG00000078721	4	42848071	42853888	*Gm12429*
ENSMUSG00000050141 [a]	4	42868004	42874234	*BC049635*
ENSMUSG00000036062	4	42916660	42944752	*N28178*
ENSMUSG00000028551	4	109660876	109667189	*Cdkn2c*
ENSMUSG00000010517 [a]	4	109676588	109963960	*Faf1*
ENSMUSG00000029722	5	137650483	137684726	*Agfg2*
ENSMUSG00000045348	5	137730883	137741607	*Nyap1*
ENSMUSG00000029723 [a]	5	137745730	137768450	*Tsc22d4*
ENSMUSG00000029659	5	149411749	149431723	*Medag*
ENSMUSG00000029660 [a]	5	149439706	149470620	*Tex26*
ENSMUSG00000029658	5	149528679	149611894	*Wdr95*
ENSMUSG00000033174	6	88724412	88828360	*Mgll*
ENSMUSG00000030083	6	88835915	88841935	*Abtb1*
ENSMUSG00000033152	6	88842558	88875044	*Podxl2*
ENSMUSG00000030314	6	114643097	114860614	*Atg7*
ENSMUSG00000030315	6	114860628	114969994	*Vgll4*
ENSMUSG00000030316	6	115004381	115037876	*Tamm41*
ENSMUSG00000009394	6	115134902	115282626	*Syn2*
ENSMUSG00000092004	6	115227343	115259294	*Gm17482*
ENSMUSG00000030317	6	115245616	115251849	*Timp4*
ENSMUSG00000000440	6	115361221	115490401	*Pparg*
ENSMUSG00000042389	6	115544664	115578350	*Tsen2*
ENSMUSG00000068011	6	115583544	115592576	*2510049J12Rik*
ENSMUSG00000000439	6	115601938	115618670	*Mkrn2*
ENSMUSG00000000441	6	115618067	115676635	*Raf1*
ENSMUSG00000055396	6	115675995	115677136	*D830050J10Rik*
ENSMUSG00000059900	6	115729131	115762466	*Tmem40*
ENSMUSG00000030319	6	115774538	115804893	*Cand2*
ENSMUSG00000071226	6	120666369	120771190	*Cecr2*
ENSMUSG00000004902	6	120773768	120793982	*Slc25a18*
ENSMUSG00000019210	6	120795245	120822685	*Atp6v1e1*

[a] indicates genes that are testis-specific.

Table S13. cont'd.

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000009112	6	120836230	120892842	*Bcl2l13*
ENSMUSG00000004446	6	120891930	120916853	*Bid*
ENSMUSG00000051586	6	120931707	121003153	*Mical3*
ENSMUSG00000003178	6	121007241	121081609	*Mical3*
ENSMUSG00000030143	6	132361041	132364134	*Gm8882*
ENSMUSG00000059934	6	132569809	132572941	*Prh1*
ENSMUSG00000058295	6	132595913	132601236	*Prp2*
ENSMUSG00000067541	6	132625111	132627511	*A630073D07Rik*
ENSMUSG00000059382	6	132656957	132657844	*Tas2r120*
ENSMUSG00000071150	6	132700090	132701007	*Tas2r121*
ENSMUSG00000078280	6	132710999	132711928	*Tas2r122*
ENSMUSG00000071149	6	132737010	132738035	*Tas2r115*
ENSMUSG00000060412	6	132754730	132755659	*Tas2r124*
ENSMUSG00000056901	6	132762131	132763174	*Tas2r102*
ENSMUSG00000053217	6	132777179	132778162	*Tas2r136*
ENSMUSG00000058349	6	132802818	132803975	*Tas2r117*
ENSMUSG00000057381	6	132847142	132848143	*Tas2r123*
ENSMUSG00000030194	6	132855438	132856355	*Tas2r116*
ENSMUSG00000062952	6	132868008	132869009	*Tas2r110*
ENSMUSG00000056926	6	132893011	132893940	*Tas2r113*
ENSMUSG00000059410	6	132909651	132910587	*Tas2r125*
ENSMUSG00000063762	6	132951102	132952064	*Tas2r129*
ENSMUSG00000057699	6	132956884	132957919	*Tas2r131*
ENSMUSG00000062528	6	132980015	132980965	*Tas2r109*
ENSMUSG00000030196	6	133036163	133037101	*Tas2r103*
ENSMUSG00000071147	6	133054817	133055816	*Tas2r140*
ENSMUSG00000072704	6	133105239	133107747	*2700089E24Rik*
ENSMUSG00000055594	6	133292216	133295790	*5530400C23Rik*
ENSMUSG00000095412	6	133529189	133532762	*Gm5885*
ENSMUSG00000032758	6	133849855	133853667	*Kap*
ENSMUSG00000030199	6	134035700	134270158	*Etv6*
ENSMUSG00000030200	6	134396318	134438736	*Bcl2l14*
ENSMUSG00000035919 [a]	9	22475715	22888280	*Bbs9*
ENSMUSG00000020052	10	87490819	87493660	*Ascl1*
ENSMUSG00000020051	10	87521795	87584136	*Pah*
ENSMUSG00000020053	10	87858265	87937042	*Igf1*
ENSMUSG00000035383	10	88091072	88092375	*Pmch*
ENSMUSG00000035365 [a]	10	88091432	88146941	*Parpbp*
ENSMUSG00000035351 [a]	10	88146992	88178388	*Nup37*
ENSMUSG00000020056 [a]	10	88201093	88246158	*Ccdc53*
ENSMUSG00000020057	10	88322804	88379080	*Dram1*
ENSMUSG00000035311	10	88379132	88447329	*Gnptab*
ENSMUSG00000060002	10	88452745	88504073	*Chpt1*
ENSMUSG00000020059 [a]	10	88459569	88473236	*Sycp3*
ENSMUSG00000020061	10	88518279	88605152	*Mybpc1*
ENSMUSG00000004359	10	88674772	88685015	*Spic*
ENSMUSG00000060904	10	88730858	88744094	*Arl1*
ENSMUSG00000004356	10	88746607	88826814	*Utp20*
ENSMUSG00000020062	10	88885992	88929505	*Slc5a8*
ENSMUSG00000035189	10	88948994	89344762	*Ano4*
ENSMUSG00000074802	10	89408823	89443967	*Gas2l3*
ENSMUSG00000047638	10	89454234	89533585	*Nr1h4*
ENSMUSG00000019935	10	89574020	89621253	*Slc17a8*
ENSMUSG00000019906 [a]	10	107271843	107425143	*Lin7a*

[a] indicates genes that are testis-specific.

Table S13. cont'd.

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000000435	10	107482908	107486134	*Myf5*
ENSMUSG00000035923	10	107492860	107494729	*Myf6*
ENSMUSG00000035916	10	107517360	107720027	*Ptprq*
ENSMUSG00000091455	10	107762223	107912134	*Otogl*
ENSMUSG00000019907	10	108162400	108277575	*Ppp1r12a*
ENSMUSG00000035873	10	108332189	108414391	*Pawr*
ENSMUSG00000035864	10	108497650	109010982	*Syt1*
ENSMUSG00000020181	10	109682660	110000219	*Nav3*

[a]indicates genes that are testis-specific.

Table S14 Genes identified in regions of overlap between the results of QTL mapping, a study of the hybrid zone, and our study in comparisons between *M. m. musculus* and *M. m. domesticus.*

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000040152	2	118111876	118127133	*Thbs1*
ENSMUSG00000027344 [a]	2	118204888	118256966	*Fsip1*
ENSMUSG00000040133	2	118277110	118373419	*Gpr176*
ENSMUSG00000005102	2	118388618	118475234	*Eif2ak4*
ENSMUSG00000009549 [a]	2	118475850	118479711	*Srp14*
ENSMUSG00000040093	2	118528757	118549687	*Bmf*
ENSMUSG00000040084	2	118598211	118641591	*Bub1b*
ENSMUSG00000074923	2	118663303	118698020	*Pak6*
ENSMUSG00000078137	2	118699103	118703963	*Ankrd63*
ENSMUSG00000040061	2	118707517	118728438	*Plcb2*
ENSMUSG00000045838	2	118754158	118762661	*A430105I19Rik*
ENSMUSG00000046804	2	118772769	118778165	*Phgr1*
ENSMUSG00000040035	2	118779719	118811293	*Disp2*
ENSMUSG00000027331	2	118814003	118853957	*Knstrn*
ENSMUSG00000027332	2	118861954	118882909	*Ivd*
ENSMUSG00000040007	2	118900377	118924528	*Bahd1*
ENSMUSG00000074916	2	118926497	118928585	*Chst14*
ENSMUSG00000039983	2	119017779	119029393	*Ccdc32*
ENSMUSG00000027324	2	119034790	119039769	*Rpusd2*
ENSMUSG00000027326 [a]	2	119047119	119105501	*Casc5*
ENSMUSG00000027323 [a]	2	119112793	119147445	*Rad51*
ENSMUSG00000070730	2	119137001	119157034	*Rmdn3*
ENSMUSG00000046814	2	119167773	119172390	*Gchfr*
ENSMUSG00000034278	2	119172500	119208795	*Dnajc17*
ENSMUSG00000055926	2	119174509	119177575	*Gm14137*
ENSMUSG00000068580 [a]	2	119208617	119217049	*Zfyve19*
ENSMUSG00000027317	2	119218119	119229906	*Ppp1r14d*
ENSMUSG00000027315	2	119237362	119249527	*Spint1*
ENSMUSG00000034226 [a]	2	119269201	119271272	*Rhov*
ENSMUSG00000034216	2	119288740	119298453	*Vps18*
ENSMUSG00000027314	2	119325784	119335962	*Dll4*
ENSMUSG00000027313	2	119351229	119354381	*Chac1*
ENSMUSG00000034154	2	119373042	119477687	*Ino80*
ENSMUSG00000048647	2	119516505	119547627	*Exd1*
ENSMUSG00000014077	2	119547697	119587027	*Chp1*
ENSMUSG00000072980	2	119609512	119618469	*Oip5*
ENSMUSG00000027306	2	119618298	119651244	*Nusap1*
ENSMUSG00000027305	2	119655446	119662827	*Ndufaf1*
ENSMUSG00000027304	2	119675068	119735407	*Rtf1*
ENSMUSG00000027296	2	119742337	119751263	*Itpka*
ENSMUSG00000027297	2	119751320	119760431	*Ltk*
ENSMUSG00000034032	2	119763304	119787537	*Rpap1*
ENSMUSG00000027298	2	119797733	119818104	*Tyro3*
ENSMUSG00000028524	4	102741297	102973628	*Sgip1*
ENSMUSG00000028523 [a]	4	102986379	103005594	*Tctex1d1*
ENSMUSG00000066090	4	103017872	103026842	*Insl5*
ENSMUSG00000035126 [a]	4	103038065	103114555	*Wdr78*
ENSMUSG00000028522 [a]	4	103114390	103165754	*Mier1*
ENSMUSG00000028521	4	103170649	103215164	*Slc35d1*

[a] indicates genes that are testis-specific.

Table S14. cont’d.

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000028520[a]	4	103230445	103290863	*4921539E11Rik*
ENSMUSG00000035069	4	103313812	103371868	*Oma1*
ENSMUSG00000028519	4	103619359	104744844	*Dab1*
ENSMUSG00000070886	4	104328252	104330557	*Gm10304*
ENSMUSG00000029656	4	104766317	104804548	*C8b*
ENSMUSG00000035031	4	104815679	104876398	*C8a*
ENSMUSG00000095386	4	104857329	104859137	*Gm17662*
ENSMUSG00000078612	4	104913456	105016863	*1700024P16Rik*
ENSMUSG00000028518	4	105029874	105109890	*Prkaa2*
ENSMUSG00000028517	4	105157347	105232764	*Ppap2b*
ENSMUSG00000029705	5	136248135	136567490	*Cux1*
ENSMUSG00000046548	5	136613702	136615328	*4731417B20Rik*
ENSMUSG00000005474[a]	5	136693146	136701094	*Myl10*
ENSMUSG00000004415	5	136741759	136883209	*Col26a1*
ENSMUSG00000007987[a]	5	136908150	136913244	*Rabl5*
ENSMUSG00000019054	5	136953275	136966234	*Fis1*
ENSMUSG00000001739	5	136966616	136975858	*Cldn15*
ENSMUSG00000059518[a]	5	136982164	136988021	*Znhit1*
ENSMUSG00000004846	5	136987019	136996648	*Plod3*
ENSMUSG00000037428	5	137030295	137033351	*Vgf*
ENSMUSG00000004849	5	137034993	137046135	*Ap1s1*
ENSMUSG00000037411	5	137061504	137072272	*Serpine1*
ENSMUSG00000043279	5	137105644	137116209	*Trim56*
ENSMUSG00000037390[a]	5	137134924	137149320	*Muc3*
ENSMUSG00000079174	5	137154030	137166001	*Gm3054*
ENSMUSG00000094840	5	137208813	137212389	*A630081J09Rik*
ENSMUSG00000023328	5	137287519	137294466	*Ache*
ENSMUSG00000051502	5	137294669	137295664	*Ufsp1*
ENSMUSG00000037364[a]	5	137295704	137307674	*Srrt*
ENSMUSG00000023348	5	137309899	137314241	*Trip6*
ENSMUSG00000037344	5	137314558	137333597	*Slc12a9*
ENSMUSG00000029710	5	137350109	137378669	*Ephb4*
ENSMUSG00000079173	5	137378637	137477064	*Zan*
ENSMUSG00000029711	5	137483020	137533242	*Epo*
ENSMUSG00000029715	5	137501438	137502518	*Pop7*
ENSMUSG00000029714	5	137518880	137527934	*Gigyf1*
ENSMUSG00000029713	5	137528127	137533510	*Gnb2*
ENSMUSG00000029712	5	137553517	137569582	*Actl6b*
ENSMUSG00000029716	5	137569851	137587481	*Tfr2*
ENSMUSG00000037221	5	137596645	137601058	*Mospd3*
ENSMUSG00000029718	5	137605103	137613784	*Pcolce*
ENSMUSG00000089984[a]	5	137612503	137629002	*Fbxo24*
ENSMUSG00000093445	5	137629121	137641099	*Lrch4*
ENSMUSG00000029720	5	137629175	137642899	*Gm20605*
ENSMUSG00000079165	5	137641334	137642902	*Sap25*
ENSMUSG00000047182	5	137643032	137645714	*Irs3*
ENSMUSG00000029722	5	137650483	137684726	*Agfg2*
ENSMUSG00000045348	5	137730883	137741607	*Nyap1*
ENSMUSG00000029723[a]	5	137745730	137768450	*Tsc22d4*
ENSMUSG00000029725[a]	5	137778849	137780110	*Ppp1r35*
ENSMUSG00000029726[a]	5	137781906	137786715	*Mepce*
ENSMUSG00000037108	5	137787798	137822621	*Zcwpw1*
ENSMUSG00000046245	5	137821952	137836268	*Pilra*

[a]indicates genes that are testis-specific.

Table S14. cont’d.

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000066684	5	137852147	137858049	*Pilrb1*
ENSMUSG00000066682	5	137865829	137871758	*Pilrb2*
ENSMUSG00000029727	5	137892932	137921619	*Cyp3a13*
ENSMUSG00000056966	5	137953809	137962959	*Gjc3*
ENSMUSG00000037053	5	137981521	137990233	*Azgp1*
ENSMUSG00000075599	5	138021276	138034665	*Smok3a*
ENSMUSG00000079156	5	138021429	138050636	*Smok3b*
ENSMUSG00000029729	5	138085084	138107822	*Zkscan1*
ENSMUSG00000037017 [a]	5	138116903	138134265	*Zscan21*
ENSMUSG00000037007	5	138139702	138155744	*Zfp113*
ENSMUSG00000019494	5	138161071	138164646	*Cops6*
ENSMUSG00000029730 [a]	5	138164583	138172422	*Mcm7*
ENSMUSG00000019518	5	138172002	138178708	*Ap4m1*
ENSMUSG00000036980 [a]	5	138178617	138187451	*Taf6*
ENSMUSG00000036968	5	138187485	138193918	*Cnpy4*
ENSMUSG00000049285	5	138194314	138195621	*Mblac1*
ENSMUSG00000089783	5	138203609	138207308	*Gm454*
ENSMUSG00000047592	5	138225898	138253363	*Nxpe5*
ENSMUSG00000050552	5	138255608	138259398	*Lamtor4*
ENSMUSG00000036948	5	138259656	138264046	*BC037034*
ENSMUSG00000091964	5	138259658	138264046	*BC037034*
ENSMUSG00000075593	5	138264921	138272840	*Gal3st4*
ENSMUSG00000029510	5	138264952	138280005	*Gpc2*
ENSMUSG00000036928 [a]	5	138280240	138312393	*Stag3*
ENSMUSG00000075591	5	138363719	138388287	*Gm10874*
ENSMUSG00000036898	5	138441468	138460694	*Zfp157*
ENSMUSG00000029526	5	138561840	138564694	*1700123K08Rik*
ENSMUSG00000058291	5	138604616	138619761	*Zfp68*
ENSMUSG00000056014	5	138622859	138648903	*A430033K04Rik*
ENSMUSG00000025854	5	138754514	138810077	*Fam20c*
ENSMUSG00000094504	5	138820080	138821619	*Gm5294*
ENSMUSG00000025856	5	138976014	138997370	*Pdgfa*
ENSMUSG00000075585	5	138995056	139000576	*6330403L08Rik*
ENSMUSG00000025855	5	139017306	139150001	*Prkar1b*
ENSMUSG00000025857	5	139150223	139186510	*Heatr2*
ENSMUSG00000036817	5	139200637	139249840	*Sun1*
ENSMUSG00000025858	5	139252324	139270051	*Get4*
ENSMUSG00000056413	5	139271876	139325622	*Adap1*
ENSMUSG00000045438	5	139336189	139345233	*Cox19*
ENSMUSG00000029541	5	139352617	139357033	*Cyp2w1*
ENSMUSG00000053553	5	139359739	139460502	*3110082I17Rik*
ENSMUSG00000044197	5	139377742	139396415	*Gpr146*
ENSMUSG00000021206	5	139378220	139379259	*D830046C22Rik*
ENSMUSG00000044092	5	139405280	139415623	*C130050O18Rik*
ENSMUSG00000053647	5	139423151	139427800	*Gper1*
ENSMUSG00000053581	5	139471211	139484549	*Zfand2a*
ENSMUSG00000029546	5	139543494	139548179	*Uncx*
ENSMUSG00000036718	5	139706693	139736336	*Micall2*
ENSMUSG00000029547	5	139751282	139775678	*Ints1*
ENSMUSG00000018143	5	139791513	139802653	*Mafk*
ENSMUSG00000036687 [a]	5	139802485	139819917	*Tmem184a*
ENSMUSG00000098140	5	139807978	139826407	*Gm26938*
ENSMUSG00000029551	5	139823592	139826885	*Psmg3*

[a] indicates genes that are testis-specific.

Table S14. cont'd.

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000048988	5	139907943	139974711	*Elfn1*
ENSMUSG00000031737	8	92357796	92361456	*Irx5*
ENSMUSG00000031738	8	92674289	92680956	*Irx6*
ENSMUSG00000031740	8	92827328	92853417	*Mmp2*
ENSMUSG00000033192	8	92855350	92919279	*Lpcat2*
ENSMUSG00000078144	8	92901395	92902409	*Capns2*
ENSMUSG00000055368	8	92960079	93001667	*Slc6a2*
ENSMUSG00000071047	8	93020214	93048192	*Ces1a*
ENSMUSG00000078964	8	93056727	93080017	*Ces1b*
ENSMUSG00000057400	8	93099015	93131283	*Ces1c*
ENSMUSG00000056973	8	93166068	93197838	*Ces1d*
ENSMUSG00000061959	8	93201218	93229619	*Ces1e*
ENSMUSG00000031725	8	93256236	93279747	*Ces1f*
ENSMUSG00000057074	8	93302369	93337308	*Ces1g*
ENSMUSG00000074156	8	93351843	93363676	*Ces1h*
ENSMUSG00000058019	8	93499213	93535707	*Ces5a*
ENSMUSG00000031748	8	93809966	93969388	*Gnao1*
ENSMUSG00000031751	8	93971588	94012663	*Amfr*
ENSMUSG00000031754 [a]	8	94017770	94037021	*Nudt21*
ENSMUSG00000033009	8	94037198	94067921	*Ogfod1*
ENSMUSG00000031755	8	94067954	94098811	*Bbs2*
ENSMUSG00000031757	8	94137204	94139031	*Mt4*
ENSMUSG00000031760	8	94152607	94154148	*Mt3*
ENSMUSG00000031762	8	94172618	94173567	*Mt2*
ENSMUSG00000031765	8	94179089	94180325	*Mt1*
ENSMUSG00000032939	8	94214597	94315066	*Nup93*
ENSMUSG00000031766	8	94329192	94366213	*Slc12a3*
ENSMUSG00000031770	8	94386438	94395377	*Herpud1*
ENSMUSG00000074151	8	94472763	94527272	*Nlrc5*
ENSMUSG00000034361	8	94532990	94570529	*Cpne2*
ENSMUSG00000031774	8	94574943	94601726	*Fam192a*
ENSMUSG00000050079 [a]	8	94601955	94660275	*Rspry1*
ENSMUSG00000031776 [a]	8	94666755	94674417	*Arl2bp*
ENSMUSG00000031775	8	94674895	94696242	*Pllp*
ENSMUSG00000031779	8	94745590	94751699	*Ccl22*
ENSMUSG00000031778	8	94772009	94782423	*Cx3cl1*
ENSMUSG00000031780	8	94810453	94812035	*Ccl17*
ENSMUSG00000031781	8	94819818	94838358	*Ciapin1*
ENSMUSG00000031782	8	94838321	94854895	*Coq9*
ENSMUSG00000031783	8	94857450	94864242	*Polr2c*
ENSMUSG00000040631	8	94863828	94876312	*Dok4*
ENSMUSG00000063605	8	94902869	94918098	*Ccdc102a*
ENSMUSG00000061577	8	94923694	94943290	*Gpr114*
ENSMUSG00000031785	8	94977109	95014208	*Gpr56*
ENSMUSG00000022295	15	38661904	38692443	*Atp6v1c1*
ENSMUSG00000022296	15	38933142	38949405	*Baalc*
ENSMUSG00000022297	15	39006280	39038186	*Fzd6*
ENSMUSG00000054196	15	39076932	39087121	*Cthrc1*
ENSMUSG00000022299	15	39094191	39112716	*Slc25a32*
ENSMUSG00000022300	15	39112874	39146856	*Dcaf13*
ENSMUSG00000037386	15	39198332	39681940	*Rims2*
ENSMUSG00000022303	15	39745932	39760934	*Dcstamp*
ENSMUSG00000022304	15	39768485	39857470	*Dpys*
ENSMUSG00000022305	15	39870603	39943994	*Lrp12*

[a] indicates genes that are testis-specific.

Table S14. cont'd.

Ensembl Gene ID	Chr	Gene Start (bp)	Gene End (bp)	Associated Gene Name
ENSMUSG00000094112	15	40142188	40148689	*9330182O14Rik*
ENSMUSG00000022306	15	40655042	41104592	*Zfpm2*
ENSMUSG00000022307	15	41447482	41861048	*Oxr1*
ENSMUSG00000042895	15	41865293	41869720	*Abra*
ENSMUSG00000022309	15	42424727	42676977	*Angpt1*
ENSMUSG00000051920	15	43020811	43170818	*Rspo2*
ENSMUSG00000022336	15	43250040	43282736	*Eif3e*
ENSMUSG00000072592	15	43430943	43477036	*Gm10373*
ENSMUSG00000022337	15	43477229	43527777	*Emc2*
ENSMUSG00000054409	15	43866695	43870029	*Tmem74*
ENSMUSG00000048915	17	62604184	62881317	*Efna5*
ENSMUSG00000090425	17	62604292	62606707	*Efna5*
ENSMUSG00000023965 [a]	17	63057452	63500017	*Fbxl17*
ENSMUSG00000045506	17	63863300	63863791	*A930002H24Rik*
ENSMUSG00000000127	17	63896018	64139494	*Fer*
ENSMUSG00000024083	17	64281005	64331916	*Pja2*
ENSMUSG00000073377	17	64514081	64555660	*AU016765*
ENSMUSG00000024085	17	64600736	64755110	*Man2a1*
ENSMUSG00000024088	17	64832523	64836071	*4930583I09Rik*
ENSMUSG00000045036 [a]	17	65256005	65540782	*Tmem232*
ENSMUSG00000024091	17	65580056	65613555	*Vapa*
ENSMUSG00000050612 [a]	17	65637505	65642204	*Txndc2*
ENSMUSG00000056515	17	65651726	65772752	*Rab31*
ENSMUSG00000061950 [a]	17	65782573	65841926	*Ppp4r1*
ENSMUSG00000024096	17	65848433	65885755	*Ralbp1*
ENSMUSG00000024098	17	65923066	65951187	*Twsg1*
ENSMUSG00000034647	17	65967501	66077089	*Ankrd12*
ENSMUSG00000024099 [a]	17	66078795	66101559	*Ndufv2*
ENSMUSG00000024101	17	66111546	66120503	*Wash*
ENSMUSG00000035842 [a]	17	66123520	66152167	*Ddx11*
ENSMUSG00000052105	17	66336982	66449750	*Soga2*
ENSMUSG00000023460	17	66494512	66519717	*Rab12*
ENSMUSG00000024105	17	66555252	66594621	*Themis3*

[a]indicates genes that are testis-specific.