Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination

A. GERALDES,* P. BASSET,† K. L. SMITH and M. W. NACHMAN

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA

Abstract

In the early stages of reproductive isolation, genomic regions of reduced recombination are expected to show greater levels of differentiation, either because gene flow between species is reduced in these regions or because the effects of selection at linked sites within species are enhanced in these regions. Here, we study the patterns of DNA sequence variation at 27 autosomal loci among populations of *Mus musculus musculus*, *M. m. domesticus*, and *M. m. castaneus*, three subspecies of house mice with collinear genomes. We found that some loci exhibit considerable shared variation among subspecies, while others exhibit fixed differences. We used an isolation-with-gene-flow model to estimate divergence times and effective population sizes (N_e) and to disentangle ancestral variation from gene flow. Estimates of divergence time indicate that all three subspecies diverged from one another within a very short period of time approximately 350 000 years ago. Overall, N_e for each subspecies was associated with the degree of genetic differentiation: *M. m. musculus* had the smallest N_e and the greatest proportion of monophyletic gene genealogies, while *M. m. castaneus* had the largest N_e and the smallest proportion of monophyletic gene genealogies. *M. m. domesticus* and *M. m. musculus* were more differentiated from each other than either were from *M. m. castaneus*, consistent with greater reproductive isolation between *M. m. domesticus* and *M. m. musculus*. F_{ST} was significantly greater at loci experiencing low recombination rates compared to loci experiencing high recombination rates in comparisons between *M. m. castaneus* and *M. m. musculus* or *M. m. domesticus*. These results provide evidence that genomic regions with less recombination show greater differentiation, even in the absence of chromosomal rearrangements.

Keywords: divergence, gene flow, polymorphism, recombination, speciation

Received 22 June 2011; revision received 3 August 2011; accepted 15 August 2011

Introduction

Understanding how new species arise is a central problem in evolutionary genetics. Complex patterns of genetic variation are expected among recently diverged lineages, and these patterns may be governed by both stochastic and deterministic processes (e.g. Maroja *et al.*

Correspondence: Michael Nachman, Fax: 520 621 9190;
E-mail: nachman@u.arizona.edu
*Present address: Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4.
†Present address: Hospital Preventive Medicine Service, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland.

2009). For example, neutral polymorphisms are expected to be shared among daughter populations as a simple consequence of the persistence of polymorphisms present in the ancestral population. Shared polymorphisms may also derive from secondary contact and gene flow between daughter populations. Both processes (persistence of ancestral variation and subsequent gene flow) may occur simultaneously, and while distinguishing between them is difficult, a variety of analytic tools now exist for assessing their relative importance (e.g. Nielsen & Wakeley 2001; Hey & Nielsen 2004; Becquet & Przeworski 2007; Hey & Nielsen 2007). Selection can also shape the distribution of variation among young lineages, and this can occur in several ways and

affect individual regions of the genome differently. Positive directional selection on individual loci may reduce variation within one or both daughter populations, leading to increased differentiation for those loci. In genomic regions with low recombination rates, the effects of selection at linked sites (owing to genetic hitchhiking or background selection) can also lead to increased differentiation (Charlesworth 1998). Gene flow between daughter populations may be reduced at some loci if particular alleles reduce fitness in the sister lineage, either as a result of epistasis or selection in a different environment (reviewed in Coyne & Orr 2004).

Studying patterns of genetic variation among newly arising lineages can thus shed light on the history of speciation for a particular group, including the relative importance of demography, gene flow and selection in shaping observed patterns. By studying many loci throughout the genome, it may be possible to identify particular genomic regions that are important in isolating newly arising species.

Regions of low recombination are of particular interest and are expected to show increased levels of differentiation for two different reasons. First, several speciation models have suggested that gene flow between species may be reduced in such regions (Noor et al. 2001; Rieseberg 2001; Navarro & Barton 2003; Faria & Navarro 2010). Noor et al. (2001) model is based on the asymmetric nature of Dobzhansky–Muller incompatibilities and the difficulty of selecting against such mutations when two incompatibilities with opposite asymmetries are locked in a single nonrecombining region. The model of Navarro & Barton (2003) emphasizes the accumulation of coadapted genes within nonrecombining regions, while the model of Rieseberg (2001) emphasizes the accumulated effect of multiple genes contributing to isolation when they occur in a single nonrecombining region. Second, the effects of selection at linked sites will extend over larger distances in regions of low recombination. Positive selection and associated genetic hitchhiking (Smith & Haigh 1974) as well as background selection (Charlesworth et al. 1993) can reduce levels of genetic variation within species and thereby lead to an inflation of measures like F_{ST} between species (Charlesworth 1998). Distinguishing reduced gene flow between species from selection within species as causes of increased differentiation is a difficult problem (Noor & Bennet 2009). Noor & Bennet (2009) suggest that coalescent models in which gene flow is estimated in a nonequilibrium context may help solve this problem, and we utilize this approach here.

The house mouse, Mus musculus, is arguably the best mammalian model for studies of the genetics of speciation. It consists of three subspecies with parapatric distributions: M. m. musculus is found in Eastern Europe and Northern Asia, M. m. castaneus is found in Southeast Asia, and M. m. domesticus is native to the Near East, Northern Africa and Western Europe and has been introduced to the Americas, Africa, and many oceanic islands in association with humans during historical times. These three lineages are young; the available data suggest that they diverged in allopatry within the last 500 000 years (e.g. Boursot et al. 1993; Geraldes et al. 2008a). Regions of secondary contact exist where the subspecies meet in nature. The best studied of these is a narrow hybrid zone between M. m. musculus and M. m. domesticus that stretches from Denmark to the Black Sea in Central Europe (e.g. Raufaste et al. 2005; Macholan et al. 2007; Teeter et al. 2008, 2010). Hybrids have also been extensively studied in the laboratory. M. m. musculus and M. m. domesticus are reproductively isolated primarily by hybrid male sterility (e.g. Forejt 1996; Britton-Davidian et al. 2005). The genetic basis of this sterility is complex and maps in part to the X chromosome (e.g. Storchova et al. 2004; Good et al. 2008). The involvement of the X chromosome in reproductive isolation is also suggested by patterns of reduced gene flow between subspecies on the X chromosome compared to the autosomes (Tucker et al. 1992). Recently, the first gene for hybrid male sterility in a vertebrate was identified in crosses between mice that were primarily derived from M. m. musculus and M. m. domesticus (Mihola et al. 2009). Finally, most classical inbred strains of laboratory mice derive from crosses between house mouse subspecies (Frazer et al. 2007; Yang et al. 2007). Thus, the genetic tools and genomic resources for laboratory mice can be applied to the study of mouse speciation, including a genome sequence, expression databases, individual gene knockouts, and many other resources and methods.

Despite our impressive understanding of mouse genetics in general, we still know remarkably little about the distribution of genetic variation in natural populations of the three house mouse subspecies. Our chief goal here is to fill this gap by providing a genome-wide picture of patterns of differentiation among M. m. musculus, M. m. castaneus, and M. m. domesticus in the context of models of speciation. We resequenced 23 autosomal loci and used these data with four loci previously sequenced in the same populations to estimate the levels of differentiation and gene flow among mouse subspecies in different regions of the genome.

Materials and methods

Samples

We included 27 Mus musculus domesticus from Western Europe, 26 M. m. musculus from Eastern Europe, and

Fig. 1 Approximate location of populations sampled in this study. Blue indicates *Mus musculus domesticus*, red indicates *M. m. musculus* and orange indicates *M. m. castaneus*. Sample sizes, sampling localities names and geographic coordinates are indicated in Table S1 (Supporting information).

Table 1 Loci surveyed, GC content, gene density and recombination rate

Locus	Chromosome	Start position (bp)[*]	GC content (%)[†]	Gene density/Mb[‡]	Recombination rate (cM/Mb)[§]
Atp6v1h	1	5 090 103	0.394	2.9	0.082
Chrng[¶]	1	89 102 229	0.455	9.7	0.364
Fbxo28	1	184 268 340	0.449	8.0	0.898
Ndufa8	2	35 893 440	0.438	13.9	0.321
Med19[¶]	2	84 522 848	0.388	25.3	0.174
Atp5e	2	174 287 027	0.448	9.9	1.232
Zfhx4	3	5 223 313	0.381	2.1	0.072
Ssr3	3	65 187 464	0.403	5.1	0.280
Prpf3[¶]	3	95 653 041	0.443	19.9	0.431
Fpgt	3	154 752 158	0.415	4.2	0.735
Clcn6[¶]	4	147 391 868	0.400	15.2	0.600
Acot7	4	151 573 815	0.496	17.1	1.112
Sfrs8	5	130 008 014	0.458	7.3	1.203
Cmas	6	142 713 639	0.425	6.5	1.153
Nomo1	7	53 333 140	0.466	22.1	0.479
Usp10	8	122 476 548	0.496	13.3	1.203
Ncapd3	9	26 842 239	0.416	5.4	0.948
Rab21	10	114 735 869	0.404	6.3	1.107
Slc39a11	11	113 294 508	0.459	15.8	1.283
Golga5	12	103 717 894	0.451	10.0	0.915
Iars	13	49 816 700	0.448	8.0	0.644
Rgs7bp	13	105 749 379	0.407	6.4	0.835
Atxn10	15	85 226 251	0.486	14.8	0.741
Lrpprc	17	85 144 001	0.445	6.6	0.987
Fbxo38	18	62 704 865	0.441	8.6	0.780
Mamdc2	19	23 518 001	0.426	5.5	1.162
Shoc2	19	54 090 657	0.431	5.2	1.178

[*]Position in NCBI mouse builds 37 of the sequenced region.

[†]In a 10-Mb window centred around the start position of the sequenced region.

[‡]In a 10-Mb window centred around the start position of the sequenced region. Only protein-coding genes were considered.

[§]Recombination rates were calculated for 10-Mb windows centred on the sequenced region by regressing the genetic position of the markers (Shifman *et al.* 2006; Cox *et al.* 2009) in the genetic map against their physical position on mouse NCBI build 37.

[¶]Data from Geraldes *et al.* 2008a;.

27 *M. m. castaneus* from India (Fig. 1 and Table S1, Supporting information). All mice were collected at least 300 m apart to avoid the sampling of related individuals. Mice from India were kindly provided by B. Harr. One *M. caroli* and one *M. spretus* were purchased from the Jackson laboratory and used as outgroups.

Selection of loci and molecular methods

We PCR-amplified and sequenced mostly intronic portions of 23 autosomal loci (Table 1 and Fig. 2). Loci were chosen to cover a wide range of recombination rates. Recombination rates were calculated by regressing marker positions on the genetic map (Shifman *et al.* 2006; Cox *et al.* 2009) against physical positions on the mouse genome sequence (NCBI build 37) for a 10-Mb window centred on each locus. These recombination rates are derived from crosses involving inbred strains that are largely of *M. m. domesticus* origin. We assume that recombination rates are broadly similar across the three subspecies (Dumont *et al.* 2011). We also assume that recombination rates estimated over 10-Mb distances are a reasonable proxy for local recombination rates. To the extent that this is not true, we will have less power to detect true differences among regions differing in recombination rate. The genomes of these three subspecies are collinear at the level of cytogenetic resolution.
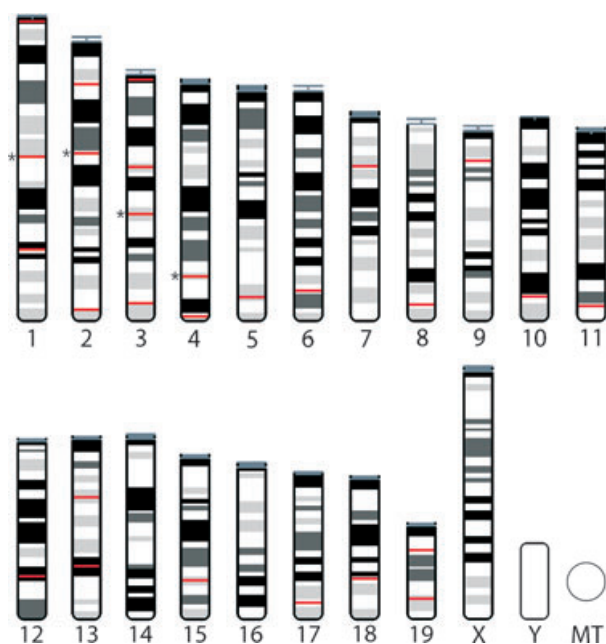
Fig. 2 Location of the loci surveyed in this study. Red lines indicate the approximate location of each locus on the ideogram of the house mouse karyotype. Exact location and locus details are provided in Table 1. Asterisks mark four loci from Geraldes *et al.* (2008a) reanalysed in this study.

We cannot exclude the possibility that small inversion differences exist between subspecies. For each locus, we amplified two overlapping fragments and sequenced both. This strategy provided at least two-fold coverage for every base, typically with one sequence from each strand. It also enabled us to detect rare instances of allele-specific PCR. In many cases, PCR primers were placed in exons to amplify intervening introns. PCR and sequencing primers are listed in Table S2 (Supporting information). Sanger sequencing was performed on an ABI 3700 automated sequencer by the University of Arizona's Genomic Analysis and Technology Core facility. An additional four autosomal loci previously sequenced in the same populations were included (Geraldes *et al.* 2008a).

Sequence assembly and editing

Sequences were trimmed to exclude all exonic regions; all data analysed are from introns. Assembly and editing of contigs was performed with phred/phrap/consed/polyphred (Nickerson *et al.* 1997; Ewing & Green 1998; Ewing *et al.* 1998; Gordon *et al.* 1998). Alignments for each locus were generated with Clustal X (Thompson *et al.* 1994) and checked manually with BioEdit (Hall 1999). All indel polymorphisms were excluded from further analyses. For each locus, individuals and sites with more than 10% missing data were excluded from further analyses. All resulting alignments were deposited in Genbank under accession numbers JF336572–JF338137. Haplotypes were inferred with Phase 2.1.1 (Stephens *et al.* 2001; Stephens & Donnelly 2003) after checking for convergence of three independent runs.

Population genetic analyses

For each locus and each subspecies, we estimated π, the average number of pairwise differences between sequences (Nei & Li 1979), and θ, the proportion of segregating sites (Watterson 1975). Divergence between each subspecies and *Mus caroli* was estimated with D_{XY} (Nei 1987), the average pairwise divergence. Deviations from the expected frequency spectrum of polymorphisms under a neutral model were assessed with Tajima's *D* (Tajima 1989) and Fu and Li's *D* (Fu & Li 1993). These analyses were performed using the program SITES (Wakeley & Hey 1997). Statistical significance for Tajima's *D* (TD) and Fu and Li's *D* (FLD) was assessed after 1000 coalescent simulations with parameters estimated from the data for each locus using the program HKA (http://genfaculty.rutgers.edu/hey/software). Evolutionary relationships among haplotypes were inferred using the neighbor-joining method

(Saitou & Nei 1987) in MEGA4 (Kumar *et al.* 2008). This was done using all of the data as well as the largest nonrecombining segment for each locus. Trees were rooted with *M. caroli*, and bootstrap values were calculated after 1000 replicates.

Differentiation among subspecies

Levels of differentiation between pairs of subspecies were assessed in several ways. First, we used the program SITES (Wakeley & Hey 1997) to estimate F_{ST} (Wright 1951) using equation 3 of Hudson *et al.* (1992). We conducted computer simulations to ask whether the mean and range of observed F_{ST} values were greater than expected under a simple model of divergence without gene flow, using the program Make Sample (Hudson 2002). We simulated gene genealogies for three populations, with current population sizes of 36 600, 82 600 and 366 000 that diverged instantaneously 326 000 years ago from an ancestral population with a size of 277 000. The parameters used for these simulations were based on the IMa results. The geometric mean mutation rate across all loci was also taken from IMa analyses. We simulated 1000 data sets with 27 gene genealogies each. We calculated the mean and range of F_{ST} values in these simulated data sets and then compared these simulated values to the observed values.

Nei (1973) noted that measures of differentiation such as F_{ST} are heavily influenced by levels of genetic variation within subpopulations, and Charlesworth (1998) pointed out that forces that reduce variation within subpopulations, such as background selection or genetic hitchhiking, will increase F_{ST}. Thus, we calculated D_{XY} between each pair of subspecies, a measure that is independent of within-subspecies diversity. We also calculated D_A, or net nucleotide divergence corrected for within-subspecies diversity (Nei 1987). Comparison between D_A and D_{XY} may help to distinguish between competing hypotheses (i.e. reduced gene flow or diversity-reducing selection) for high levels of differentiation in some genomic regions (Noor & Bennet 2009).

The measures of genetic differentiation described above provide a summary of the data but cannot be used to directly make inferences about processes such as gene flow, except under a number of simplifying assumptions (Whitlock & McCauley 1999). For example, $F_{ST} = 1/(4Nm + 1)$, where Nm is the number of migrants per generation, assumes that populations are at equilibrium with respect to migration and drift, a condition that is almost certainly not met between subspecies of house mice. One solution to this problem is to use more complex models that relax some of these assumptions (Noor & Bennet 2009). In particular,

coalescent-based methods have been developed to distinguish on-going gene flow from ancestral shared variation in the context of recently diverging populations (e.g. Hey & Nielsen 2004; Becquet & Przeworski 2007; Hey & Nielsen 2007). Here, we used IM and IMa, which implement a Markov Chain Monte Carlo method for analysis of genetic data under an isolation-with-migration model (Nielsen & Wakeley 2001; Hey & Nielsen 2004, 2007), to obtain maximum-likelihood estimates of population sizes, divergence times and migration rates under nonequilibrium conditions. This model assumes that there is no recombination within loci and free recombination between loci. We used IMgc (Woerner *et al.* 2007) to obtain the longest region within each locus without four gametic types. Estimates from IMa have been shown to be robust when data sets are trimmed to apparently nonrecombining blocks (Strasburg & Rieseberg 2010). Using this nonrecombining data set (Table S3, Supporting information), we used IMa to compare different models of population divergence. The full isolation-with-migration model has six parameters: the effective population size of each contemporary population (N_{e1} and N_{e2}), the ancestral population size (N_{ea}), the time since divergence (t) and migration rates between populations in each direction ($2Nm_1$ and $2Nm_2$). These parameters are estimated jointly in the model. Thus, estimates of gene flow take into account past and current N_e, and the model does not require that populations be at migration–drift equilibrium. We first estimated parameters for each of the three subspecies in pairwise comparisons. Then, using a likelihood ratio test, we compared the log likelihood of our data under this fully parameterized model to the log likelihood of our data under nested models in which migration was set to zero in both directions (no gene flow), or in each direction separately (fully asymmetric gene flow). For each analysis, we ran the program under Metropolis Coupled Monte Carlo Markov Chains using at least 28 chains with a two-step heating scheme and parameters that allowed for proper chain swapping. We ran the program for at least two million steps. For each analysis, we checked for convergence between two replicates, and we present results from just one replicate of each analysis. The geometric mean of the mutation rate of all loci was used to convert the estimated parameters (θ and t, which are scaled to the mutation rate) to demographic parameters (N_e and t). We estimated mutation rates per year for each locus assuming that the divergence to *M. caroli* represents 4.3 MY (Suzuki *et al.* 2004). Estimating the mutation rate per generation further requires knowledge of the number of generations per year. Wild mice are expected to have between one and two generations per year (Bronson 1979; Geraldes *et al.* 2008a). As the precise

generation time is unknown, we give estimates of N_e assuming both one and two generations per year. Finally we used IM to estimate per locus migration rates into each subspecies (except for *M. m. domesticus*, as a model with no gene flow into this subspecies was not significantly different from the fully parameterized model). Migration rates per locus are given in terms of the joint parameter 2Nm.

Results

Overall levels of genetic differentiation among subspecies

Gene genealogies for 27 autosomal loci sampled among 26 *M. m. musculus*, 27 *M. m. castaneus* and 27 *M. m. domesticus* are shown in Fig. S1 (Supporting information) and are summarized in Fig. 3. We observed different genealogical patterns among different loci, as expected for taxa at an early stage of divergence. For some loci such as *Prpf3*, alleles corresponding to the three subspecies were fully sorted (i.e. each was monophyletic, with alleles within each subspecies more closely related to other alleles within that subspecies than to any alleles in other subspecies). For other loci, subspecies were intermingled on the gene genealogy and were either paraphyletic such as *Rab21* (i.e. with one subspecies nested within another) or polyphyletic such as *Golga5* (with multiple unrelated lineages for each subspecies). We compared each subspecies to each other subspecies and calculated the pro-

portion of gene trees that were reciprocally monophyletic, paraphyletic or polyphyletic (Fig. 3a). In comparisons between *M. m. domesticus* and *M. m. musculus* (DM), a greater proportion of loci were fully sorted than in comparisons between *M. m. domesticus* and *M. m. castaneus* (DC) or between *M. m. musculus* and *M. m. castaneus* (MC) (Fig. 3a), indicating that DM are more differentiated than DC or MC. The same pattern was observed when only the largest nonrecombining blocks for each locus were used to generate genealogies (Fig. S2, Supporting information). Similar patterns were seen using only nodes with high bootstrap support (>80%; Fig. S1, Supporting information).

Levels of population differentiation between pairs of subspecies were assessed using F_{ST} for sequence data (Hudson *et al.* 1992). F_{ST} was significantly higher for DM (average $F_{ST} = 0.72$) than for DC (average $F_{ST} = 0.43$; Wilcoxon signed-rank test $P < 0.0001$) or MC (average $F_{ST} = 0.49$; Wilcoxon signed-rank test $P < 0.0001$) (Table 2), consistent with the observed genealogical patterns (Fig. 3). Higher overall differentiation for DM could be due to lower levels of gene flow, smaller effective population sizes or some combination of these factors.

To explore these issues, we fitted the polymorphism data to a model of divergence-with-gene-flow (IMa) (Nielsen & Wakeley 2001; Hey & Nielsen 2004, 2007) and obtained maximum-likelihood estimates (MLE) of demographic parameters. These analyses indicate that DM started to diverge about 321 000 years ago, DC about 314 000 years ago and MC about 346 000 years
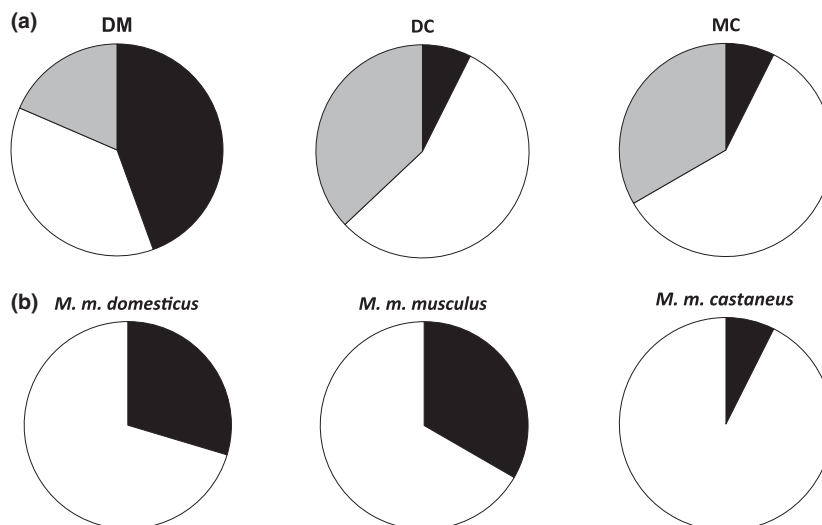
Fig. 3 Summary of genealogical relationships among subspecies of house mouse for the 27 loci surveyed in this study. (a) Proportion of loci that are reciprocally monophyletic (black), paraphyletic (grey) or polyphyletic (white), in pairwise comparisons among subspecies. (b) proportion of loci where each subspecies is either monophyletic (black) or not (white), in relation to the other two subspecies.

Table 2 Estimates of population differentiation and gene flow among house mouse subspecies

| Locus | Rec. Rate‡ | F_{ST} | | | Fixed/total* | | | 2Nm† | | | | | |
| | | | | | | | | Into *Mus musculus musculus* | | | Into *Mus musculus castaneus* | | |
		DM	DC	MC	DM	DC	MC	From *M. m. domesticus*	From *M. m. castaneus*	Total	From *M. m. domesticus*	From *M. m. musculus*	Total
Zfhx4	0.072	0.75	0.62	0.80	0.44	0.14	0.36	0.00	0.00	0.00	0.01	0.02	0.03
Atp6v1h	0.082	0.50	0.40	0.27	0.00	0.00	0.00	0.34	0.00	0.34	0.95	1.39	2.34
Med19	0.174	0.80	0.73	0.70	0.17	0.00	0.04	0.00	0.00	0.00	0.27	0.02	0.29
Ssr3	0.280	0.81	0.39	0.55	0.33	0.02	0.00	0.01	0.00	0.01	0.36	0.61	0.97
Ndufa8	0.321	0.83	0.83	0.80	0.50	0.07	0.02	0.01	0.00	0.01	0.01	0.02	0.03
Chrng	0.364	0.69	0.32	0.45	0.13	0.00	0.03	0.00	0.00	0.00	0.01	0.02	0.03
Prpf3	0.431	0.94	0.75	0.90	0.74	0.08	0.46	0.00	0.01	0.01	0.01	0.02	0.03
Nomo1	0.479	0.89	0.54	0.46	0.36	0.00	0.00	0.00	0.20	0.20	0.56	0.02	0.58
Clcn6	0.600	0.50	0.38	0.18	0.00	0.01	0.00	0.15	0.54	0.69	0.01	0.05	0.06
Iars	0.644	0.41	0.26	0.37	0.00	0.00	0.00	0.00	0.01	0.01	0.30	0.64	0.94
Fpgt	0.735	0.86	0.45	0.30	0.11	0.00	0.00	0.00	0.00	0.00	0.53	2.90	3.43
Atxn10	0.741	0.93	0.39	0.40	1.08	0.00	0.00	0.00	0.00	0.00	0.16	1.39	1.55
Fbxo38	0.780	0.92	0.58	0.36	1.14	0.01	0.00	0.00	0.10	0.10	0.04	0.95	0.99
Rgs7bp	0.835	0.88	0.38	0.55	0.44	0.00	0.03	0.00	0.00	0.00	0.67	0.20	0.87
Fbxo28	0.898	0.63	0.28	0.63	0.08	0.00	0.00	0.00	0.01	0.01	0.13	0.52	0.65
Golga5	0.915	0.48	0.20	0.32	0.00	0.00	0.00	0.29	0.35	0.64	1.47	0.02	1.49
Ncapd3	0.948	0.62	0.41	0.42	0.08	0.00	0.00	0.00	0.02	0.02	0.98	0.02	1.00
Lrpprc	0.987	0.94	0.36	0.58	0.54	0.00	0.03	0.00	0.00	0.00	1.15	0.02	1.17
Rab21	1.107	0.55	0.27	0.48	0.09	0.00	0.04	0.00	0.00	0.00	0.01	0.02	0.03
Acot7	1.112	0.84	0.46	0.62	0.31	0.00	0.07	0.00	0.01	0.01	0.10	0.49	0.59
Cmas	1.153	0.67	0.32	0.27	0.00	0.00	0.00	0.21	0.16	0.37	0.58	0.64	1.22
Mamdc2	1.162	0.87	0.60	0.67	0.35	0.05	0.06	0.00	0.00	0.00	0.01	0.02	0.03
Shoc2	1.178	0.77	0.64	0.39	0.00	0.00	0.00	0.27	0.11	0.38	0.36	0.02	0.38
Usp10	1.203	0.77	0.52	0.61	0.00	0.01	0.00	0.21	0.24	0.45	0.07	0.17	0.24
Sfrs8	1.203	0.57	0.06	0.47	0.00	0.00	0.00	0.35	0.08	0.43	6.68	0.02	6.70
Atp5e	1.232	0.47	0.19	0.32	0.00	0.00	0.00	0.01	0.10	0.11	1.98	0.02	2.00
Slc39a11	1.283	0.49	0.32	0.30	0.00	0.00	0.00	0.37	0.00	0.37	0.01	1.24	1.25
Average Low (8)	0.275	0.78	0.57	0.62	0.32	0.03	0.07	0.05	0.03	0.07	0.27	0.26	0.54
Average High (19)	0.985	0.69	0.37	0.43	0.15	0.00	0.01	0.10	0.09	0.19	0.80	0.49	1.29
Average all (27)	0.775	0.72	0.43	0.49	0.19	0.01	0.02	0.08	0.07	0.15	0.65	0.42	1.07

*Ratio of fixed differences between subspecies to all other polymorphisms (shared, among subspecies and exclusive to either one).

†M locus-specific estimates of the effective rate at which genes come into each subspecies from the gene pool of the other.

‡Recombination rates were calculated for 10-Mb windows centred on the sequenced region by regressing the genetic position of the markers (Shifman et al. 2006; Cox et al. 2009) in the genetic map against their physical position on mouse NCBI build 37.

Table 3 Maximum-likelihood estimates (MLE) and 90% posterior density intervals (in parentheses) of demographic parameters obtained with IMa for subspecies of house mice. Estimates of effective population size (N_e) assuming a generation length of 1 and 0.5 years are given

Generation length	Subspecies 1	Subspecies 2	$N_{e\ \text{species 1}}$	$N_{e\ \text{species 2}}$	$N_{e\ \text{Ancestral}}$	t*	2Nm(species1)[†]	2Nm(species 2)[‡]	Average 2Nm
1 year	*domesticus*	*musculus*	82 552 (67 951–97 153)	38 096 (28 679–49 226)	272 365 (133 094–408 266)	320 764	0.003 (0.000–0.023)	0.057 (0.031–0.093)	0.030
	domesticus	*castaneus*	82 653 (67 552–97 575)	370 264 (322 590–427 472)	220 887 (166 857–284 451)	313 822	0.000 (0.000–0.025)	0.193 (0.088–0.336)	0.097
	musculus	*castaneus*	35 060 (24 672–46 314)	363 101 (312 802–419 689)	190 196 (111 603–278 221)	345 752 (247 268–372 981)	0.058 (0.027–0.106)	0.190 (0.078–0.346)	0.124
0.5 years	*domesticus*	*musculus*	165 104 (135 902–194 306)	76 193 (57 359–98 451)	544 730 (266 187–816 533)				
	domesticus	*castaneus*	165 126 (135 103–195 149)	740 527 (645 180–854 943)	441 774 (333 714–568 903)				
	musculus	*castaneus*	70 120 (49 344–92 628)	726 202 (625 603–839 377)	380 391 (223 205–556 441)				

In cases where the posterior density distribution failed to return to zero, confidence intervals cannot be reliably estimated and were left blank.

*Number of years since species begun to diverge.

[†]The effective rate at which genes come into the subspecies one gene pool from subspecies two.

[‡]The effective rate at which genes come into the subspecies two gene pool from subspecies one.

ago (Table 3). Although the uncertainty in these estimates is fairly high [e.g., 90% of the posterior probability density (HPD90) for the estimate of divergence time between DC falls between 247 000 and 373 000 years ago], these analyses indicate that all three pairs of subspecies started to diverge at approximately the same time.

Higher differentiation for DM compared to DC or MC could arise if *M. m. domesticus* and *M. m. musculus* have lower N_e than *M. m. castaneus*. Genetic drift will be greater in smaller populations, leading to greater changes in allele frequencies from the ancestral population. Levels of nucleotide polymorphism (π) were significantly lower in *M. m. domesticus* (average π = 0.213%) and in *M. m. musculus* (average π = 0.166%) compared to *M. m. castaneus* (average π = 0.585%; Wilcoxon signed-rank tests $P < 0.0001$ for both), suggesting that N_e is higher in *M. m. castaneus* (Table 4). Maximum-likelihood estimates of N_e were obtained with IMa (Table 3). N_e was approximately 82 500 for *M. m. domesticus*, approximately 36 500 for *M. m. musculus* and approximately 366 500 for *M. m. castaneus*, assuming a generation time of one year. If a generation time of 0.5 years is assumed, all values are twice as large (Table 3). These estimates of N_e are in good agreement with the proportion of monophyletic gene genealogies for each subspecies (Fig. 3b). *M. m. musculus* has the greatest proportion of monophyletic loci and the smallest N_e, while *M. m. castaneus* has the smallest proportion of monophyletic loci and the largest N_e. Thus, it seems likely that genetic drift plays an important role in shaping genome-wide differentiation between house mouse subspecies.

Differences in migration rate among subspecies could also lead to differences in the levels of overall differentiation. Maximum-likelihood estimates of gene flow between each pair of subspecies were generally low (Table 3). Nonetheless, nested models in which gene flow was constrained to be 0 in one or both directions were rejected as significantly worse than models with gene flow in both directions using likelihood ratio tests except for DM, where a model without gene flow could not be rejected (Table 5). The average migration rate into each population was lower for DM (2Nm = 0.030), than for DC (2Nm = 0.097) or MC (2Nm = 0.124). Gene flow into *M. m. castaneus* was higher from both *M. m. domesticus* (2Nm = 0.193) and from *M. m. musculus* (2Nm = 0.190) than gene flow into *M. m. domesticus* (2Nm = 0.003 from *M. m. musculus* and 2Nm = 0.000 from *M. m. castaneus*) or into *M. m. musculus* (2Nm = 0.057 from *M. m. domesticus* and 2Nm = 0.058 from *M. m. castaneus*). Thus, low migration into *M. m. domesticus* and *M. m. musculus* may also contribute to the high overall differentiation observed for DM.

Heterogeneity among loci in the levels of differentiation and variation in recombination rate

F_{ST} showed considerable variation among loci. F_{ST} ranged from 0.06 to 0.83 for DC, from 0.18 to 0.90 for MC and from 0.41 to 0.94 for DM (Table 2 and Fig. 4). To determine whether the mean and range of F_{ST} values were higher than expected under a simple model without gene flow, we performed coalescent simulations of 1000 sets of 27 gene genealogies and compared

Table 4 Average levels of polymorphism within subspecies of house mice and divergence (D_{XY}) between these and *Mus caroli*, for the eight low and 19 high recombination loci and the entire data set

Locus	Subspecies	n[†]	L[‡]	S[§]	π (%)	θ (%)	Tajima's D[¶]	Fu and Li's D[¶]	D_{XY} (%)	θ/D_{XY}
Average low (8)	*Mus musculus domesticus*	58	1867	15	0.139	0.179	−0.5667	−0.4564	3.408	0.052
	M. m. musculus	34	1865	7	0.087	0.100	−0.2222	−0.4227	3.470	0.029
	M. m. castaneus	43	1833	41	0.470	0.540	−0.7017*	−0.4791	3.393	0.159
Average high (19)	*M. m. domesticus*	53	1653	16	0.244	0.227	−0.2493	−0.1305	3.886	0.058
	M. m. musculus	48	1652	12	0.199	0.168	0.3950*	0.5877**	3.918	0.043
	M. m. castaneus	42	1634	48	0.634	0.716	−0.4240	−0.0968	3.863	0.185
Average all (27)	*M. m. domesticus*	55	1717	16	0.213	0.213	−0.3370	−0.2271	3.744	0.059
	M. m. musculus	44	1716	11	0.166	0.148	0.2121	0.2883*	3.785	0.041
	M. m. castaneus	42	1693	46	0.585	0.664	−0.5063**	−0.2101	3.724	0.183

*0.05> P > 0.01; **0.01> P > 0.001; ***P < 0.001
[†]Number of chromosomes analysed.
[‡]Number of bp analysed.
[§]Number of segregating sites.
[¶]Statistical significance for Tajima's D and Fu and Li's D was assessed after 10 000 coalescent simulations under the program HKA.

Table 5 Likelihood ratio tests (2LLR) of nested IMa models against the fully parameterized (five parameters) isolation-with-migration model. Presented are the results for symmetric migration rates (four parameters), unidirectional gene flow (four parameters) and no gene flow (three parameters)

Subspecies comparison	Model	log (P)	2LLR	P-value
Mus musculus domesticus – Mus musculus musculus	mD[*] = mM[†]	1.4328	11.2443	0.0008
	mD mM = 0	−28.9782	72.0654	<0.0001
	mD = 0 mM	6.8006	0.5079	0.4760
	mD = mM = 0	−28.9777	72.0644	<0.0001
M. m. domesticus–M. m. castaneus	mD = mC[‡]	5.0164	4.6939	0.0303
	mD mC = 0	−18.4373	51.6011	<0.0001
	mD = 0 mC	7.3601	0.0065	0.9357
	mD = mC = 0	−26.3559	67.4384	<0.0001
M. m. musculus–M. m. castaneus	mM = mC	3.8046	2.8861	0.0894
	mM mC = 0	−4.2169	18.9291	<0.0001
	mM = 0 mC	−12.1585	34.8123	<0.0001
	mM = mC = 0	−55.7118	121.9188	<0.0001

[*]mD = migration into *M. m. domesticus*.
[†]mM = migration into *M. m. musculus*.
[‡]mC = migration into *M. m. castaneus*.
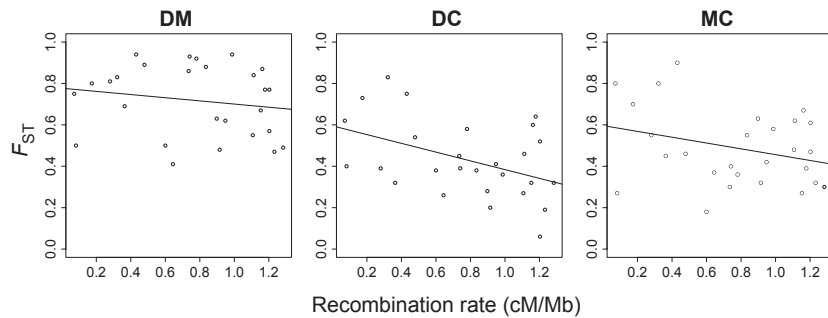
Fig. 4 Scatter plots showing the distribution of F_{ST} values between pairs of house mouse subspecies as a function of the recombination rate for the 27 loci included in this study. From left to right, DM (*Mus musculus domesticus* and *M. m. musculus*; $r = -0.209$, $P = 0.150$), DC (*M. m. domesticus* and *M. m. castaneus*; $r = -0.398$, $P = 0.020$) and MC (*M. m. musculus* and *M. m. castaneus*; $r = -0.210$, $P = 0.147$).

observed values with the simulated distributions. Despite the inference of gene flow from IMa (Table 5), we found that the observed range of F_{ST} falls within 95% of the simulated ranges for both DM and MC (DM, $P = 0.717$, MC, $P = 0.191$). For DC, the result was marginally significant, with 5.7% of the simulations showing F_{ST} ranges greater than the observed range (i.e. $P = 0.057$). These simulations highlight that a wide range of F_{ST} is expected even in the absence of gene flow.

Recombination rates for the 27 loci varied between 0.072 and 1.283 cM/Mb (Table 2). We compared recombination rate and F_{ST} in each of the three pairwise comparisons (Fig. 4). There was a significant negative correlation between F_{ST} and recombination rate for DC, and a nonsignificant trend in the same direction for the other two comparisons (DC $r = -0.398$, $P = 0.020$; MC

$r = -0.210$, $P = 0.147$; DM $r = -0.209$, $P = 0.150$). We also divided all loci into two groups, those with recombination rates above the genomic average and those with recombination rates below the genomic average, and we obtained similar results. F_{ST} for low recombination loci was significantly greater than F_{ST} for high recombination loci for DC (average $F_{ST} = 0.57$ for low and $F_{ST} = 0.37$ for high, Mann–Whitney U test $P = 0.008$) and for MC (average $F_{ST} = 0.62$ for low and $F_{ST} = 0.43$ for high, Mann–Whitney U test $P = 0.026$) (Table 2). For DM, the difference was not significant but tended in the same direction (average $F_{ST} = 0.78$ for low and $F_{ST} = 0.69$ for high, Mann–Whitney U test $P = 0.164$). We also compared the ratio of fixed differences to total polymorphism between each pair of subspecies for low and high recombination loci in a 2×2 contingency table and found that in each case, the ratio

Table 6 Fisher's exact tests comparing polymorphisms and fixed differences for regions of low and high recombination rates

Comparison	Recombination rate	Polymorphisms	Fixed differences	*P*-value
Mus musculus domesticus–Mus musculus musculus	Low	169	54	
	High	502	73	8.2×10^{-05}
M. m. domesticus–M. m. castaneus	Low	410	11	
	High	1094	5	8.2×10^{-04}
M. m. musculus–M. m. castaneus	Low	360	24	
	High	1036	10	9.0×10^{-08}

of fixed differences to polymorphism was significantly higher for low recombination loci ($P < 0.001$) (Table 6). These results show that regions of low recombination are more differentiated, on average, than regions of high recombination, but they do not speak directly to the cause of this greater differentiation.

To explore whether selection at linked sites (i.e. background selection or genetic hitchhiking) is reducing variation within subspecies and thereby increasing differentiation in regions of low recombination, we compared the levels of nucleotide diversity (π) within each subspecies to recombination rate among loci. There was a significant positive correlation for *M. m. musculus* (Spearman rank order correlation $r = 0.383$; $P = 0.02$), but not for the other two subspecies. Thus, the effects of selection at linked sites within subspecies appear to be modest. Similarly, standard tests of a neutral model based on the distribution of allele frequencies were generally not significant. We calculated Tajima's D (Tajima 1989) and Fu and Li's D (Fu & Li 1993) for each of the three subspecies (Table S4, Supporting information) and found that only two loci in *M. m. domesticus* showed deviations from neutral expectations after a Bonferroni correction for multiple tests (*Iars*: FLD = 1.9284, $P = 0.0014$ and *Lrpprc*: TD = −2.2657, $P = 0.0011$ and FLD = −4.0980, $P = 0.0005$).

To explore whether reduced gene flow is contributing to increased differentiation in regions of low recombination, we calculated D_{XY} and D_A between pairs of subspecies. Reduced gene flow should increase both D_{XY} and D_A, while genetic hitchhiking or background selection should primarily increase D_A and may reduce D_{XY} (Noor & Bennet 2009). We found that both D_{xy} and D_a are higher in regions of low recombination for DC and MC (but not for DM; Table S5, Supporting information), although these differences were not significant (Mann–Whitney U tests $P > 0.05$ for all).

Finally, we compared estimates of Nm from IM with estimates of recombination rate for all loci. We found a significant positive correlation between recombination rate and estimates of 2Nm into *M. m. musculus* (Spearman rank order correlation $r = 0.359$ $P = 0.033$) and a

marginally significant correlation between recombination rate and estimates of 2Nm into *M. m. castaneus* (Spearman rank order correlation $r = 0.290$ $P = 0.071$). We also found that for each of the unidirectional estimates of gene flow (i.e. from a particular subspecies to a particular subspecies), the average Nm was approximately twofold higher for loci in regions of high recombination compared to loci in regions of low recombination (Table 2).

Discussion

This study represents the largest survey of DNA sequence variation in wild populations of the three *M. m.* subspecies. Our main observations are as follows. (i) All three subspecies started to diverge at approximately the same time around 350 000 years ago, yet *M. m. musculus* and *M. m. domesticus* were more differentiated from each other than either were from *M. m. castaneus*. (ii) There is heterogeneity among loci in the levels of population differentiation between pairs of subspecies, with F_{ST} ranging from almost zero to one. Genomic regions with lower rates of recombination are more differentiated, on average, than regions of higher recombination.

History of speciation in house mice

The data and analyses presented here provide insight into the temporal and demographic history of speciation in house mice. Models of isolation-with-gene-flow enabled us to estimate parameters while allowing for gene flow following divergence from an ancestral population. Previous studies suggested that house mouse subspecies started to diverge within the last 500 000 years (She *et al.* 1990; Geraldes *et al.* 2008a). Our analyses, based on considerably more data, suggest that the divergence may be even younger and closer to 350 000 years ago. Interestingly, the three estimates of *t* in pairwise comparisons between subspecies were very similar, suggesting that all three subspecies diverged at approximately the same time. This is consistent with a recent study in which the genome sequences of one

M. m. domesticus, one *M. m. musculus* and one *M. m. castaneus* were analysed to estimate the species phylogeny (White *et al.* 2009). White *et al.* (2009) found highly discordant patterns among loci, with only 39% supporting the best-supported topology (*M. m. domesticus* as the sister group to a clade containing *M. m. musculus* and *M. m. castaneus*), which is close to the 33% expected if all three lineages diverged at approximately the same time as suggested by our results.

Maximum-likelihood estimates of N_e were different for the three subspecies. N_e for *M. m. castaneus* (approximately 366 500) was substantially larger than N_e for *M. m. domesticus* (approximately 82 500) which in turn was larger than N_e for *M. m. musculus* (approximately 36 500). These values are based on a generation time of one year and are twice as large if a generation time of 0.5 years is assumed. Probably the biggest source of error in estimates of N_e is the uncertainty in estimates of generation time. While laboratory mice can have four or more generations per year, wild mice typically breed seasonally (Bronson 1979). Moreover, abundant food for commensal mice, which is associated with more continuous breeding, occurred only after the development of agriculture about 10 000 years ago. This represents a small fraction of the approximately 350 000 year history of these lineages. It thus seems likely that wild *M. musculus* have had a generation time of closer to one year over much of their history. However, uncertainty in the exact value precludes more precise estimates of N_e. The estimates of N_e here are slightly different from previous estimates which were based on fewer loci and more individuals (Geraldes *et al.* 2008a). Geraldes *et al.* (2008a) sampled eight loci (of which only four were autosomal) and two or three major geographic regions for each subspecies, while the present study sampled 27 loci and one major geographic region for each subspecies. These factors are likely to account for the modest differences in estimates of N_e. Nonetheless, the two studies are in good agreement about the relative values of N_e for the three subspecies.

Despite the fact that all three subspecies diverged from each other within a very short period of time, differentiation was greater for DM than for DC or MC. This was seen in the higher average F_{ST} for DM (0.72) compared to either DC or MC ($F_{ST} = 0.43$ and 0.49, respectively) and in the greater proportion of monophyletic gene genealogies in DM compared to DC or MC (Fig. 3). Similarly, the ratio of fixed differences to intrasubspecific polymorphisms was 19% for DM but only 1% for DC and 2% for MC. The proportion of monophyletic gene trees within each subspecies (Fig. 3b) accords well with the relative N_e for each subspecies, suggesting that genetic drift plays a major role in determining the amount of differentiation between subspe-

cies. It is also likely that gene flow following divergence accounts for some of this overall pattern, as gene flow between DM is less than between DC or MC (Table 3).

Interestingly, the greater differentiation observed for DM compared to DC or MC is consistent with the known greater degree of reproductive isolation for DM than for DC or MC. *M. m. musculus* and *M. m. domesticus* form a narrow hybrid zone in Central Europe (Teeter *et al.* 2010), and laboratory crosses between these taxa produce sterile hybrid males (Forejt 1996). In contrast, *M. m. domesticus* and *M. m. castaneus* appear to have hybridized freely in California (Orth *et al.* 1998), and *M. m. musculus* and *M. m. castaneus* hybridized to form *M. m. molossinus* in Japan (Boursot *et al.* 1993).

Recombination rate variation and patterns of differentiation

Recent speciation models have invoked the importance of suppressed recombination in rearranged regions of the genome as a mechanism to facilitate species divergence (Noor *et al.* 2001; Rieseberg 2001; Navarro & Barton 2003). House mice provide a useful system for testing the generality of these models, as there is considerable genomic variation in recombination rate (Shifman *et al.* 2006), but the three subspecies are not distinguished by chromosomal rearrangements. F_{ST} was significantly greater at loci experiencing low recombination rates compared to loci experiencing high recombination rates between *M. m. musculus* and *M. m. castaneus* and between *M. m. domesticus* and *M. m. castaneus*. In the comparison between *M. m. musculus* and *M. m. domesticus*, we observed a trend in the same direction. We found a significant negative correlation between F_{ST} and recombination rate for DC and a nonsignificant trend in the same direction for DM and MC. For all three comparisons, the average F_{ST} was higher for loci in regions of low recombination.

These differences in F_{ST} could be due to differences in gene flow or due to diversity-reducing selection, or some combination of both processes. Teasing apart these processes is exceptionally difficult (Noor & Bennet 2009), and many previous studies that have documented a correlation between recombination rate and F_{ST} have made no attempt to consider these alternatives (e.g. Takahashi *et al.* 2004). The IM analyses performed here show that gene flow is occurring between DC and MC. Moreover, genes with high recombination rates (and low F_{ST}) had the highest IM estimates of gene flow. However, caution is warranted in interpreting this pattern; F_{ST} and IM estimates of gene flow are not independent. Thus, while this pattern is consistent with reduced gene flow in regions of low recombination, we cannot rule out the alternative that selection at linked

sites is also a contributing factor. One approach to disentangling these processes will come from mapping genes underlying reproductive isolation. The speciation models of Noor *et al.* (2001), Rieseberg (2001) and Navarro & Barton (2003) predict that such genes will be found in regions of low recombination.

Several studies have now documented correlated variation in recombination rate and levels of differentiation among populations or closely related species. In *Drosophila melanogaster*, F_{ST} between Africa and North America is greatest in regions of low recombination (Begun & Aquadro 1993). In humans, F_{ST} among African, Asian and European populations is negatively correlated with local rates of recombination (Keinan & Reich 2010). Similarly, a study based on two inbred strains from each subspecies of house mice revealed that G_{ST} was negatively correlated with recombination rate (Takahashi *et al.* 2004). In all of these cases, patterns were attributed to selection at linked sites reducing variation in regions of the genome with low recombination. Other studies have documented variation in the levels of differentiation across the genome but placed the emphasis on the interplay between recombination rate and gene flow (e.g. Rieseberg *et al.* 1999; Machado *et al.* 2002; Geraldes *et al.* 2006). These studies include some taxa in which rearranged chromosomes are believed to suppress recombination (e.g. sunflowers, Rieseberg *et al.* 1999; *Drosophila*, Machado *et al.* 2002) and some studies in which the suppression of recombination is associated with proximity to a centromere (e.g. rabbits, Geraldes *et al.* 2006, 2008b; Carneiro *et al.* 2009). The genomes of the three subspecies of house mice are collinear, and the loci studied here do not lie near centromeres. Therefore, the observation of variation in the levels of differentiation cannot be attributed to chromosomal rearrangements or to proximity to centromeres. Greater differentiation in regions of low recombination therefore appears to be a fairly general phenomenon.

Acknowledgements

References

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, **17**, 1505–1519.

Begun DJ, Aquadro CF (1993) African and North-American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature*, **365**, 548–550.

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

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

Bronson FH (1979) The reproductive ecology of the house mouse. *Quarterly Review of Biology*, **54**, 265–299.

Carneiro M, Ferrand N, Nachman MW (2009) Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics*, **181**, 593–606.

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–543.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Cox A, Ackert-Bicknell CL, Dumont BL *et al.* (2009) A new standard genetic map for the laboratory mouse. *Genetics*, **182**, 1335–1344.

Coyne JA, Orr HA (2004) *Speciation*, Sinauer Associates Inc., Sunderland.

Dumont BL, White MA, Steffy B, Wiltshire T, Payseur BA (2011) Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Research*, **21**, 114–125.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Faria R, Navarro A (2010) Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends in Ecology and Evolution*, **25**, 660–669.

Forejt J (1996) Hybrid sterility in the house mouse. *Trends in Genetics*, **12**, 412–417.

Frazer KA, Eskin E, Kang HM *et al.* (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, **448**, 1050–1053.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Geraldes A, Ferrand N, Nachman MW (2006) Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics*, **173**, 919–933.

Geraldes A, Basset P, Gibson B, Smith KL, Harr B *et al.* (2008a) Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Molecular Ecology*, **17**, 5349–5363.

Geraldes A, Carneiro M, Delibes-Mateos M, Villafuerte R, Nachman MW *et al.* (2008b) Reduced introgression of the Y chromosome between subspecies of the European rabbit (*Oryctolagus cuniculus*) in the Iberian Peninsula. *Molecular Ecology*, **17**, 4489–4499.

Good JM, Handel MA, Nachman MW (2008) Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution*, **62**, 50–65.

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Research*, **8**, 195–202.

Hall TA (1999) BioEdit: a user friendly biological sequence alignment editor and analyses program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2785–2790.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337.

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from sequence data. *Genetics*, **132**, 583–589.

Keinan A, Reich D (2010) Human population differentiation is strongly correlated with local recombination rate. *PLoS Genetics*, **6**, e1000886.

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinformatics*, **9**, 299–306.

Machado CA, Kliman RM, Markert JA, Hey J (2002) Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Molecular Biology and Evolution*, **19**, 472–488.

Macholan M, Munclinger P, Sugerkova M, Dufkova P, Bimova B *et al.* (2007) Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. *Evolution*, **61**, 746–771.

Maroja LS, Andres JA, Harrison RG (2009) Genealogical discordance and patterns of introgression and selection across a cricket hybrid zone. *Evolution*, **63**, 2999–3015.

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

Navarro A, Barton NH (2003) Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, **57**, 447–459.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 3321–3323.

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

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–5273.

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, **25**, 2745–2751.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Noor MA, Bennet SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species *Heredity*, **103**, 439–444.

Noor MA, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 12084–12088.

Orth A, Adama T, Din W, Bonhomme F (1998) Natural hybridization of two subspecies of house mice, *Musculus domesticus* and *Mus musculus castaneus*, near Lake Casitas (California). *Genome*, **41**, 104–110.

Raufaste N, Orth A, Belkhir K, Senet D, Smadja C *et al.* (2005) Inferences of selection and migration in the Danish house mouse hybrid zone. *Biological Journal of the Linnean Society*, **84**, 593–616.

Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, **16**, 351–358.

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

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

She JX, Bonhomme F, Boursot P, Thaler L, Catzeflis F (1990) Molecular phylogenies in the genus *Mus* – comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. *Biological Journal of the Linnean Society*, **41**, 83–103.

Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW *et al.* (2006) A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *Plos Biology*, **4**, 2227–2237.

Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, **73**, 1162–1169.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Storchova R, Gregorova S, Buckiova D, Kyselova V, Divina P, Forejt J (2004) Genetic analysis of X-linked hybrid sterility in the house mouse. *Mammalian Genome*, **15**, 515–524.

Strasburg JL, Rieseberg LH (2010) How robust are ''Isolation with migration'' analyses to violations of the IM model? A simulation study *Molecular Biology and Evolution*, **27**(2), 297–310.

Suzuki H, Shimada T, Terashima M, Tsuchiya K, Aplin K (2004) Temporal, spatial, and ecological modes of evolution of Eurasian Mus based on mitochondrial and nuclear gene sequences. *Molecular Phylogenetics and Evolution*, **33**, 626–646.

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

Takahashi A, Liu YH, Saitou N (2004) Genetic variation versus recombination rate in a structured population of mice. *Molecular Biology and Evolution*, **21**, 404–409.

Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM *et al.* (2008) Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Research*, **18**, 67–76.

Teeter KC, Thibodeau LM, Gompert Z, Buerkle CA, Nachman MW *et al.* (2010) The variable genomic architecture of isolation between hybridizing species of house mice. *Evolution*, **64**, 472–485.

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W – improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Tucker PK, Sage RD, Warner J, Wilson AC, Eicher EM (1992) Abrupt cline for sex-chromosomes in a hybrid zone between 2 species of mice. *Evolution*, **46**, 1146–1163.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

White MA, Ane C, Dewey CN, Large BR, Payseur BA (2009) Fine-scale phylogenetic discordance across the house mouse genome. *PLoS Genetics*, **5**, e1000729.

Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: Fst ≠ 1/(4Nm +1). *Heredity*, **82**, 117–125.

Woerner AE, Cox MP, Hammer MF (2007) Recombination-filtered genomic datasets by information maximization. *Bioinformatics*, **23**, 1851–1853.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F (2007) On the subspecific origin of the laboratory mouse. *Nature Genetics*, **39**, 1100–1107.

Data accessibility

DNA sequences: Individual by individual sequences deposited in Genbank under accessions JF336572–JF338137.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Gene genealogies for each locus surveyed in this study.

Fig. S2 Summary of genealogical relationships among subspecies of house mouse for the 27 loci surveyed in this study.

Table S1 Geographic origin and sex of the samples used in this study.

Table S2 Primer details.

Table S3 Comparison of the total amount of data generated and the largest non-recombining blocks used for IM and IMa analyses.

Table S4 Levels of polymorphism within subspecies of house mice and divergence (DXY) between these and Mus caroli.

Table S5 Estimates of population divergence between pairs of house mouse subspecies.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.