Functional Organization of the Genome May Shape the Species Boundary in the House Mouse

Václav Janoušek,[1,2] Pavel Munclinger,[1] Liuyang Wang,[‡,3] Katherine C. Teeter,[4] and Priscilla K. Tucker[*,3]

[1]Department of Zoology, Faculty of Science, Charles University in Prague, Prague, Czech Republic

[2]Institute of Vertebrate Biology, ASCR, Brno, Czech Republic

[3]Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan

[4]Department of Biology, Northern Michigan University

‡Present address: Department of Molecular Genetics and Microbiology, School of Medicine, Duke University, Durham, NC

***Corresponding author:** E-mail: ptuck@umich.edu.

Associate editor: Michael Rosenberg

Abstract

Genomic features such as rate of recombination and differentiation have been suggested to play a role in species divergence. However, the relationship of these phenomena to functional organization of the genome in the context of reproductive isolation remains unexplored. Here, we examine genomic characteristics of the species boundaries between two house mouse subspecies (*Mus musculus musculus*/*M. m. domesticus*). These taxa form a narrow semipermeable zone of secondary contact across Central Europe. Due to the incomplete nature of reproductive isolation, gene flow in the zone varies across the genome. We present an analysis of genomic differentiation, rate of recombination, and functional composition of genes relative to varying amounts of introgression. We assessed introgression using 1,316 autosomal single nucleotide polymorphism markers, previously genotyped in hybrid populations from three transects. We found a significant relationship between amounts of introgression and both genomic differentiation and rate of recombination with genomic regions of reduced introgression associated with higher genomic differentiation and lower rates of recombination, and the opposite for genomic regions of extensive introgression. We also found a striking functional polarization of genes based on where they are expressed in the cell. Regions of elevated introgression exhibit a disproportionate number of genes involved in signal transduction functioning at the cell periphery, among which olfactory receptor genes were found to be the most prominent group. Conversely, genes expressed intracellularly and involved in DNA binding were the most prevalent in regions of reduced introgression. We hypothesize that functional organization of the genome is an important driver of species divergence.

Key words: **hybrid zone, mouse genome, speciation.**

Article

Introduction

Thanks to advances in high-throughput technologies research on the genetics of speciation has shifted in the last few years from a study of a few genes to a study of whole genomes. The shift from the genic to genomic level enables us to better understand the patterns of genomic differentiation between diverging species which ultimately give rise to reproductive isolation (Nosil and Feder 2012). Genomic regions of high differentiation—varyingly referred to in the literature as "islands of speciation" or "islands of differentiation" (Turner et al. 2005; Harr 2006; Nosil et al. 2009)—have been attributed to limited gene flow due to reduced recombination and/or diversifying selection in sympatry (reviewed in Butlin 2005; Feder et al. 2012). Alternatively, genomic regions of high differentiation may result from the faster sorting of alleles due to selection at linked sites in allopatry (Noor and Bennett 2009; Nachman and Payseur 2012; Cruickshank and Hahn 2014). Reduced recombination is viewed in both scenarios to play an important role in species divergence. However, its role in the evolution of reproductive isolation is still unclear.

Recent genome-wide scans in species which evolved at least partially in allopatry provide conflicting results on the relationship between recombination, differentiation, and reproductive isolation. For example, Ruegg et al. (2014), studying two parapatric species of songbirds, reported no overlap between loci linked to putative isolating traits and genome differentiation. In this study, genome differentiation was attributed to reduced recombination. Ellegren et al. (2012) found no overlap between regions of high differentiation and reduced recombination for parapatric populations of flycatchers. Interestingly, Renaut et al. (2013) found the same genomic regions of high differentiation between pairs of sympatric as well as allopatric populations of sunflowers. These authors attributed differences in differentiation across the genome to heterogeneous rates of recombination, possibly associated with the functional organization of the genome.

Only a few speciation genes have been identified, and they are primarily from taxa for which intrinsic reproductive isolation is the primary cause of their isolation (Orr et al. 2004; Presgraves 2010). The majority of these genes were identified in the genus *Drosophila* (Sawamura and Yamamoto 1997;

Ting et al. 1998; Barbash et al. 2003; Presgraves et al. 2003; Brideau et al. 2006; Bayes and Malik 2009; Ferree and Barbash 2009; Phadnis and Orr 2009; Tang and Presgraves 2009). The only mammalian speciation gene (*Prdm9*) identified to date is in the house mouse system (Mihola et al. 2009). Functionally, these genes are often involved in DNA binding (Presgraves 2010) and thus they may be a cause of Dobzhansky–Muller Incompatibilities (DMIs; Dobzhansky 1936; Muller 1940, 1942; Coyne and Orr 2004). Some authors hypothesize that they are likely to be involved in various forms of genetic conflict (reviewed in Presgraves 2010). Despite these recent advances, there has been no systematic study assessing the role of the higher functional organization of the genome in species differentiation and in the evolution of reproductive isolation.

Hybrid zones represent places where newly emerging species contact and interbreed. As such, they provide an opportunity to identify genomic regions contributing to speciation (Barton and Hewitt 1985). Analysis of differential introgression is used for this purpose (Payseur 2010). In the context of hybrid zone analysis the term introgression differs slightly from the original definition by Anderson and Hubricht (1938), that is, it is widely used to describe the degree of gene flow and its directionality.

The two house mouse subspecies (*Mus musculus musculus*/*M. m. domesticus*) contact and interbreed in Central Europe forming a narrow hybrid zone (fig. 1) which represents a secondary contact after some period of time spent in allopatry. The subspecies diverged less than 0.5 Ma (Salcedo et al. 2007; Geraldes et al. 2008; Duvaux et al. 2011), but were likely subject to episodes of gene flow long before the secondary contact in central Europe (Duvaux et al. 2011). However, as Duvaux et al. (2011) pointed out, there has still been enough time in allopatry for incompatibilities to evolve. Indeed, there is substantial evidence for intrinsic postzygotic reproductive isolation from laboratory crosses (Forejt and Iványi 1974; Storchová et al. 2004; Britton-Davidian et al. 2005; Good, Dean, et al. 2008; Good, Handel, et al. 2008; White et al. 2011) as well as from wild-sampled hybrid individuals (Turner et al. 2011; Albrechtová et al. 2012), in both cases affecting mostly males. In general, DMIs are suspected to be a cause of hybrid male unfitness in laboratory crosses between the two subspecies (White et al. 2011).

Here, we combine data for 1,316 autosomal single nucleotide polymorphism (SNP) markers from hybrid mice collected from three transects (fig. 1) along with genomic characteristics obtained from publicly available mouse genomic resources to study the relationship between gene flow, genomic differentiation, rate of recombination, and the functional composition of genes. Our goal is to elucidate the role these relationships play in shaping the species boundary in a mammal.

We report a significant relationship between amounts of introgression and both rate of recombination and genomic differentiation. The genomic regions of reduced introgression exhibit a lower male-specific (MS) rate of recombination and higher genomic differentiation, the opposite of regions with elevated introgression. Also, we found striking differences in the functional composition of genes with respect to amounts of introgression. Genomic regions of high introgression tend to have a higher prevalence of genes with functions associated with cell surface and signal transduction, whereas genomic regions of reduced introgression tend to have a higher prevalence of genes with their functions acting at the intracellular level and most notably in transcription factor activity. Our work thus supports recent findings on the importance of genomic functional organization in species divergence (Renaut et al. 2013; Ruegg et al. 2014) and is the first to report these patterns for mammals. In addition, our work suggests that functional characteristics and their distribution in the genome may also be important in the evolution of reproductive isolation in the house mouse. We find that some functional classes of genes are more prone to be involved in reproductive isolation and other classes are more prone to cross species boundaries.

Results

Bayesian Genomic Cline Analysis Output

To assess patterns of introgression at 1,316 autosomal SNP markers in three house mouse hybrid zone transects (see Materials and Methods for details), we used two genomic cline model parameters (α, β) inferred by Bayesian Genomic Cline Analysis (BGCA) as implemented by Gompert and Buerkle (2011, 2012; see Materials and Methods for details). We found a considerable number of outliers, both positive and negative values, for both parameters (α, β) (table 1; see fig. 2 for examples of genomic clines). The distributions of the actual α and β parameter values merged over the five Markov chain Monte Carlo (MCMC) chains along with their most conservative credible intervals are provided (supplementary fig. S1, Supplementary Material online). In general, α values were more symmetrically distributed around zero than β parameter values. The β parameter exhibited more outliers for positive β parameter values (i.e., those under putative selection against hybrids) than outliers for negative β parameter values (i.e., those which escape the effect of a reproductive barrier). This may result from using differentially fixed SNP markers between two house mouse subspecies which may bias marker choice to those with reduced amounts of introgression (i.e., those under putative selection in the house mouse hybrid zone). The size of credible intervals (supplementary fig. S1, Supplementary Material online) for the two parameters varies between the three transects and it likely reflects the number of samples in each transect (supplementary table S1, Supplementary Material online).

Distribution of the α and β Parameter Outliers in the Mouse Genome and Between-Transect Comparison

The distribution of outliers in the mouse genome for both parameters was plotted (supplementary fig. S2*A* and *B*, Supplementary Material online). In all three transects, the outliers for both parameters tend to be clustered according to the parameter category. When compared with the amount of clustering based on a randomized distribution of SNP markers the observed clustering is highly significant

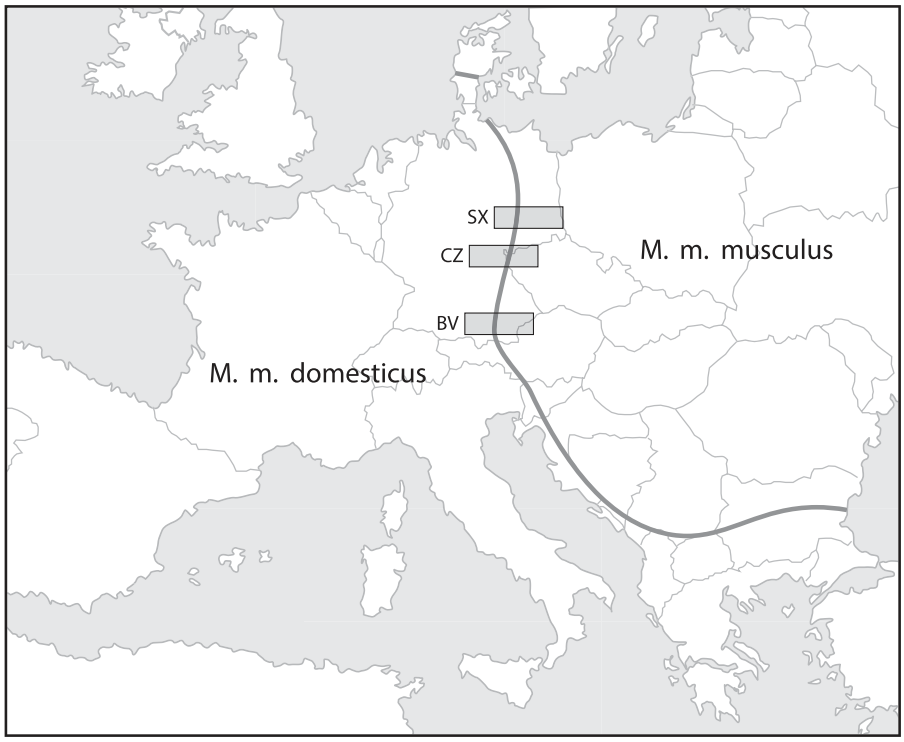


Fig. 1. The location of the house mouse hybrid zone in Europe. The black line demarcates the hybrid zone and the shaded rectangles indicate the location of the three transects studied. Collecting localities for the Czech-Bavarian (CZ) and Bavarian-Austrian (BV) transects can be found in Wang et al. (2011) and for the Saxony (SX) transect in Teeter et al. (2008).

Table 1. Number and Percentage of Outliers for α and β Parameters Inferred Using BGCA.

Parameter	Transect	Positive		Nonsignificant (zero)		Negative	
		n	%	*n*	%	*n*	%
α	CZ	371	28.21	541	41.14	403	30.65
	BV	318	24.18	561	42.66	436	33.16
	SX	242	18.40	700	53.23	373	28.37
β	CZ	505	38.40	510	38.78	300	22.81
	BV	566	43.04	489	37.19	260	19.77
	SX	446	33.92	637	48.44	232	17.64

(supplementary fig. S2C and D, Supplementary Material online) and is likely due to physical linkage between individual markers.

We found highly significant positive correlations between all three transects for both parameters (Spearman's correlation coefficient; $\alpha_{CZ\times BV}$: $\rho = 0.196$, $P = 7.462$ E-13; $\alpha_{CZ\times SX}$: $\rho = 0.2132$, $P = 5.559$ E-15; $\alpha_{BV\times SX}$: $\rho = 0.2034$, $P = 9.67$ E-14; $\beta_{CZ\times BV}$: $\rho = 0.194$, $P = 1.289$ E-12; $\beta_{CZ\times SX}$: $\rho = 0.2592$, $P = 1.241$ E-21; $\beta_{BV\times SX}$: $\rho = 0.2521$, $P = 1.617$ E-20; supplementary fig. S3, Supplementary Material online, where CZ is Czech-Bavarian; BV, Bavarian-Austrian; and SX, Saxony). However, the explained variance in the data was always lower than 10% ($\alpha_{CZ\times BV}$: $\rho^2 = 0.0384$, $\alpha_{CZ\times SX}$: $\rho^2 = 0.0455$, $\alpha_{BV\times SX}$: $\rho^2 = 0.0414$, $\beta_{CZ\times BV}$: $\rho^2 = 0.0376$, $\beta_{CZ\times SX}$: $\rho^2 = 0.0672$, $\beta_{BV\times SX}$: $\rho^2 = 0.0636$). This suggests variability in reproductive isolation between different parts of the house mouse hybrid zone and/or a strong influence of stochastic processes.

Relationship between α and β Parameter Values

We used a nonlinear regression to assess the relationship between the two parameters, direction of introgression (α) and amounts of introgression (β). The α and β parameters were chosen to be explanatory and response variables, respectively. For nonlinear regression, the second-order polynomial function was fitted for all three transects together with the interactions between the α parameter and the transect as well as the quadratic member of the α parameter and the transect (fig. 3 and supplementary table S3, Supplementary Material online). We found all the tested factors highly significant including interactions. When we tested whether the second-order polynomial model explains significantly more variance than a model without the quadratic member (i.e., linear regression), we found the model with the quadratic member explained significantly more variance ($F_{3,3936} = 246$, $P \leq 2$ E-16; supplementary table S3, Supplementary Material online). We also conducted the test for each transect separately with the second-order polynomial model explaining significantly more variance than a linear regression for all three transects (CZ: $F_{1,1312} = 483.1$, $P \leq 2$ E-16; BV: $F_{1,1312} = 77$, $P \leq 2$ E-16; SX: $F_{1,1312} = 287.85$, $P \leq 2$ E-16; supplementary table S3, Supplementary Material online).

The second-order polynomial model had an inverted U-shape with differences between the three transects.

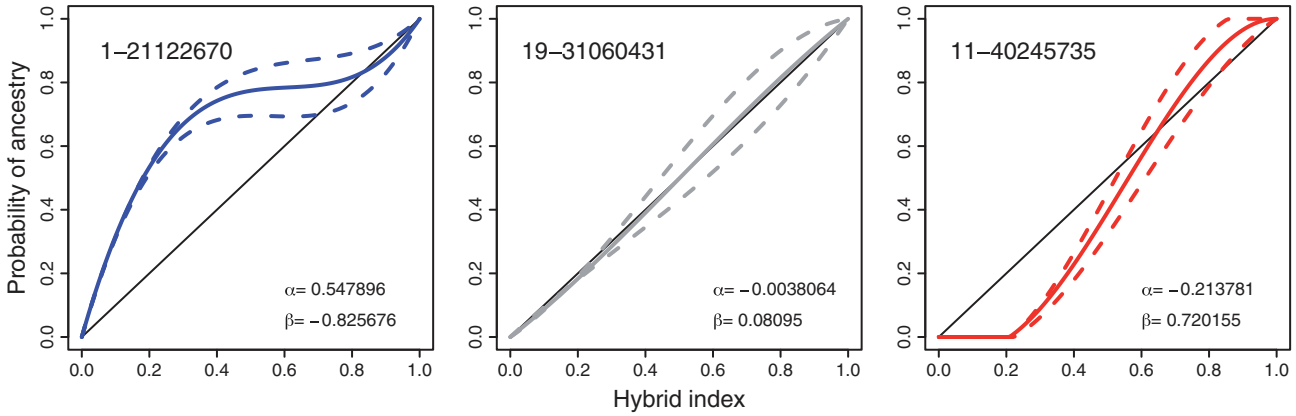


Fig. 2. Example of genomic clines for three SNP markers (1-21122670, 19-31060431, and 11-40245735) which differ in their pattern of introgression. The genomic cline for the 1-21122670 SNP marker (blue) represents a case where the cline is wide ($\beta < 0$) and at the same time is shifted to *Mus musculus domesticus* range ($\alpha > 0$). The genomic cline for the 19-31060431 SNP marker (gray) represents a case where the genomic cline does not differ from the null model (black diagonal). The genomic cline for the 11-40245735 SNP marker represents a likely case of selection against hybrids where the genomic cline is much steeper than the null model ($\beta > 0$). This genomic cline is also slightly shifted to the *M. m. musculus* range ($\alpha < 0$).

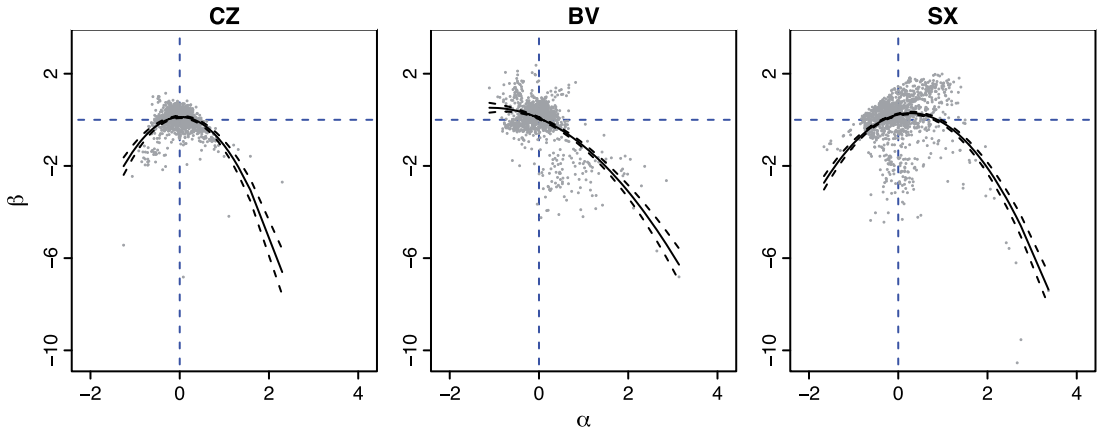


Fig. 3. The nonlinear relationship between α and β parameters for each transect. The black curves represent the polynomial model ($y = a + bx + cx^2$) and its 95% confidential intervals.

The positive or zero β parameter values (i.e., those under selection against hybrids or not deviating from the genome-wide expectation) represent the peak of the inverted U-shaped curve and are mostly associated with α parameter values close to zero. For negative β parameter values, α parameter values are generally deviating from zero, the lower the β parameter, the greater the deviation of the α parameter from zero. In general, this means that genomic clines for markers associated with parts of the genome under selection against hybrids do not shift their position from the genome-wide expectation. This observation is likely a result of coincidence of markers associated with genes involved in reproductive isolation. In contrast, markers associated with genomic regions which freely cross the hybrid zone, that is, those that are not affected by reproductive barriers, tend to be shifted on either side of the genome-wide expectation.

We also found significant differences between the three transects. The CZ and SX transects exhibited generally more symmetry with their negative β tails having both positive and negative α values, with the SX transect having the inverted U-shaped curve wider likely as a result of higher variation in the data due to poorer sampling in this transect. The BV transect was rather asymmetric with negative β values having mostly positive α. This discrepancy between the BV and the other two transects can be explained by their different dynamics as shown by Wang et al. (2011), that is, the BV transect exhibits genome-wide patterns suggesting hybrid zone movement in the direction of the *M. m. domesticus* range. In figure 3, the BV transect lacks SNP markers having β negative and introgression into *M. m. musculus* range ($\alpha < 0$). Wang et al. (2011) showed that the BV transect has moved as a result of demographic changes in a westward direction (i.e., into *M. m. domesticus* range) and as such it has left behind a footprint comprising genomic regions not involved in reproductive isolation. This interferes with distinguishing potential adaptively introgressing alleles from the ones that remain behind as the zone is moving.

The variance explained by the second-order polynomial model is approximately 25% (supplementary table S3,

Supplementary Material online) suggesting that the relationship between α and β parameters is much more complex and a simple polynomial model is far from explaining all of the variation.

Genomic Differentiation and Rate of Recombination Relative to Amounts of Introgression (β parameter category)

We found a highly significant effect of the β parameter category reflecting amounts of introgression on genomic differentiation for all three window sizes (250, 500 kb, and 1 Mb) when tested across the three transects (250 kb: Log-Likelihood$_{null}$ [LL$_{null}$] = 4,078, LL$_{\beta}$ = 4,088, Log-Likelihood Ratio [LLR] = 18.63, P = 0.0001; 500 kb: LL$_{null}$ = 4,896, LL$_{\beta}$ = 4,906, LLR = 21.32, $P \leq 0.0001$; 1 Mb: LL$_{null}$ = 5,717, LL$_{\beta}$ = 5,728, LLR = 20.66, $P \leq 0.0001$; supplementary table S4, Supplementary Material online). When the analysis was conducted for each transect separately the effect of the β parameter category proved to be highly significant for the SX transect for all three window sizes; however, for the other two transects its significance varied depending on window size with some comparisons significant and others not (supplementary table S4, Supplementary Material online). Nevertheless, the overall pattern was consistent across all three transects and window sizes (fig. 4*A* and supplementary fig. S4*A*, Supplementary Material online). For models with a significant effect for the β parameter category, the genomic differentiation for SNP markers which do not cross the house mouse hybrid zone tends to be higher than for those corresponding to the genome-wide expectation. The opposite is true for SNP markers which cross the hybrid zone more extensively than the genome-wide expectation, even though the difference is not as pronounced as for SNP markers with reduced amounts of introgression.

For rate of recombination on the 1-Mb window scale, the difference in explanatory power between the model with the β parameter as a fixed effect and the model without this effect was largest and statistically significant for the MS recombination map and less so, though still significant, for the sex-averaged (SA) recombination map. There was no difference for the female-specific (FS) recombination map (MS: LL$_{null}$ = −639.6, LL$_{\beta}$ = −632.5, LLR = 14.16, P = 8 E-04; SA: LL$_{null}$ = −256.2, LL$_{\beta}$ = −251.2, LLR = 9.974, P = 0.0068, FS: LL$_{null}$ = −774.5, LL$_{\beta}$ = −772.5, LLR = 3.899, P = 0.1423; supplementary table S5, Supplementary Material online). On the 10-Mb scale, the difference was marginally significant for only the MS map (SA: LL$_{null}$ = 2,697, LL$_{\beta}$ = 2,698, LLR = 1.953, P = 0.3766; MS: LL$_{null}$ = 1,590, LL$_{\beta}$ = 1,593, LLR = 5.055; P = 0.0799; FS: LL$_{null}$ = −493.2, LL$_{\beta}$ = −493.0, LLR = 0.3789; P = 0.8274). We found a significant effect for the SA and MS rate of recombination for the BV and the SX transects; however, no significance was found for the CZ transect. No effect was found for the FS rate of recombination (fig. 4*B* and supplementary fig. S4*B*, Supplementary Material online). For SA and MS rates of recombination, we found that SNP markers exhibiting limited introgression tend to be associated with a lower rate of recombination than those with introgression corresponding to the genome-wide expectation in the BV and SX transects. In contrast, SNP markers with greater amounts of introgression tend to be associated with rates of recombination that are higher than the genome-wide expectation. The absence of an association between FS rates of recombination and various β categories likely results from the known differences in sex-specific rates of recombination.

For both genomic differentiation and rate of recombination, the explanatory effect of the β category in the CZ transect is either nonsignificant or the least significant of the three transects. This inconsistency may stem from relatively complex geography and/or evolutionary processes occurring in this transect (Macholán et al. 2008), making proper analysis of introgression more challenging than for the other two transects.

Functional Composition of Genes Relative to Amounts of Introgression (β parameter category)

Using Ward's hierarchical agglomerative clustering method we found differences in the functional composition of genes for β negative and the other two β categories (i.e., zero and positive) for all three transects and window sizes (i.e., 250, 500 kb, and 1 Mb). Based on 1,000 bootstraps we found this distinction to be significant (supplementary fig. S5*A*, Supplementary Material online). The similar functional distinction between β categories was also found using other agglomerative clustering methods (supplementary fig. S5*B*, Supplementary Material online). The average-linkage and complete-linkage agglomerative methods provided very good resolution between the β negative and the other two β categories. The single-linkage method provided a less clear picture; however, this method is generally considered to be more sensitive to outliers (Tan et al. 2005). We also conducted a principal component analysis on GO profiles, identifying again a clear distinction between the gene lists of β negative and the other two β categories (supplementary fig. S5*C*, Supplementary Material online). Further, we constructed a heat map based on GO profiles to visualize differences in individual GO terms between gene lists of various β categories (fig. 5*A*). We found that some GO terms that are prevalent in gene lists associated with negative β values are rare in the other two β categories and vice versa. This pattern was the most pronounced in two clusters (#112 and #113; supplementary fig. S8*A*, Supplementary Material online) consisting of 7 and 24 GO terms, respectively, with a considerable number of them having significantly higher/lower prevalence for various window sizes ($P \leq 0.05$; fig. 5*B*). For other GO terms (e.g., #114 with 86 GO terms; supplementary fig. S8*A*, Supplementary Material online), the contrast between the β categories was less striking or none. We explored all terms with their P values equal or lower than 0.05 as we were looking for trends and consistencies between transects and window sizes. Nevertheless, we considered two other more conservative thresholds ($P \leq 0.01$, $P \leq 0.001$) and found the most prominent GO terms significant in at least one transect (data not shown).

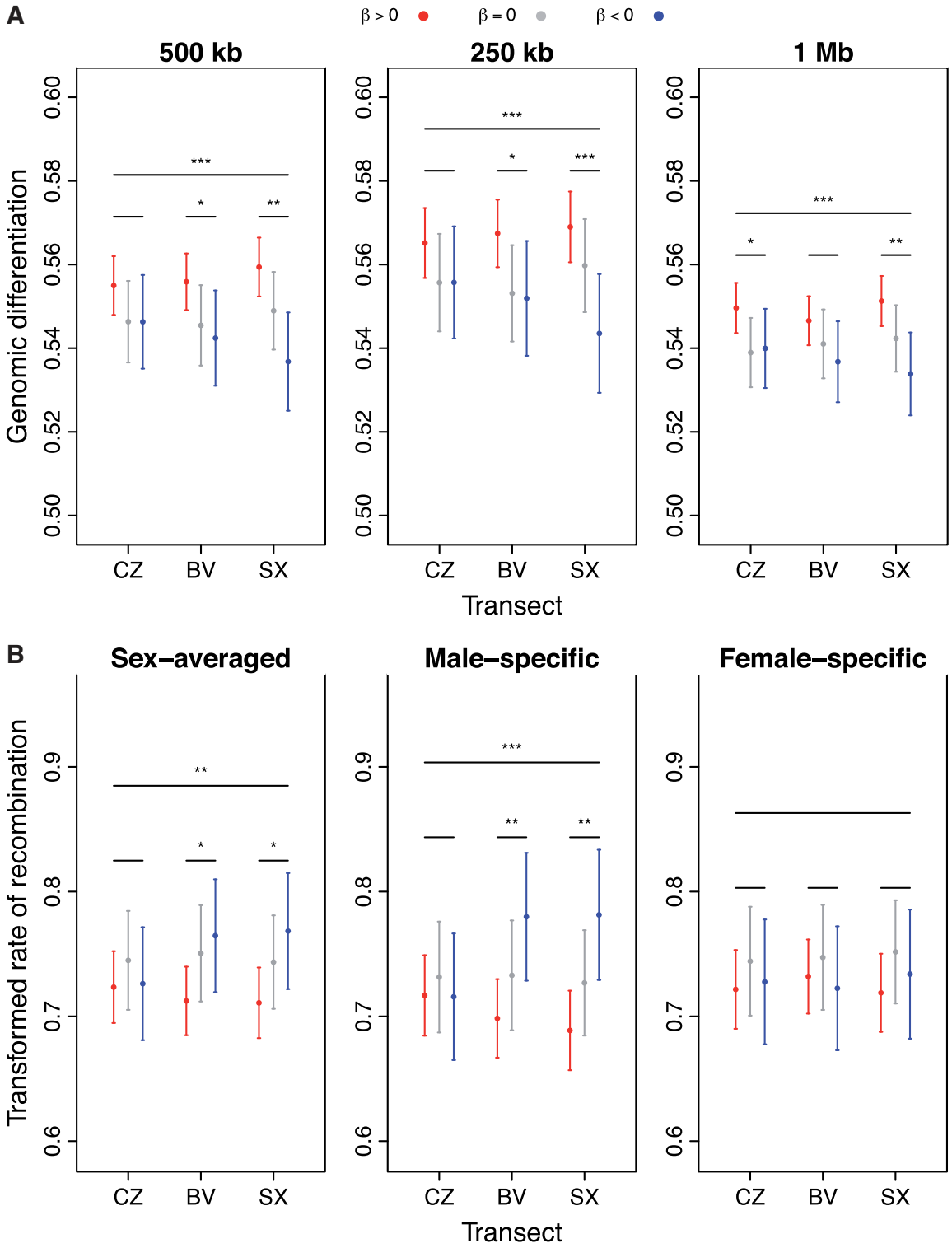
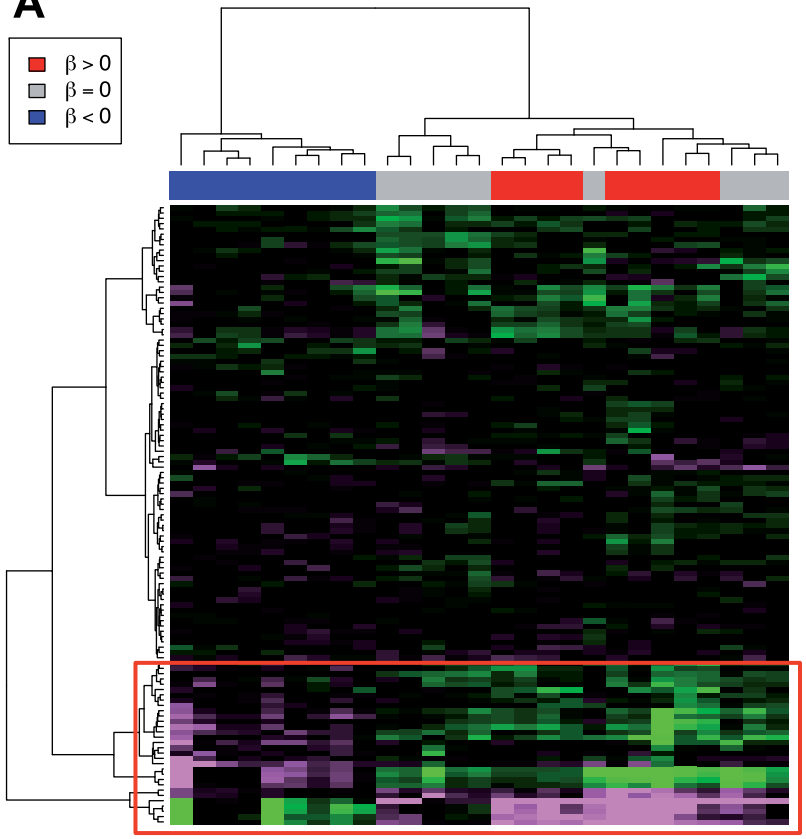


FIG. 4. The difference in genomic differentiation (*A*) and rate of recombination (*B*) between SNPs markers of positive (red), zero (gray), and negative (blue) β values, respectively. Genomic differentiation was plotted for three window sizes (250, 500 kb, and 1 Mb) and rate of recombination was plotted for three conditions (SA, MS, and FS). The average estimates with 95% confidential intervals were plotted along with significance (***$P \leq 0.001$; **$P \leq 0.01$; *$P \leq 0.05$) of the effect of the β category for each transect separately. Significance of the β category effect over all three transects is plotted as well. All of the measures were estimated using mixed-effect linear models where individual markers were nested within genomic regions to treat for underlying linkage structure (see Materials and Methods). Rate of recombination was transformed using a Box–Cox transformation to normalize the data.

A



B

Term accession	Term name	Beta > 0			Beta = 0			Beta < 0		
		CZ	BV	SX	CZ	BV	SX	CZ	BV	SX
GO:0004871	signal transducer activity	-3	-3	-3	-3	0	-3	2	0	2
GO:0005886	plasma membrane	-3	-3	-3	-2	0	-2	2	0	3
GO:0031224	intrinsic to membrane	-2	-3	-3	-3	0	-1	3	0	2
GO:0051716	cellular response to stimulus	-3	-3	-3	-3	0	-3	2	0	2
GO:0071944	cell periphery	-3	-3	-3	-1	0	-2	2	0	3
GO:0003823	antigen binding	-3	-2	-2	-1	-2	-2	-1	0	0
GO:0019882	antigen processing and presentation	-2	-2	-1	-2	-1	-2	-1	0	0
GO:0005549	odorant binding	1	-1	0	-1	1	0	-3	-2	-2
GO:0016740	transferase activity	3	0	3	1	2	0	0	0	0
GO:0032993	protein-DNA complex	1	0	0	0	1	1	-1	0	0
GO:0036094	small molecule binding	3	0	1	0	3	1	0	-1	-1
GO:0043021	ribonucleoprotein complex binding	0	0	0	0	1	0	0	-1	-1
GO:0043167	ion binding	2	0	3	0	1	0	0	0	-1
GO:0043234	protein complex	3	0	0	0	0	1	-2	-1	-2
GO:0097060	synaptic membrane	3	0	1	1	3	1	0	0	0
GO:0097367	carbohydrate derivative binding	3	0	2	0	2	2	1	-1	-1
GO:0002520	immune system development	1	1	1	0	0	0	0	0	0
GO:0003682	chromatin binding	2	1	0	0	0	0	0	0	-1
GO:0006807	nitrogen compound metabolic process	3	1	2	0	1	2	-1	-1	-3
GO:0009058	biosynthetic process	3	1	1	0	0	0	-1	0	0
GO:0016491	oxidoreductase activity	1	1	0	0	0	0	-1	-1	0
GO:0043227	membrane-bounded organelle	3	1	1	3	1	3	-2	0	-2
GO:0044237	cellular metabolic process	3	1	2	0	1	3	-1	0	-2
GO:0016043	cellular component organization	3	2	0	2	2	2	-2	-1	0
GO:0019827	stem cell maintenance	2	2	0	0	0	1	-1	0	0
GO:0003700	sequence-specific DNA binding transcription	3	3	2	0	0	0	0	0	1
GO:0005515	protein binding	3	3	1	3	1	3	-1	0	0
GO:0005622	intracellular	3	3	2	3	3	3	-3	0	-2
GO:0043229	intracellular organelle	3	3	1	2	1	3	-2	0	-2
GO:0044424	intracellular part	3	3	3	3	3	3	-2	0	-2
GO:0097159	organic cyclic compound binding	3	3	3	0	2	1	-1	-1	-1

FIG. 5. Functional analysis of genes. (*A*) Heat map based on hierarchical clustering of functional profiles (*x* axis) for genes from around SNP markers for a given transect, window size and β parameter category ($\beta > 0$: red, $\beta < 0$: blue, $\beta = 0$: gray) and GO terms (*y* axis). The degree of GO term enrichment and depletion for each functional profile reflected by the varying intensity of green and purple colors, respectively. The red rectangle marks the GO terms that are the most polarized for genes around SNPs markers with β negative values and genes from the other two β categories. This figure shows that for all three transects the functional composition is distinct for genomic regions belonging to a different β category. (*B*) Significantly more/less prevalent GO terms marked by red rectangle in (*A*). They correspond to two clusters (#112 and #113 in supplementary fig. S8A, Supplementary Material online). Green and purple colors depict significantly higher and lower prevalence, respectively. The positive and negative numbers represent the number of window sizes for a given combination of transect and β category for which the GO term was significantly more (positive) or less (negative) prevalent.

In general, the functional polarization we observed in the data is mainly due to differences in the cell compartment in which the genes are acting. Genes associated with SNP markers of negative β are mainly associated with GO terms such as plasma membrane (GO:0005886), intrinsic to membrane (GO:0031224), cellular response to stimulus (GO:0051716), cell periphery (GO:0071944), and signal transducer activity (GO:0004871). These GO terms are rarer for genes associated with SNP markers with a positive β or a β equal to zero. This suggests that genes in genomic regions introgressing at a higher rate than the genome-wide expectation tend to act mainly at the cell periphery. The only exception is the BV transect, for which genomic regions with greater amounts of introgression ($\beta < 0$) differed slightly in function from the other two transects. As mentioned earlier, the asymmetry in the BV transect due to hybrid zone movement reported by Wang et al. (2011) might have caused problems in distinguishing adaptive introgression. Despite the difference, both hierarchical clustering and principal component analyses grouped genomic regions exhibiting a greater amount of introgression ($\beta < 0$) in the BV transect with the other two transects (fig. 5*A* and supplementary fig. S5, Supplementary Material online). In contrast, genes associated with SNP markers of positive β or β equal to zero tend to act mainly at the intracellular level as suggested by GO terms such as intracellular (GO:0005622), intracellular organelle (GO:0043229), and intracellular part (GO:0044424). As for GO terms that would distinguish between genes associated with SNP markers of positive β and those with β equal to zero, we found sequence-specific DNA binding transcription activity (GO:0003700) as the GO term most consistently showing higher prevalence for genes associated with SNP markers of positive β. Given the polarization of GO terms, we suggest that the underlying functional structure of the genome is an important player in forming the species genomic boundaries.

We found a high prevalence of genes for olfactory receptors associated with GO terms around SNP markers with negative β. An independent test to assess the prevalence of genes for olfactory receptors in genomic regions of varying values of β showed that these genes were significantly more prevalent in genomic regions of negative β and significantly less prevalent in genomic regions of positive β (supplementary table S6, Supplementary Material online). To treat for the possible influence of olfactory receptors genes in the clustering analysis, we removed them from the data set and reran the analysis (supplementary figs. S6, S7, and S8*B*, Supplementary Material online). We were still able to find a distinction between gene sets associated with SNP markers of negative β and the other two categories suggesting the main axis (i.e., intracellular vs. cell periphery) is more important than the olfactory receptors themselves. The difference was, however, not as pronounced as in the original data set.

Discussion

Properties of the genome such as functional composition of genes, genomic differentiation, and recombination rate are thought to be relevant in the evolution of reproductive barriers; however, their precise roles still need to be elucidated. The aim of our study was to characterize the relationship between amounts of introgression and the abovementioned genomic characteristics to better understand their role in shaping reproductive barriers to gene flow in the house mouse.

Overall Patterns of Introgression in the House Mouse Hybrid Zone

We assessed introgression for 1,316 autosomal SNP markers in three house mouse hybrid zone transects. We found highly significant correlations between pairs of transects for both the α as well as the β parameter. This suggests the importance of common reproductive barriers acting in different parts of house mouse hybrid zone. On the other hand, the correlation was quite low, with less than 10% of the explained variance which is in accord with an assertion by Janoušek et al. (2012) suggesting the influence of stochastic processes on the data or the variable nature of reproductive isolation in different parts of house mouse hybrid zone. Geographic variability in reproductive isolation was also found in laboratory crosses (Vyskočilová et al. 2005, 2009). The fact that we found a significant positive correlation between transects for the α parameter (i.e., cline shift) suggests an overall tendency of alleles to move in the same direction among transects.

We found a highly significant nonlinear relationship between the α and β parameter values for all three transects. The relationship very likely reflects the nature of hybrid zones as described by theoretical models (Barton 1983; Barton and Bengtsson 1986; Bierne et al. 2011). In general, we found that SNP markers with steeper clines form the center of the hybrid zone with no shift from the genome-wide expectation, whereas markers with elevated amounts of introgression tend to be shifted on either side of the genome-wide expectation. The SNP markers with steep clines are hypothesized to be near genes involved in reproductive isolation and, as such, are possibly under selection in hybrids. According to theory (Barton 1983; Barton and Bengtsson 1986; Bierne et al. 2011) they tend to be coupled and form a barrier to gene flow. Conversely, SNP markers with shallow clines shifted away from the genome-wide expectation are likely near genes which actively cross the hybrid zone due to processes such as adaptive introgression. Such genomic regions have been reported from house mouse populations by Staubach et al. (2012). When alleles move adaptively from one species to another one assumes it occurs in the same direction in different parts of the hybrid zone. This is suggested by the correlation of α parameter values between transects. Some of the patterns observed here may also result from neutral ancestral polymorphism. However, our observation of a functional link with the amounts of introgression (see below) makes it unlikely to be simply a result of neutral processes.

Genomic Properties and the Species Boundary

Our comparison of genomic characteristics obtained from publicly available resources revealed a nonrandom association

with amounts of introgression in the house mouse hybrid zone. We found systematically higher differentiation in genomic regions of reduced introgression and, in some instances, lower differentiation in regions of extensive introgression. Also, the rate of recombination tends to be reduced for genomic regions of limited introgression and the opposite for genomic regions of extensive introgression. Our data are in accord with results by Geraldes et al. (2011) who found genomic regions of reduced recombination to have higher differentiation for 27 autosomal loci in the mouse genome. The same result was shown by Phifer-Rixey et al. (2014) for the testes transcriptome. When we compare our observations of varying genomic differentiation and rate of recombination with amounts of introgression in the context of current speciation theory, there are two ways to interpret our data according to the geographic context. In general, genomic regions of higher differentiation have been hypothesized to harbor genes involved in reproductive isolation (Turner et al. 2005; Harr 2006; Nosil et al. 2009). Higher differentiation has been previously attributed to locally reduced gene flow due to reduced recombination and/or diversifying selection in sympatric and parapatric models of speciation (reviewed in Butlin 2005; Feder et al. 2012). In the face of gene flow reduced recombination can prevent the break-up of coadapted genes involved in reproductive isolation (reviewed in Butlin 2005), leading to elevated genomic differentiation. A similar pattern of genomic differentiation by a different process, that is, the faster sorting of alleles due to selection at linked sites, can be observed in cases of allopatry (Noor and Bennett 2009; Nachman and Payseur 2012; Cruickshank and Hahn 2014). In allopatry, the association between rate of recombination, genomic differentiation, and reproductive isolation may result from genetic conflict. This model suggests that incompatibilities involved in genetic conflict may preferentially accumulate in genomic regions of reduced recombination as the linkage between distorter and responder loci in these regions is less likely to break away (reviewed in Seehausen et al. 2014). Given the observations by Duvaux et al. (2011), gene flow between the two house mouse subspecies occurred long before the secondary contact in Central Europe. However, since divergence, the two subspecies have spent most of the time in allopatry, suggesting the majority of incompatibility loci accumulated during this period. Although a role for gene flow in the evolution of reproductive isolation in the house mouse has been suggested (see Vošlajerová-Bímová et al. 2011), direct observations from the hybrid zone suggest that the main cause of hybrid inferiority is postzygotic isolation due to reduced fertility of hybrid males (Turner et al. 2011; Albrechtová et al. 2012). Regardless of the speciation model, our observations suggest rate of recombination and genomic differentiation are important players in species divergence.

In addition to the role of genomic differentiation and rate of recombination, we also found striking differences in functional composition of genes between genomic regions of varying amounts of introgression. The most interesting finding was a functional polarization of genes with regard to in which part of the cell they act. Genomic regions of reduced introgression exhibited a higher prevalence of genes acting at the intracellular level. In contrast, genomic regions of extensive introgression exhibited a higher prevalence of genes acting at the cell periphery.

Among the molecular functions significantly enriched in genomic regions with reduced introgression, "sequence-specific DNA binding transcription activity" is the most consistent among the three transects. This agrees with previous findings by Janoušek et al. (2012) who found molecular functions related to transcription activity enriched in genomic regions under negative epistasis. Indeed, some speciation genes, identified to date, are known to be involved in DNA binding activity (Presgraves 2010). For regions of extensive introgression, "signal transducer activity" was the most prevalent molecular function. We also found that olfactory receptor genes were a prominent group of genes enriched in genomic regions of extensive introgression but depleted in genomic regions of reduced introgression. The functional polarization remained even after the removal of olfactory genes suggesting that cellular compartment (i.e., intracellular vs. cell periphery) is driving the pattern. Interestingly, similar to our findings, Staubach et al. (2012) found olfactory receptor genes and other sensory perception genes acting at the cell periphery to be enriched in introgressed haplotypes between the two house mouse subspecies.

More generally, in humans, genes expressed at the cell periphery have been shown to correspond to those at the periphery of a protein network. They undergo positive selection and segmental duplication more frequently than genes expressed intracellularly (Kim et al. 2007). They are more likely to be involved in interactions with the environment and, as such, they are likely to experience selection in the face of environmental perturbations (Kim et al. 2007). These genes may traverse species boundaries as a result of balancing selection (e.g., transpecific polymorphisms) or adaptive introgression (reviewed in Hedrick 2013). Olfactory receptor genes in the house mouse may represent a good example of these genes as they are known to be involved in chemical communication with the external environment and they often undergo rapid molecular evolution and extensive segmental duplications (Godfrey et al. 2004). Conversely, intracellular genes tend to be more interconnected, constrained, and essential. When discussed in light of DMIs, putatively responsible for intrinsic reproductive isolation in the house mouse, intracellular highly interconnected genes are more likely the cause than loosely interconnected genes on the cell periphery.

A correlational analysis assessing rate of recombination and various genomic features by Frazer et al. (2007) found an association between rate of recombination and various GO terms in humans. These authors found genes expressed mainly at the cell periphery (i.e., receptors) to have a high rate of recombination, whereas genes acting mainly at the intracellular level (i.e., nucleic acid binding) to have lower rates of recombination. The functional axis associated with varying levels of recombination found by Frazer et al. (2007) corresponds to the functional polarization (i.e., intracellular vs. cell periphery) we found between genomic regions of varying introgression. Therefore, this represents a link between

functional composition of genes, rate of recombination and, indirectly, genomic differentiation. Given these facts, we hypothesize that the functional organization of the genome may be an important player shaping species boundaries and ultimately the evolution of reproductive barriers to gene flow in the house mouse.

Materials and Methods

House Mouse Hybrid Zone Data Sets

We used previously published data sets from the CZ and the BV house mouse hybrid zone transects (Wang et al. 2011) and an unpublished data set from the SX transect. Counts of samples differ between the three transects (supplementary table S1, Supplementary Material online) and their description and geographic distribution can be found in Teeter et al. (2008, 2010) for the BV and the SX transects and in Wang et al. (2011) for the BV and the CZ transects. All of these samples were genotyped at 1,316 autosomal SNP markers evenly distributed in the mouse genome with mean distance between neighboring markers 1.86 Mb (for further details, see Wang et al. 2011). When transformed into the Build 38 coordinate system using an in-house Ensembl PERL API script, the average, minimal and maximal distances between neighboring markers on autosomes are 1.79, 0.13, and 10.47 Mb, respectively.

Bayesian Genomic Cline Analysis

The BGCA implemented by Gompert and Buerkle (2011, 2012) was used to assess introgression in the house mouse hybrid zone. It utilizes genomic cline models to describe patterns of introgression between two parental species at focal markers. The models describe the probability of ancestry of one parental species at a focal marker given the hybrid index with respect to a null model (see Gompert and Buerkle 2011 for details). The genomic cline model has two basic parameters—α and β. Following Gompert and Buerkle (2011), the genomic cline parameter α measures the change in probability of ancestry relative to a null expectation. For example, increasing (positive values) or decreasing (negative values) α parameter values reflect shifts in genomic clines either to one of the two parental populations or vice versa. The genomic cline parameter β reflects the rate of change in probability of ancestry from one parental population to the other. Positive values of the β parameter denote an increase in the rate of change (steeper cline), whereas negative values denote a decrease in the rate of change (wider clines). For both parameters, a zero value corresponds to the null model.

These two parameters dissect the pattern of introgression at a focal marker into two components which are not, however, fully independent. In our study, we use the α parameter as a proxy for "direction of introgression" and the β parameter as a proxy for "amounts of introgression." Positive and negative α parameter values represent directional movement of alleles into *M. m. domesticus* and *M. m. musculus*, respectively. The β parameter reflects the strength of the barrier to gene flow between the two species. For positive values of β, the higher the β, the stronger the barrier. Negative β values reflect the ability of alleles at a focal marker to escape the effect of the barrier, the greater the negative β, the greater the ability to escape the barrier.

To assess both parameters, we ran five independent MCMC chains for each data set (transect) for 50,000 steps with 25,000 steps as a burn-in period with random seed. The output was recorded each 25th step to obtain a sample of size 1,000. We merged output of all five MCMC chains by averaging estimates over all chains for each marker. The 99% credible intervals were merged over the five MCMC chains by choosing the most conservative (i.e., the widest) intervals. These merged 99% credible intervals were used to detect outliers. Based on this procedure we defined three categories for each parameter (α, β): Positive (significantly higher than zero: α, $\beta > 0$), zero (does not deviate significantly from zero: α, $\beta = 0$), and negative (significantly lower than zero: α, $\beta < 0$). For further analyses we used the β parameter, that is, the relative amount of introgression reflecting putative selection against hybrids. For all subsequent statistical analyses and visualization of the data, we used the statistical environment of the R-project (R Development Core Team 2014) along with appropriate statistical packages (cited below). For the data visualization, we used mainly the "lattice" package (Sarkar 2008). To efficiently work with the genomic data, we used the BEDTools suite (Quinlan and Hall 2010) and in-house PERL scripts.

Analysis of Rate of Recombination and Genomic Differentiation in Relation to Amount of Introgression (β parameter category)

In our study, we analyzed the relationship between amount of introgression (defined according to the β parameter value) and the two other measures, rate of recombination and genomic differentiation. To assess the two measures, we used publicly available resources. The rate of recombination was obtained from the genetic map built by Shifman et al. (2006) and revisited later by Cox et al. (2009). The necessary scripts to extrapolate the genetic map used to estimate rate of recombination in sliding windows of a given size (1 and 10 Mb) were kindly provided by Gary Churchill, The Jackson Laboratory, Bar Harbor, ME. We used SA and sex-specific rates of recombination. We analyzed sex-specific rates of recombination due to known differences in the overall rate of recombination (Davisson et al. 1989) as well as in local variation in rate of recombination between the two sexes (Shifman et al. 2006; Cox et al. 2009). To get an estimate of genomic differentiation between the two house mouse subspecies, we used publicly available genotypes of the Mouse Diversity Array (~500,000 SNPs; Yang et al. 2011). From the array, we pulled out only European samples of the two house mouse subspecies (*M. m. musculus*, *M. m. domesticus*; supplementary table S2, Supplementary Material online). SNPs for the analysis were chosen by two criteria: 1) Those for which genotypes were available for all nineteen samples and 2) those that were polymorphic within these nineteen samples. Because the samples of the two house mouse subspecies were taken from across their range in Europe, they do not

represent any particular mouse population. To measure genomic differentiation, we used a simple index that does not carry any population genetic assumptions. The index characterizes the distribution of genotypes among the 19 samples. It ranges from 0 to 1 according to whether the genotypes are distributed completely randomly among samples of the two subspecies (i.e., it is uninformative with respect to subspecies boundaries) or whether their distribution reflects fixed difference between the two subspecies (i.e., it is diagnostic). The average index of genomic differentiation was calculated for sliding windows of three sizes (250, 500 kb, and 1 Mb) using a step size of one-fifth of the window size (i.e., 50, 100, and 200 kb).

Before we explored the relationship between genomic differentiation, rate of recombination, and amounts of introgression, neighboring SNP markers with the same β category (i.e., $\beta < 0$, $\beta = 0$, $\beta > 0$) were collapsed into genomic regions. This step enabled us to treat for statistical dependence of SNP markers due to underlying physical linkage. In further analyses, this division into genomic regions represented additional hierarchical structure which was used to treat for mutual dependence of neighboring SNP markers.

We used the "lme" method implemented in the "nlme" R package (Pinheiro et al. 2014) to test for the relationship between amounts of introgression, genomic differentiation, and rate of recombination. The method enables one to construct mixed-effect linear models which were used to take into account the hierarchical structure of genomic data. The SNP markers were nested within genomic regions which represented levels of random effect (i.e., chosen from a larger population of regions). When calculated over all three transects, individual transects were treated as random effects as well. The β category for each SNP marker was treated as a fixed effect. To evaluate its significance, we fitted models with and without the β category as a fixed effect and compared them using LLR test.

Functional Composition of Genes in Relation to Amounts of Introgression (β parameter category)

To assess the functional composition of genes associated with SNP markers of varying amounts of introgression (i.e., $\beta < 0$, $\beta = 0$, $\beta > 0$), we used methods of multidimensional statistics implemented in "FactoMineR" and "pvclust" R packages (Suzuki and Shimodaira 2006; Lê et al. 2008). We assessed functional composition of genes for all three transects separately and also included genes within three window sizes (250, 500 kb, and 1 Mb) around the SNP markers. The genes were retrieved from the Ensembl Core database (Database Version 75; Flicek et al. 2014). To characterize functional composition of these genes, we used the Gene Ontology (GO) term classification which was retrieved from the Gene Ontology MySQL server (downloaded March 24, 2014; Ashburner et al. 2000). To reduce redundancy in functional terms, we used only GO terms at the second level of the GO hierarchy. The genes formed gene lists based on transect, window size and β category (i.e., $\beta < 0$, $\beta = 0$, $\beta > 0$), and their functional composition was assessed for each list separately. In our analysis, we calculated functional profiles for each gene list in the manner of Sánchez-Pla et al. (2007). The functional profiles are composed of differences between expected and observed prevalence of genes within a gene list for each GO term separately. In order to use some statistically meaningful entity, we chose the negative decadic logarithm of P values ($-\log_{10}(P)$) based on a Fisher exact test of GO term enrichment/depletion. The appropriate sign (+/−) was further assigned according to whether the difference between expected and observed gene counts was negative or positive. The functional profile for each gene list thus represented a vector of modified P values and we worked with this set of functional profiles as multidimensional data. Each functional profile represented a response variable and the set of GO terms represented explanatory variables.

For the analysis of functional composition, we used two approaches. First, we applied hierarchical clustering implemented in the R project as function "hclust" (a function included in basic R release; R Development Core Team 2014). For the clustering of functional profiles, we used agglomerative methods of which the Ward's minimum variance method based on the Euclidean distances between individual profiles provided the best results. However, we used other agglomerative methods including complete-linkage, average-linkage, and single-linkage methods to test whether the results are robust enough with respect to the choice of method. The data were plotted using a "heatmap" function in R (a function included in the basic R release). The significance and robustness of the tree were based on 1,000 bootstraps and were obtained using the "pvclust" function implemented in the "pvclust" R package (Suzuki and Shimodaira 2006). For the second approach, we used principal component analysis implemented as a "PCA" function in the "FactoMineR" R package (Lê et al. 2008).

Supplementary Material

Supplementary tables S1–S6 and figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Data Accessibility

The genotype data matrix from the SX transect is available from http://hdl.handle.net/2027.42/83674

Acknowledgments

The SX genotype data were collected with support from NSF DEB 0746560 to P.K.T. V.J. was supported by the Grant Agency of Charles University in Prague Grant No. 6421/2012, from the NextGenProject Reg. No. CZ.1.07/2.3.00/20.0303 and from the Institutional Research Support Grant No. SVV 260 087/2014. The authors thank Gary Churchill who kindly provided scripts to obtain estimates on rate of recombination, Michael Nachman for a discussion on methods of data analysis, and Libor Mořkovský for helping with UNIX server at Charles University in Prague.

References

Albrechtová J, Albrecht T, Baird SJ, Macholán M, Rudolfsen G, Munclinger P, Tucker PK, Piálek J. 2012. Sperm-related phenotypes implicated in both maintenance and breakdown of a natural species barrier in the house mouse. *Proc Biol Sci.* 279:4803–4810.

Anderson E, Hubricht L. 1938. Hybridization in *Tradescantia*. III. The evidence for introgressive hybridization. *Am J Bot.* 25:396–402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. *Nat Genet.* 25:25–29.

Barbash DA, Siino DF, Tarone AM, Roote J. 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. *Proc Natl Acad Sci U S A.* 1009:5302–5307.

Barton N, Bengtsson BO. 1986. The barrier to genetic exchange between hybridising populations. *Heredity* 57:357–376.

Barton NH. 1983. Multilocus clines. *Evolution* 37:454–471.

Barton NH, Hewitt GM. 1985. Analysis of hybrid zones. *Annu Rev Ecol Syst.* 16:113–148.

Bayes JJ, Malik HS. 2009. Altered heterochromatin binding by a hybrid sterility protein in *Drosophila* sibling species. *Science* 326: 1538–1541.

Bierne N, Welch J, Loire E, Bonhomme F, David P. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol Ecol.* 20:2044–2072.

Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA. 2006. Two Dobzhansky-Muller genes interact to cause hybrid lethality in *Drosophila*. *Science* 314:1292–1295.

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. *Biol J Linn Soc.* 84:379–393.

Butlin RK. 2005. Recombination and speciation. *Mol Ecol.* 14:2621–2635.

Cox A, Ackert-Bicknell CL, Dumont BL, Ding Y, Bell JT, Brockmann GA, Wergedal JE, Bult C, Paigen B, Flint J, et al. 2009. A new standard genetic map for the laboratory mouse. *Genetics* 182:1335–1344.

Coyne JA, Orr HA. 2004. Speciation. Sunderland (MA): Sinauer Associates.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol.* 23:3133–3157.

Davisson MT, Roderick TH, Doolittle DP. 1989. Recombination percentages and chromosomal assignments. In: Lyon MF, Searle AG, editors. Genetic variants and strains of the laboratory mouse. Oxford: Oxford University Press. p. 432–505.

Dobzhansky T. 1936. Studies on hybrid sterility. 11. Localization of sterility factors in Drosophila pseudoobscura hybrids. *Genetics* 21: 113–135.

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

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H, Nadachowska-Brzyska K, Qvarnström A, et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491:756–760.

Feder JL, Gejji R, Yeaman S, Nosil P. 2012. Establishment of new mutations under divergence and genome hitchhiking. *Philos Trans R Soc Lond B Biol Sci.* 367:461–474.

Ferree PM, Barbash DA. 2009. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS Biol.* 7:e1000234.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42:D749–D755.

Forejt J, Iványi P. 1974. Genetic studies on male sterility of hybrids between laboratory and wild mice (*Mus musculus* L.). *Genet Res.* 24: 189–206.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.

Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, Bulatova N, Ziv Y, Nachman MW. 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, Basset P, Smith KL, Nachman MW. 2011. Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. *Mol Ecol.* 20:4722–4736.

Godfrey PA, Malnic B, Buck LB. 2004. The mouse olfactory receptor gene family. *Proc Natl Acad Sci U S A.* 101:2156–2161.

Gompert Z, Buerkle CA. 2011. Bayesian estimation of genomic clines. *Mol Ecol.* 20:2111–2127.

Gompert Z, Buerkle CA. 2012. bgc: software for Bayesian estimation of genomic clines. *Mol Ecol Resour.* 12:1168–1176.

Good JM, Dean MD, Nachman MW. 2008. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics* 179:2213–2228.

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.

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

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol Ecol.* 22:4606–4618.

Janoušek V, Wang L, Luzynski K, Dufková P, Vyskočilová MM, Nachman MW, Munclinger P, Macholán M, Piálek J, Tucker PK. 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.

Kim PM, Korbel JO, Gerstein MB. 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A.* 104:20274–20279.

Lê S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. *J Stat Softw.* 25:1–18.

Macholán M, Baird SJ, Munclinger P, Dufková P, Bímová B, Piálek J. 2008. Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone? BMC Evol Biol. *8* 271.

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

Muller HJ. 1940. Bearing of the *Drosophila* work on systematics. In: Huxley JS, editor. The new systematics. Oxford: Clarendon Press. p. 185–268.

Muller HJ. 1942. Isolating mechanisms, evolution, and temperature. *Biol Symp.* 6:71–125.

Nachman MW, Payseur BA. 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.

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

Nosil P, Feder JL. 2012. Genomic divergence during speciation: causes and consequences. *Philos Trans R Soc Lond B Biol Sci.* 367:332–342.

Nosil P, Funk DJ, Ortiz-Barrientos D. 2009. Divergent selection and heterogeneous genomic divergence. *Mol Ecol.* 18:375–402.

Orr HA, Masly JP, Presgraves DC. 2004. Speciation genes. *Curr Opin Genet Dev.* 14:675–679.

Payseur BA. 2010. Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Mol Ecol Resour.* 10:806–820.

Phadnis N, Orr HA. 2009. A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. *Science* 323:376–379.

Phifer-Rixey M, Bomhoff M, Nachman MW. 2014. Genome-wide patterns of differentiation among house mouse subspecies. *Genetics* 198:283–297.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Development Core Team. 2014. nlme: linear and nonlinear mixed effects models. R package

version 3.1-117. Available from: http://CRAN.R-project.org/package=nlme.
Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nat Rev Genet.* 11:175–180.
Presgraves DC, Balagopalan L, Abmayr SM, Orr HA. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* 423:715–719.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
R Development Core Team. 2014. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, Bowers JE, Burke JM, Rieseberg LH. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat Commun.* 4: 1827.
Ruegg K, Anderson EC, Boone J, Pouls J, Smith TB. 2014. A role for migration-linked genes and genomic islands in divergence of a songbird. *Mol Ecol.* 23:4757–4769.
Salcedo T, Geraldes A, Nachman MW. 2007. Nucleotide variation in wild and inbred mice. *Genetics* 177:2277–2291.
Sánchez-Pla A, Salicrú M, Ocaña J. 2007. Statistical methods for the analysis of high-throughput data based on functional profiles derived from the gene ontology. *J Stat Plan Inference.* 137: 3975–3989.
Sarkar D. 2008. Lattice: multivariate data visualization with R. New York: Springer.
Sawamura K, Yamamoto MT. 1997. Characterization of a reproductive isolation gene, zygotic hybrid rescue, of *Drosophila melanogaster* by using minichromosomes. *Heredity* 79:97–103.
Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Saetre GP, Bank C, Brännström A, et al. 2014. Genomics and the origin of species. *Nat Rev Genet.* 15:176–192.
Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, Mott R, Flint J. 2006. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol.* 4:e395.
Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. 2012. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genet.* 8:e1002891.
Storchová R, Gregorová S, Buckiová D, Kyselová V, Divina P, Forejt J. 2004. Genetic analysis of X-linked hybrid sterility in the house mouse. *Mamm Genome.* 15:515–524.
Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22:1540–1542.
Tan PN, Steinbach M, Kumar V. 2005. Introduction to data mining. Boston: Addison-Wesley Longman Publishing Co., Inc.
Tang S, Presgraves DC. 2009. Evolution of the *Drosophila* nuclear pore complex results in multiple hybrid incompatibilities. *Science* 323: 779–782.
Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE, Krenz JG, Sans-Fuentes MA, Nachman MW, Tucker PK. 2008. Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res.* 18:67–76.
Teeter KC, Thibodeau LM, Gompert Z, Buerkle CA, Nachman MW, Tucker PK. 2010. The variable genomic architecture of isolation between hybridizing species of house mice. *Evolution* 64:472–485.
Ting CT, Tsaur SC, Wu ML, Wu CI. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282:1501–1504.
Turner LM, Schwahn DJ, Harr B. 2012. Reduced male fertility is common but highly variable in form and severity in a natural house mouse hybrid zone. *Evolution* 66:443–458.
Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* 3:e285.
Vošlajerová-Bímová B, Macholán M, Baird SJ, Munclinger P, Dufková P, Laukaitis CM, Karn RC, Luzynski K, Tucker PK, Piálek J. 2011. Reinforcement selection acting on the European house mouse hybrid zone. *Mol Ecol.* 20:2403–2424.
Vyskočilová M, Pražanová G, Piálek J. 2009. Polymorphism in hybrid male sterility in wild-derived *Mus musculus musculus* strains on proximal chromosome 17. *Mamm Genome.* 20: 83–91.
Vyskočilová M, Trachtulec Z, Forejt J, Piálek J. 2005. Does geography matter in hybrid sterility in house mice? *Biol J Linn Soc* 84: 663–674.
Wang L, Luzynski K, Pool JE, Janoušek V, Dufková P, Vyskočilová MM, Teeter KC, Nachman MW, Munclinger P, Macholán M, et al. 2011. Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. *Mol Ecol.* 20: 2985–3000.
White MA, Steffy B, Wiltshire T, Payseur BA. 2011. Genetic dissection of a key reproductive barrier between nascent species of house mice. *Genetics* 189:289–304.
Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH, Nachman MW, Piálek J, et al. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet.* 43:648–655.