Molecular Ecology (2009) **18**, 1207–1224 doi: 10.1111/j.1365-294X.2009.04098.x

A powerful regression-based method for admixture mapping of isolation across the genome of hybrids

ZACHARIAH GOMPERT and C. ALEX BUERKLE

Department of Botany and Program in Ecology, University of Wyoming, Department 3165, 1000 E. University Ave., Laramie, WY 82071, USA

Abstract

We propose a novel method that uses natural admixture between divergent lineages (hybridization) to investigate the genetic architecture of reproductive isolation and adaptive introgression. Our method employs multinomial regression to estimate genomic clines and to quantify introgression for individual loci relative to the genomic background (clines in genotype frequency along a genomic admixture gradient). Loci with patterns of introgression that deviate significantly from null expectations based on the remainder of the genome are potentially subject to selection and thus of interest to understanding adaptation and the evolution of reproductive isolation. Using simulations, we show that different forms of selection modify these genomic clines in predictable ways and that our method has good power to detect moderate to strong selection for multiple forms of selection. Using individual-based simulations, we demonstrate that our method generally has a low false positive rate, except when genetic drift is particularly pronounced (e.g. low population size, low migration rates from parental populations, and substantial time since initial admixture). Additional individual-based simulations reveal that moderate selection against heterozygotes can be detected as much as 50 cM away from the focal locus directly experiencing selection, but is not detected at unlinked loci. Finally, we apply our analytical method to previously published data sets from a mouse (*Mus musculus* and *M. domesticus*) and two sunflower (*Helianthus petiolaris* and *H. annuus*) hybrid zones. This method should be applicable to numerous species that are currently the focus of research in evolution and ecology and should help bring about new insights regarding the processes underlying the origin and maintenance of biological diversity.

Keywords: admixture, genetic mapping, hybridization, introgression, reproductive isolation

Received 22 October 2008; revision received 5 December 2008; accepted 11 December 2008

Introduction

Hybrid zones are recognized as ‘natural laboratories’ that provide an opportunity to investigate the evolution of reproductive isolation and thus the origin of species (Endler 1977; Hewitt 1988; Harrison 1990). For example, hybridization between divergent gene pools can generate recombinant individuals that provide the basis for examining the genetic architecture of reproductive isolation between populations or species (Rieseberg *et al*. 1999b; Buerkle & Rieseberg 2001; Rogers *et al*. 2001; Lexer *et al*. 2007). Information on the genetic architecture of reproductive isolation may allow us to discriminate among competing models of speciation and infer the sequence and class of genetic changes that accompany, or even facilitate, speciation (Bradshaw *et al*. 1995; Wu 2001; Coyne & Orr 2004; Martin *et al*. 2008; Minder & Widmer 2008).

The use of natural admixture or hybridization to identify genomic regions that contribute to reproductive isolation between divergent gene pools is complementary to the use of artificial crosses. Artificial crosses have identified genomic regions associated with divergent phenotypes and hybrid inferiority (Coyne & Orr 1998; Liu *et al*. 1996; Mackay 2001; Good *et al*. 2008a; Simon *et al*. 2008); however, the precise fitness consequences of alternative alleles at these genomic regions in nature is often ambiguous. Conversely, the distribution of genotypic variation in hybrid

Correspondence: C. Alex Buerkle, Fax: 307-766-2851; E-mail: buerkle@uwyo.edu

zones is directly influenced by selection in nature, and thus, can be used to identify genomic regions that limit gene flow under natural conditions (Rieseberg *et al*. 1999b). Unlike many artificial crosses, natural admixture often involves multiple generations of crossing and genetic recombination, which facilitates fine-scale mapping of reproductive isolation (Rieseberg & Buerkle 2002; Lexer *et al*. 2007; Buerkle & Lexer 2008). Furthermore, natural admixture may allow for mapping isolation in species that are not amenable to laboratory crosses (e.g. due to large physical size or long generation time).

Previous approaches to utilizing natural admixture to investigate the genetic architecture of reproductive isolation have examined population allele frequencies or individual genotypes as a function of geographic location (e.g. Barton & Hewitt 1985; Mallet *et al*. 1990; Barton & Gale 1993; Porter *et al*. 1997; Yanchukov *et al*. 2006; Teeter *et al*. 2008). These geographic cline methods have contributed substantially to our understanding of hybridization and reproductive isolation. For example, the number of loci contributing to reproductive isolation has been estimated for a number of hybrid zones based on multilocus-geographic clines. These studies have suggested that the number of loci contributing to reproductive isolation varies substantially, from just a few (Porter *et al*. 1997) to 50 or more (Szymura & Barton 1991; Macholan *et al*. 2007). Despite their overall utility, this class of methods assumes a restrictive model of selection (additive underdominance) and may not be applicable for many hybridizing taxa, particularly those that form 'mosaic' hybrid zones (e.g. *Bombina*, *Populus*, and *Vermivora*; Vines *et al*. 2003; Lexer *et al*. 2007; Vallender *et al*. 2007).

Natural admixture can also be used to study hybrid vigour and adaptive introgression. Previous studies have found evidence of increased fitness of hybrid individuals due to both overdominance and recombinant hybrid vigour (Rieseberg *et al*. 1999a; Fitzpatrick & Shaffer 2007). Additionally, hybrids can serve as intermediaries for the transfer of adaptive genetic variation between parental populations (Rieseberg & Wendel 1993; Martin *et al*. 2006; Whitney *et al*. 2006). As adaptation may be limited by the rate of mutation (e.g. Lenski *et al*. 1991; Burch & Chao 1999), introgression has the potential to speed adaptation. While hybrid vigour and adaptive introgression are likely important evolutionary phenomena, unlike reproductive isolation, a general conceptual framework for their study in the context of natural hybridization has not been developed.

We propose a novel 'genomic clines method' for investigating the genetic architecture of reproductive isolation and adaptive introgression between divergent gene pools that utilizes genome-wide admixture as a basis for expectations at individual loci. This individual-based method uses multinomial regression to predict the probability of a given genotype as a function of genome-wide admixture (e.g. hybrid index; Buerkle 2005). These probabilities can be equated with clines in genotypic frequency along a genomic admixture gradient. The level of genome-wide admixture will be only infinitesimally influenced by individual loci and is likely to approximate the pattern of neutral introgression. Therefore, deviations at specific loci from neutral expectations based on genome-wide admixture implicate selective forces acting on the individual loci or closely linked genes. Alternatively, under circumstances that are conducive to substantial genetic drift in hybrids (e.g. small population size), deviations at focal loci can be generated stochastically. A detailed comparison of patterns of introgression for linked and unlinked markers allows discrimination between these hypotheses. This method is particularly suited for identifying intrinsic isolating barriers, which commonly contribute to reproductive isolation (Presgraves 2002; Coyne & Orr 2004; Fitzpatrick 2004). Additionally, as this is a regression-based method, geographic and or environmental variables can easily be incorporated into the model as co-variates to facilitate the identification of genomic regions subject to extrinsic selection pressures.

In this study, we present the genomic clines method for the detection of loci with patterns of introgression that depart from expectations based on the remainder of the genome and may be subject to different types of selection in hybrids. We provide details of the underlying models and evaluate the method's performance (power to detect selected loci and rate of false-positives) with simulated data. We also illustrate the analytical approach by applying it to two previously published data sets and use this as a basis to discuss the prospects for future empirical research.

Methods

Hybrid index and allelic classes

We quantify genome-wide admixture for each individual using a hybrid index. For markers with fixed differences between the parental populations, the hybrid index is simply the proportion of alleles inherited from each of the parental populations (Buerkle 2005). When loci do not exhibit fixed differences between the parental populations, the hybrid index takes into account uncertainty in inheritance. Hybrid indexes are roughly equivalent to Bayesian admixture proportions (e.g. *Q* from Structure; Pritchard *et al*. 2000) when two populations are assumed. However, unlike Bayesian admixture proportions, hybrid indexes are based on parental populations that are defined a priori to estimate parental allele frequencies. Parental populations can be identified using population genetic analyses (e.g. Bayesian assignment tests; Pritchard *et al*. 2000), as well as phenotypic and distributional data. We use hybrid indexes instead of Bayesian admixture proportions, as more experience exists with the former in hybrid zones, and their interpretation is simple when parental taxa are well-defined (Rieseberg *et al*. 1998, 1999b; Buerkle 2005).

To maximize the accuracy with which genomic clines can be estimated and the ease with which they can be interpreted, for each locus with more than two alleles, we combine alleles into two allelic classes with frequency differentials between populations (δ; Gregorius & Roberds 1986; Zhu *et al*. 2005) equal to those observed when each allele is considered separately. This procedure reduces multi-allelic data to a bi-allelic classification, without a loss of information or distortion of the relationship between the parental populations. Individuals are recognized as interclass heterozygotes or homozygotes. This procedure has been used previously in admixture mapping (Lexer *et al*. 2007) and is analogous to single-locus analyses of ancestry that are part of the linkage model of Structure (Falush *et al*. 2003).

Whereas our method was developed with reference to codominant molecular markers, it can easily be adapted to dominant data. This should prove useful for non-model systems for which obtaining a data set with dominant markers (e.g. amplified fragment length polymorphisms) may be much more feasible than obtaining a data set with a large number of codominant markers. We describe the application of this method to dominant markers below (see Analysis of empirical data sets, *Helianthus* hybrid zones).

Regression model

Our genomic clines method uses multinomial regression to estimate individual-based clines in genotypic frequency along a genomic admixture gradient (i.e. hybrid index). For each marker locus, the log probabilities of genotypes containing the allele at higher frequency in *population 1* (A_1) relative to homozygous genotypes containing the allele at higher frequency in *population 2* (A_2) are estimated as a function of hybrid index (h) in logistic regression models:

$$\ln\frac{f_{A_1A_1}}{f_{A_2A_2}} = \alpha_{A_1A_1} + \beta_{A_1A_1}h$$

$$\ln\frac{f_{A_1A_2}}{f_{A_2A_2}} = \alpha_{A_1A_2} + \beta_{A_1A_2}h$$

After rearrangement of terms (as in Tan *et al*. 2005), these probability distributions give the conditional probability of the observed genotypes across all individuals for each locus given the multinomial regression model, denoted $Pr(X \mid M_1)$, where X is the vector of observed genotypes and M_1 is the regression model. This conditional probability is equivalent to the likelihood of the regression model, $L(M_1 \mid X)$.

For each locus, we can determine whether the estimated genomic clines are consistent with a null model of neutral introgression. We simulate populations under the null model by permuting the observed genotypes for each individual among loci. This procedure retains the overall genomic and spatial structure of the admixed population, while treating individual loci as interchangeable. The modelling of loci as exchangeable is consistent with a model of introgression in which each locus contributes an equal, small amount to the barrier to gene exchange. We use multinomial regression on the permuted data to estimate the probability of observing each genotype as a function of hybrid index. This permutation procedure is repeated a large number of times (e.g. 1000) to calculate the mean probability of observing each genotype for each individual across all loci under the null model of neutral introgression (M_0). The conditional probability of the observed genotypic data under this null model of neutral introgression is then calculated, as $Pr(X \mid M_0)$, which is equivalent to the likelihood of the null model given the observed data, $L(M_0 \mid X)$. This allows us to calculate the log-likelihood ratio of the best-fit model (M_1) to the null model (M_0) for the observed data.

This permutation-based procedure for generating null populations assumes that the allele frequencies in the parental populations (*population 1* and *population 2*) are similar for all loci examined, and thus the probability of a given genotype under the null model of neutral introgression is the same for all loci. This assumption is easily met if all loci examined exhibit fixed differences between the parental populations. Choosing markers with fixed differences is increasingly feasible in studies of hybrid zones as the number of highly variable molecular markers available continues to increase (e.g. Teeter *et al*. 2008). However, even when markers do not display fixed differences, this assumption can be met by choosing markers with similar allele frequencies in the parental populations; small differences in allele frequencies should not be a problem for the modelling assumption that they are interchangeable. The choice of informative loci with high allele frequency differentials (including fixed differences) between parental populations does not introduce any bias, as these differences do not imply an association between fitness and these loci in hybrids. Population genetic differentiation alone is not an indicator of selection and methods that use genome-scans of population differentiation to detect loci with a history of selection do so by identifying outlier loci that exhibit exceptional differentiation for a given level of polymorphism (subpopulation heterozygosity; Beaumont & Balding 2004; Beaumont 2005).

Additionally, we have developed an alternative parametric procedure for estimating a null model and generating null populations when the assumption about the uniformity of allele frequency differences across loci cannot be met.

This procedure utilizes the allele frequency estimates for a locus from the parental populations and the genome-wide hybrid index and heterozygosity of individuals to calculate the expected probability of genotypes (a related method was previously applied in Lexer *et al*. 2007). For diploid organisms, the expected frequency of the A_1 allele is given by: $E(f_{A_1}) = f_{A_1,\text{pop1}} + (f_{A_1,\text{pop2}} - f_{A_1,\text{pop1}})h$, where $f_{A_1,\text{pop1}}$ and $f_{A_1,\text{pop2}}$ denote the frequencies of the A_1 allele in *population 1* and *population 2*, respectively. The allele at a higher frequency in *population 1* (population with $h = 0$) is designated A_1. For a locus with fixed differences between *population 1* and *population 2*, this reduces to $E(f_{A_1}) = 1 - h$. We then calculate the expected distribution of genotypes under the null model on the basis of these allele frequency estimates and each individual's observed interclass heterozygosity as follows:

$$E(f_{A_1A_1}) = E(f_{A_1})^2 - (\Delta_H \times f_{A_1A_1}) \qquad \text{(eqn 1)}$$

$$E(f_{A_1A_2}) = 2E(f_{A_1})(1 - E(f_{A_1})) + \Delta_H \qquad \text{(eqn 2)}$$

$$E(f_{A_2A_2}) = (1 - E(f_{A_1}))^2 - (\Delta_H \times f_{A_2A_2}). \qquad \text{(eqn 3)}$$

In equations 1–3, $E(f_{A_1A_1})$, $E(f_{A_1A_2})$, and $E(f_{A_2A_2})$ denote the expected frequencies of individuals homozygous for the A_1 allele, heterozygous individuals, and individuals homozygous for the A_2 allele, respectively, and Δ_H is equal to an individual's observed mean interclass heterozygosity minus an individual's expected mean interclass heterozygosity $\{\Delta_H = H_O - 2E(f_{A_1})[1 - E(f_{A_1})]\}$. We construct a simulated admixed population based on this expected genotype frequency distribution with individual hybrid index and interclass heterozygosity values equal to those of our real population, but with a population size sufficiently large to exclude sampling error. We then use multinomial regression to obtain a null model (M_0) of the probability of observing each genotype as a function of hybrid index for the large simulated population. The conditional probability of our observed genotypic data under this null model of neutral introgression is then calculated, as $Pr(X \mid M_0)$, which is equivalent to $L(M_0 \mid X)$. This allows us to calculate the log-likelihood ratios of M_1 to M_0 for our observed data in a similar manner to the permutation-based procedure described above.

For both the permutation and parametric methods, to determine whether the log-likelihood ratio $L(M_1 \mid X)/L(M_0 \mid X)$ is greater than would be expected by chance, we simulate a distribution of log-likelihood ratios where all deviations of the observed genomic clines from neutral expectations are due to sampling error. For the permutation method, we do this by repeatedly permuting each individual's observed genotypes as described above and using multinomial regression to estimate genomic clines. Similarly, for the parametric approach, we repeatedly sample n genotypes as a function of hybrid index, interspecific heterozygosity, and the parental allele frequencies assuming neutral introgression (as in equations 1–3), where n is the number of potentially admixed individuals in the original data set (i.e. individuals from parental populations are not included). For both approaches, we calculate the probability of each replicate data set given the regression model derived from that data set, M_N, as well the idealized null model, M_0. This allows us to calculate the log-likelihood ratio of M_N relative to M_0 for each of the simulated data sets, and thus to obtain a distribution of likelihood ratios under a null model. This null distribution of likelihood ratios is the basis for comparison to the log-likelihood ratio statistic from the observed data set ($\ln L(M_1 \mid X)/L(M_0 \mid X)$) and allows significance testing and rank ordering of loci on the basis of the magnitude of deviations.

The parametric procedure is not expected to capture completely the genomic composition of an admixed population and may lead to an increase in the rate of false positives, particularly when spatial population structure is present in the area of admixture. In contrast, the permutation procedure retains the population genetic structure (or lack thereof) of the sampled hybrids. Thus, the permutation method is expected to be preferred when parental allele frequencies are homogeneous among loci. The permutation and parametric approaches both assume that most of the genome is nearly neutral or at minimum not experiencing the same form and strength of selection. If the entire genome is subject to a given type and magnitude of selection, this method will not detect deviations from the neutral, null model, as genotypic expectations in the null model are generated on the basis of the genome. This situation would be encountered if isolation between parental taxa were nearly complete and the lack of recombination between chromosomal blocks with different parental ancestry would prevent mapping of any type. However, we do not view this as a significant limitation, as we are interested in the contributions of individual loci to isolation, and in cases in which isolation is not complete and the genic basis of isolation can be dissected.

All regressions and plots were performed in R (R Development Core Team 2008), using the *nnet* (Venables & Ripley 2002) and *genetics* packages and additional code by the authors (available upon request). The analytical methods will be implemented and distributed in a forthcoming R package (*introgress*). Simulations were performed in R and in specialized software written in C (the model of junctions, below).

Simulations: genomic clines and power for different selection models

We investigated the effects of selection on the expected genomic clines for admixed individuals under four

Table 1 Simple models of selection were compared to a model of neutral mixing between parental genomes. Each selection model incorporates a selection parameter (s) that scales the expected genotypic frequencies in a manner appropriate for the genotypic effects on fitness. The basic, neutral model here is slightly simpler than equations 1–3 because we made the simplifying assumptions that parental taxa possessed distinct alleles and that the observed mean interclass heterozygosity was equal to the expected mean interclass heterozygosity ($\Delta_H = 0$)

Model	$E(f_{A_1A_1})$	$E(f_{A_1A_1})$	$E(f_{A_1A_1})$
Neutrality	$(1 - h)^2$	$2(1 - h)h$	h^2
Underdominance	$(1 - h)^2$	$2(1 - h)h\ (1 - s)$	h^2
Overdominance	$(1 - h)^2 \cdot (1 - s)$	$2(1 - h)h$	$h^2\ (1 - s)$

Epistasis	$(1 - h)^2 \cdot (1 - hs)$	if $h < 0.5$ $2(1 - h)h \cdot [1 - (1 - h)s]$
		if $h \geq 0.5$ $2(1 - h)h \cdot (1 - hs)$
			$h^2\ [1 - (1 - h)s]$

Directional selection	$(1 - h)^2 \cdot (1 - s)$	$2(1 - h)h \cdot (1 - s/2)$	h^2

different selection models: underdominance, overdominance, epistasis (single-locus interaction with the genetic background), and directional selection (Table 1). Our model of epistasis assumes that the fitness of a given genotype at a locus under selection is dependent on an individual's overall genomic composition, such that the A_1A_1 and A_2A_2 genotypes have the highest fitness in pure *population 1* and *population 2* genomic backgrounds respectively, and the A_1A_2 genotype has the highest fitness in intermediate genomic backgrounds. Underdominance and epistasis are the most relevant models with respect to the genetic architecture of reproductive isolation, but overdominance and directional selection are relevant for the study of hybrid vigour and adaptive introgression. For each of these selection models, the expected genotype frequencies deviate from a simple model of neutrality according to a selection parameter (s; Table 1). The selection parameter (s) can be thought of either as the strength of selection in a single generation, or, particularly in the case of the directional selection model, as the cumulative effect of selection over some number of generations. For underdominance, overdominance and epistasis selection, these simple models do not capture the potential for these modes of selection to change allele frequencies through time.

We performed simulations to assess qualitatively how genomic clines produced under each of these selection models deviated from genomic clines expected under a model of neutral introgression. The results of these simulations can be used as a heuristic for understanding deviations observed in data from hybrid populations. We first generated an admixed population by randomly sampling 100 individuals, each with a hybrid index drawn from a uniform distribution bounded by 0 and 1. We then sampled genotypes based on the hybrid indexes to create 1000 replicate populations of 100 individuals under the null model of neutral introgression according to equations 1–3, and assuming fixed allelic differences between the two parental populations ($f_{A_1,pop1} = 1\ and\ f_{A_1,pop2} = 0$) and $\Delta_H = 0$ (these simplifying assumptions yield the functions in Table 1). This simulation scheme models a hybrid zone without spatial structure. Multinomial regression was used to estimate the predicted genotypic probabilities as a function of hybrid index for each of the 1000 replicates. We then generated another 1000 replicate populations using the same hybrid index values, but with the expected genotype frequencies adjusted according to the four selection models (Table 1). Sampling was conducted with s set to 0.25, 0.5, and 0.75. Multinomial regression was used to estimate the predicted genotypic probabilities as a function of hybrid index for each of the 1000 replicates under each model of selection for all three values of s. To visualize the influence of selection on the predicted genotypic probabilities, we plotted the empirical 95 % confidence intervals for the predicted probabilities of the A_1A_1 and the A_1A_2 genotypes for each of the models.

We performed additional simulations to estimate the power of our genomic clines method to detect locus-specific deviations from a null model of neutral introgression based on genome-wide admixture. Populations were simulated in a similar manner to that described above, but with the following modifications. We varied the size (n) of the sampled admixed population from 50 to 200 by steps of 50, and for each sample size, we varied the strength of selection, s, from 0.1 to 0.9 by steps of 0.1, for each of the four previously described models of selection. For each selection model (i.e. type of selection, s, and n), we simulated 1000 populations with one locus under selection and 999 unlinked neutral loci. For each simulated population, we estimated genomic clines for the locus under selection using multinomial regression. We then performed 1000 permutations of the genotypic data for each simulated population in accordance with our permutation-based procedure to obtain mean genotypic probabilities for the focal locus under a null model of neutral introgression. We then calculated the log-likelihood ratio of M_1 to M_0 given the observed data at the focal locus, as described above. Next, we obtained a distribution of expected log-likelihood ratios given neutral introgression by conducting 1000 additional genotypic permutations for each population and assessing the relative likelihood of a regression model based on the permuted sample to the likelihood of the model based on mean genotypic probabilities given the data at the focal locus for the permuted sample. Finally, for each combination of selection model and sample size, we determined the power to detect selection by calculating the proportion of log-likelihood ratios involving the focal locus under selection that exceeded the 95% quantile of the distribution of log-likelihood ratios obtained from the permuted data.

Simulations: genetic drift and false positives

In an admixed population, genetic drift at individual loci will lead to changes in allele frequencies that are not the result of selection and could mislead our method by identifying false signals of selection (false positives). To assess and quantify this possibility, we simulated admixed populations without selection but allowing for drift and tested for departures from a neutral model of introgression. We modelled diploid hermaphroditic populations of fixed size, with individuals harbouring 10 chromosome pairs, each 1 Morgan in length, and fixed allelic differences between parental lineages across the entire genome. An initial admixed population was created with equal representation of chromosomes from each of two parental lineages. Mating occurred at random within the admixed population with a given rate of migration from each of the parental populations. Similar to the simulations described above, this model simulates a hybrid zone without spatial structure. The simulations were conducted for 50 generations, and 100 individuals were sampled for analysis from the admixed population at 5, 10, 20, 30, and 50 generations. The simulated individuals were assumed to have discrete, non-overlapping generations. The model tracked junctions along each chromosome (i.e. locations where chromosomal segments derived from different parental populations meet; Fisher 1954) rather than markers and thereby maintained full information about the recombination history (as in Buerkle & Rieseberg 2008). At the end of a simulation, each of the individuals was scored for 110 bi-allelic loci, spaced evenly at 10-cM intervals along each of the 10 chromosomes (11 loci per chromosome). Simulations were conducted with the size of the admixed population (N) set to 100, 500, and 1000 and with migration (m, the proportion of gametes in the admixed population each generation that originated from the parental populations) set to 0, 0.05, 0.2, and 0.4.

For each parameter combination, we simulated 100 replicate populations and then used our method to determine the false-positive rate, which here is the proportion of replicate populations with a significant deviation from neutrality ($P < 0.05$) at a focal locus, *c1.f* (where *c1* specifies chromosome 1 and *f* the focal locus, at the centre of chromosome 1). In each case, we used both the parametric and permutation methods to generate populations under a null model of neutrality (1000 replicates for each). Hybrid index and interclass heterozygosity were calculated based on all 110 loci; these 110 loci were also used for permutations.

Simulations: selection and hitchhiking

We assessed the ability of our method to detect selection on an individual locus while allowing for drift, and evaluated the relative potential of these processes to produce patterns of introgression that deviate from expectations given genome-wide admixture. This was done using the simulation method described in the preceding section, but with several modifications. We modelled underdominance or epistatic selection on a focal locus, *c1.f.* For underdominance, the fitness of homozygous individuals at this locus was set to 1 and heterozygous individuals were given a fitness of $1 - s$. For epistatic selection, the fitness of genotypes varied with hybrid index as in Table 1. For both models, we performed simulations with $s = 0.1, 0.3, 0.5, 0.7$, and 0.9. We conducted simulations with both fertility selection (i.e. reduced probability of producing functional gametes) and viability selection (i.e. reduced probability of survivorship of zygotes). The effects of viability selection should be easier to detect than fertility selection, because the distortions relative to Hardy–Weinberg expectations will be seen in the adult population. We fixed the size for the admixed population (N) at 500 and only sampled the admixed population once (this was done at the fifth generation). Migration (m) was set to 0, 0.05, 0.2, and 0.4. We limited these simulations to a single population size and sampling generation to reduce the dimensionality of the simulations. We simulated 100 replicate populations for each combination of parameters (i.e. selection intensity, type of selection, and migration). We then employed our method to determine the proportion of the 100 replicates for each set of simulations that differed significantly from neutral introgression, as described in the preceding paragraph.

To determine the effect of selection at our focal locus (*c1.f*) on other loci due to linkage and genetic hitchhiking, we tested for departures from neutrality at 12 loci using our simulated data with underdominance selection. We examined variation at 10 loci linked to the focal locus (*c1.1*, *c1.2*, ... , *c1.9*, and *c1.10*) and two additional unlinked loci (*c2.6* and *c3.6*, at the centre of chromosomes 2 and 3). Markers *c1.1–c1.5* and *c1.6–c1.10* were spaced every 10 cM along chromosome 1, with *c1.f* at the centre of the chromosome (at the 50 cM position). Tests of non-neutral introgression were conducted as previously described, but using only the permutation method to generate null expectations.

Analysis of empirical data sets

We analysed two previously published data sets using our genomic clines method for investigating the genetic architecture of reproductive isolation and adaptive introgression. One data set consists of 13 diagnostic X-linked, codominant microsatellite markers scored for 254 females sampled from a 174.6-km geographic transect along a European hybrid zone between mouse species (*Mus domesticus* and *M. musculus*; Payseur *et al*. 2004). Payseur *et al*. (2004) found that population allele frequencies varied with geographic location in this hybrid zone, consistent with geographic cline models. The other data set consists of 61 dominant markers [random amplified polymorphic DNA (RAPD)] from narrow (less than 50 m) hybrid zones between sunflower species (*Helianthus* annuus and *H. petiolaris*), a set of three in Nebraska (these were pooled into a single sample of 139 individuals, as before) and one in California (89 individuals; Buerkle & Rieseberg 2001). The *Mus* data set was selected because our results can be compared to previous analyses of the contribution of the X-chromosome to isolation that were conducted using this data set based on classical methods for geographic clines (Payseur *et al.*, 2004). The fact that the *Mus* data set contains only 13 markers is not problematic, as we are concerned only with making inferences regarding patterns of introgression for individual markers relative to the X-chromosome, not the entire genome. Our analysis of the sunflower data allows us to revisit, with a more refined analytical method, the question of the variability among hybrid zones in the genetic architecture of isolating barriers, and demonstrate an extension of our proposed method.

We analysed the *Mus* hybrid zone using our genomic clines method as described above (see Regression model). It was not necessary to define allelic classes, as only two alleles were segregating at each locus (Payseur *et al*. 2004). Hybrid index values (h) were calculated based on the 13 X-chromosome markers. As these markers exhibited fixed differences between *M. domesticus* and *M. musculus,* hybrid index values were calculated simply as the proportion of alleles inherited from *M. domesticus.* Our permutation-based procedure was used to generate expectations under the null model of neutral introgression. Because only X-linked markers were used, this analysis will detect markers that show patterns of introgression that are significantly different than the X-chromosome taken as a whole, not the entire genome.

We analysed the *Helianthus* marker data from each of the hybrid zones using our genomic clines method as described above (see Regression model), but with modifications to deal with dominant markers. For all 61 RAPD markers, the marker bands (PCR product present) were found at high frequency in *H. petiolaris*, but were absent from *H. annuus* populations (Buerkle & Rieseberg 2001). Thus, we calculated the probability of observing no marker band as $E(f_{A_1}) = f_{A_1,\text{annuus}} + (f_{A_1,\text{petiolaris}} - f_{A_1,\text{annuus}})h$, where $f_{A_1,\text{annuus}}$ and $f_{A_1,\text{petiolaris}}$ denote the frequencies of the absence allele in parental *H. annuus* and *H. petiolaris* populations. The markers used did not exhibit fixed differences and marker band frequencies in the parental populations were not considered to be sufficiently similar among the loci to allow use of our permutation-based method (range of absence allele frequencies: $f_{A_1,\text{petiolaris}} = 0 - 0.5$, $f_{A_1,\text{annuus}} = 1$). Therefore, we used our parametric-based method to obtain neutral expectations. As the *Helianthus* hybrid zones were sampled over very small geographic distances (several hundred metres), populations structure is unlikely to be a problem. We obtained maximum-likelihood estimates of hybrid index values for *Helianthus* individuals using the software *hindex* (Buerkle 2005).

To compare patterns of locus-specific introgression between the Nebraska and California *Helianthus* hybrid zones, we contrasted the likelihood of the regression model derived from the California hybrid zone to that from the Nebraska hybrid zone given the data from the California hybrid zone ($L(M_{\text{Cal}} \mid X_{\text{Cal}})$ and $L(M_{\text{Neb}} \mid X_{\text{Cal}})$). The data from the California hybrid zone were used for these comparisons because they had a narrower range of hybrid indexes. Thus, we could estimate probabilities for the observed genotypes for the California individuals without extending the predictions of either regression model beyond the range of hybrid indexes used to generate the models. We calculated the log-likelihood ratio of $L(M_{\text{Cal}} \mid X_{\text{Cal}})$ to $L(M_{\text{Neb}} \mid X_{\text{Cal}})$ for each locus to determine the degree of concordance in patterns of introgression between the two hybrid zones.

Results

Simulations: genomic clines and power for different selection models

The four models of selection (underdominance, overdominance, epistasis, and directional selection with incomplete dominance) shifted the predicted genomic clines as expected, with a more pronounced deviation with increasing intensity of selection (Fig. 1). These simple models for fitness of hybrids produce genomic clines that are distinguishable graphically from one another and from a neutral model, and they provide a basis for interpreting deviations observed in empirical data sets. The underdominance and overdominance models resulted in the expected shifts in the frequency of heterozygotes relative to homozygotes (Fig. 1A, B). Whereas the overall predicted probabilities for each genotype were not substantially altered under the epistasis model, the rate of change in probabilities for each genotype increased, resulting in sharper clines (Fig. 1C).

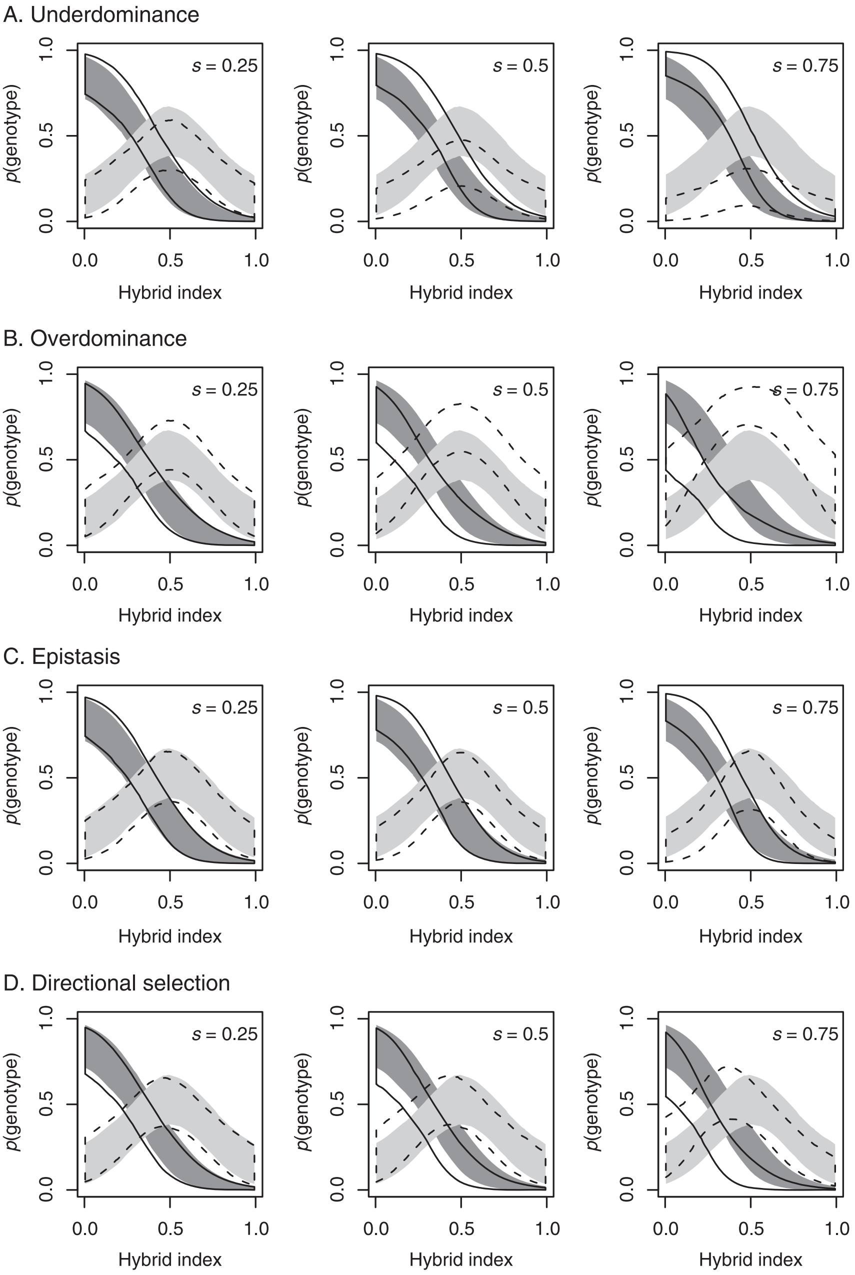

Fig. 1 Different models of fitness and intensities of selection shift genomic clines away from expectations under neutrality. Genomic clines are shown for underdominance (A), overdominance (B), epistasis (C), and directional selection (D). See Table 1 for details on selection models. The selection intensity is given for each model. Solid grey regions represent the 95% confidence intervals for A_1A_1 (dark grey) and A_1A_2 (light grey) genomic clines given neutral introgression. Different forms of selection shift these probabilities, as indicated by the 95% confidence intervals for the A_1A_1 (solid lines) and A_1A_2 (dashed lines) genotypes.

Directional selection against the A_1 allele shifted the A_1A_1 and A_1A_2 probability distributions markedly to the left (Fig. 1D). Overall, the graphical analysis suggests that relatively strong selection would be required to distinguish among different models of selection with confidence.

Quantitative estimates of the power to detect selection varied among the models examined and increased with increasing values of *s* and *n* (Fig. 2). Assuming *s* is equivalent among the different selection models, our genomic clines method has more power to detect deviations from neutral introgression due to underdominance, overdominance, and directional selection than deviations due to epistasis (Fig. 2). This difference in power reflects the varying degree to which these forms of selection alter the expected genomic clines (Fig. 1).

Simulations: genetic drift and false positives

Genetic drift led to a false signal of selection in some instances, but the prevalence of this phenomenon was dependent on the size of the admixed populations, the migration rate from the parental gene pools to the admixed population, and the number of generations since initial admixture (Fig. 3). When the admixed population was

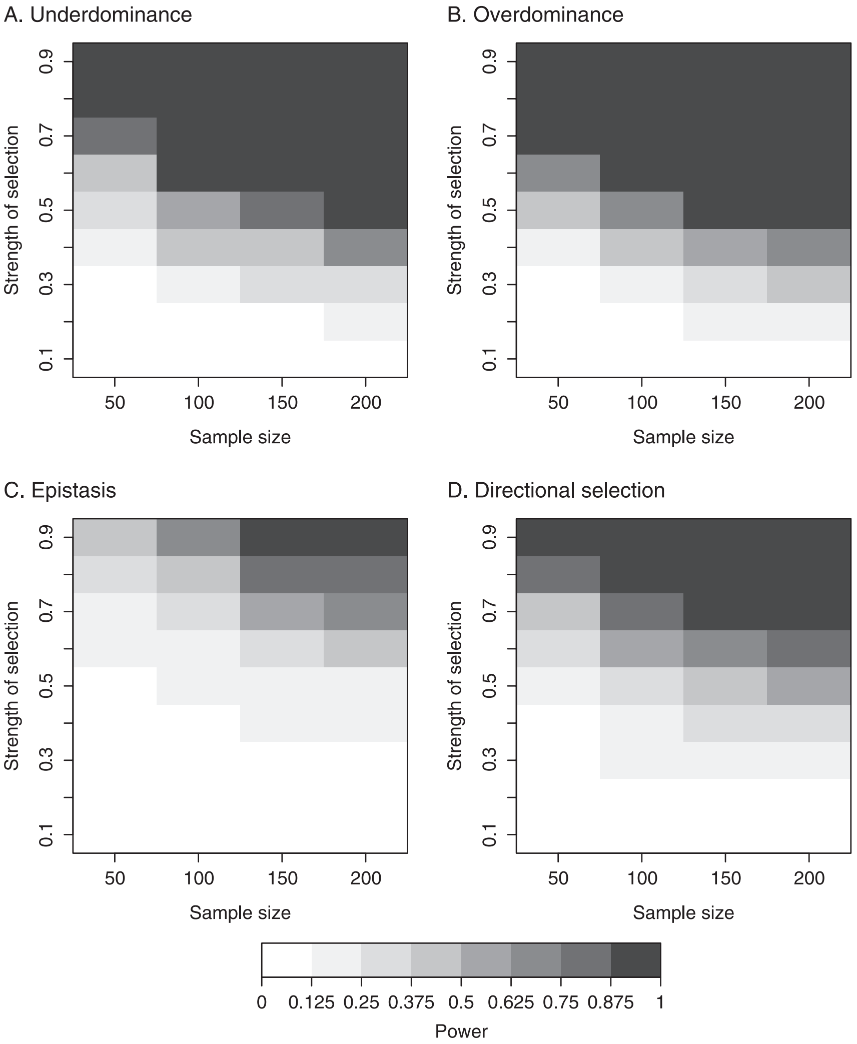

Fig. 2 Power to detect selection as a function of selection intensity and sample size for underdominance (A), overdominance (B), epistasis (C), and directional selection with codominance (D). Model specifications are in Table 1.

composed of 100 individuals, only very high migration (i.e. $m = 0.4$) resulted in the proportion of false positives being equal to or less than 0.1. However, when the size of the admixed population was 500 or 1000 (sample size fixed at 100), the proportion of false positives was greatly reduced (Fig. 3). For these larger population sizes, the proportion of replicates that showed a signal of non-neutral introgression (false positives) at our focal locus was generally less than 0.2 and often between 0.01 and 0.10. Even with population sizes of 500 or 1000, no migration and 50 generations since admixture resulted in a high proportion of replicates with evidence of non-neutral introgression at our focal locus. The proportion of replicates giving significant results was similar for the permutation and parametric methods.

Simulations: selection and hitchhiking

When we simulated admixed populations that experienced underdominance or epistatic selection at a focal locus ($N = 500$, five generations after initial admixture), our ability to detect selection was dependent on the migration rate, strength of selection, and type of selection (fertility or viability; Fig. 4). Our ability to detect viability-based underdominance selection was greater than our ability to detect fertility-based underdominance selection. This difference was most pronounced when the migration rate was high ($m = 0.4$) and the effects of fertility-based selection were reduced, because of the substantial fraction of gametes that were produced by parental individuals with full fitness (Fig. 4). Similarly, epistatic selection resulted in consistently greater departures from neutrality when fitness differences affected the viability of zygotes rather than their fertility (Fig. 4). As expected, stronger selection led to a greater frequency of replicates that were distinguishable from neutral introgression. For underdominance, with $s = 0.9$ and low levels of migration, all 100 replicates produced a significant signal of selection at the focal locus. Epistatic viability selection similarly resulted in nearly all replicates deviating significantly from neutrality when selection was highest ($s = 0.9$).

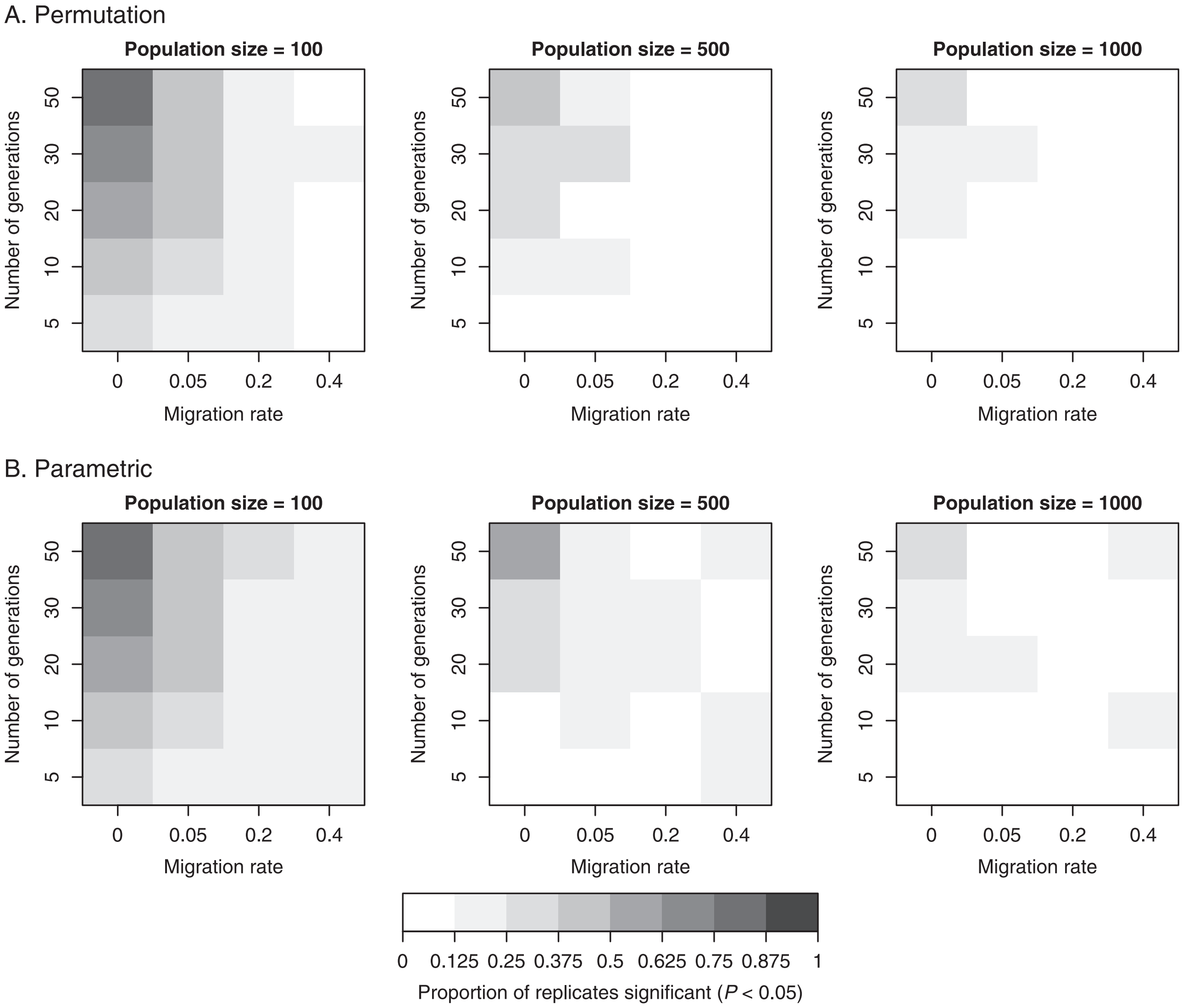

Fig. 3 The probability of detecting non-neutral introgression at a focal locus as a result of genetic drift (false positives) for both the permutation method (A) and the parametric method (B). The effect of drift is greatest when recurrent backcrossing is rare, the size of the admixed population is low, or the number of generations since initial admixture is great (sample size fixed at 100 for all).

In most cases, the rate of detection of true positives (selected loci) was substantially higher than the false-positive rate (neutral loci that varied from expectations due to chance). The exception is epistatic fertility selection, which only leads to substantial departures from neutrality with strong selection and low migration. True positives are represented by the proportion of replicates with significant results under different models of selection (Fig. 4) and the comparable false- positive rate comes from admixed populations of the same size ($N = 500$) and five generations following initial admixture in the absence of selection (Fig. 3). The rate of detection of true and false positives was similar for the permutation and parametric methods (Fig. 3 and additional data not shown).

Underdominance at the focal locus (*c1.f*) affected the probability of detecting selection at linked loci, but not at unlinked loci (Fig. 5). Departures from neutrality were most pronounced at markers *c1.5* and *c1.6*, each 10 cM from the focal locus, and decreased with increasing map distance from *c1.f*. However, even markers *c1.1* and *c1.10*, which were 50 cM from the focal locus, showed a significant signal of selection in a greater proportion of replicates than completely unlinked markers (*c2.6* and *c3.6*; Fig. 5). Loci linked to the selected locus were more likely to depart from neutrality when the migration rate was 0 than when it was 0.4, but this likely reflects an overall greater ability to detect selection in the absence of migration.

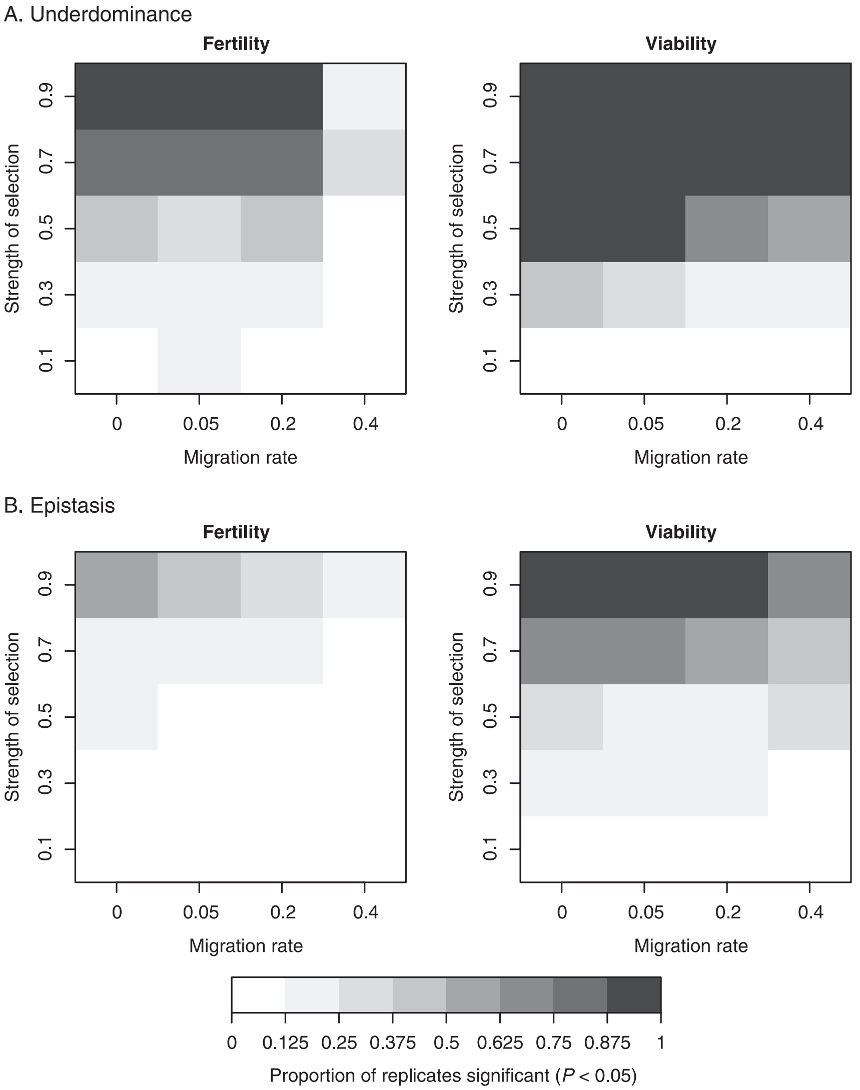

Fig. 4 The probability of detecting true selection at a focal locus that is subject to underdominance (A) or epistatic selection (B) over five generations. Panels on the left give results for fertility selection and panels on the right are for viability selection.

Analysis of empirical data sets

Of the 13 X-linked codominant markers examined for the mouse hybrid zone, eight showed significant deviations from expectations given a null model of neutral introgression ($P < 0.031$). After correcting for multiple testing using the false discovery rate procedure (Benjamini & Hochberg 1995) seven of the eight markers retained their significance. The specific manner in which each locus deviated from neutral expectations varied. For several loci (e.g. *Sep6b*) the homozygote and heterozygote genomic clines were steeper than predicted by the neutral model (Fig. 6), consistent with our model of epistasis (Fig. 1C). Genotypic variation at the locus *DXmit18.2* also involved steep transitions between parental homozygotes, but very few heterozygotes were observed (Fig. 6), consistent with underdominance and potentially strong selection (Fig. 1A). The *Xist* locus exhibited an aberrant pattern of introgression, with almost parallel clines for the homozygous and heterozygous genotypes. Our method also reveals contrasts among loci in terms of directional introgression, with *M. musculus* alleles introgressing farther into the *M. domesticus* genetic background at several loci (e.g. *Emd*, DXmit18.2, *Trrp5*) and biased introgression from *M. domesticus* into the *M. musculus* genetic background at locus *Plp*. These results are generally consistent with the findings of Payseur *et al.* (2004) based on geographic clines methods.

For the *Helianthus* data set, 33 out of 61 (Nebraska, $P < 0.026$) and 24 out of 61 (California, $P < 0.019$) dominant markers deviated from a model of neutral introgression following correction for multiple independent tests (Benjamini & Hochberg 1995). We detected both genomic clines that were steeper and shallower than expected given neutral introgression (Fig. 7). Steep and shallow clines are expected for markers with decreased and increased rates of introgression relative to neutral expectations, respectively.

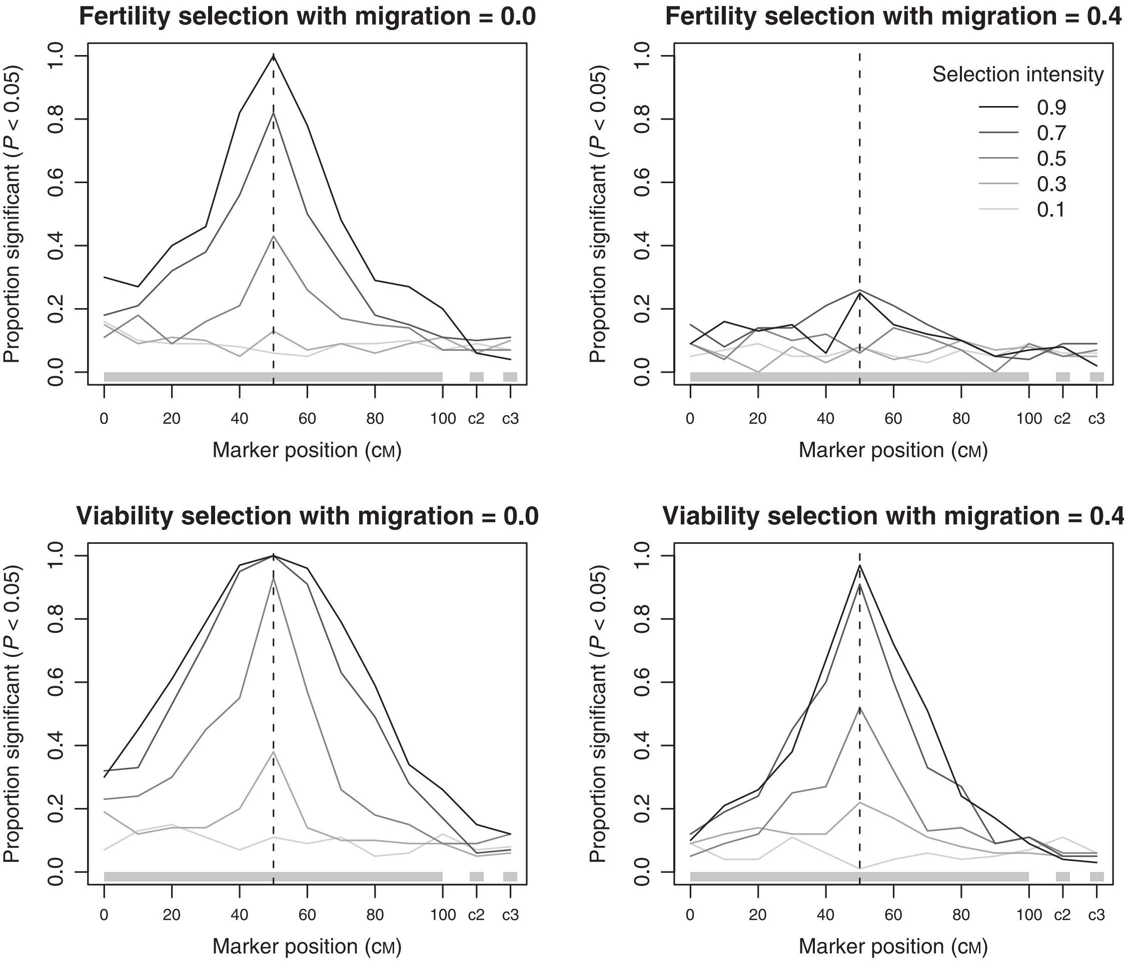

Fig. 5 Plots of the probability of detecting true underdominance at a focal locus after five generations of selection. Probabilities are for ten neighbouring, linked loci, and two unlinked loci. Horizontal grey bars connect physically linked loci, and the dotted vertical line denotes the focal locus that was the target of selection.

Introgression for 45 out of 61 loci did not differ significantly between the California and Nebraska hybrid zones ($\ln L(M_{\mathrm{Cal}} \mid X_{\mathrm{Cal}}) - \ln L(M_{\mathrm{Neb}} \mid X_{\mathrm{Cal}}) < 2$. However, several loci showed substantially larger differences in log likelihoods and very different patterns of introgression in Nebraska and California; the log-likelihood ratio was greater than 8 for five loci and the highest log-likelihood ratio was 72.36 (locus *226-07*; Fig. 7).

Discussion

The results from our analysis of admixed populations simulated under four different selection models (underdominance, overdominance, epistasis, and directional selection with incomplete dominance) indicate that our genomic clines method holds promise for detecting loci that deviate from presumably neutral, genome-wide patterns of introgression. This conclusion is strengthened by the method's demonstrated ability to detect non-neutral introgression in a European hybrid zone between species of mice (*Mus domesticus* and *M. musculus*) and two hybrid zones between sunflower species (*Helianthus annuus* and *H. petiolaris*). Furthermore, our simulation results suggest that different forms of selection affect genomic clines in different and predictable ways. This information should facilitate the interpretation of non-neutral patterns of introgression from hybrid zones in nature. The number of individuals sampled from these simulated admixed populations (50–200 individuals) is within the range of achievable sample sizes from natural hybrid zones. As would be expected, the ability of this method to detect non-neutral introgression at individual loci is highly dependent on both the strength and type of selection that the locus is experiencing. There is little empirical data on the strength of selection experienced by different regions of the genome in hybrid zones, but our analyses of data from *Mus* and *Helianthus* hybrid zones suggest that selection pressures experienced in nature are likely large enough to

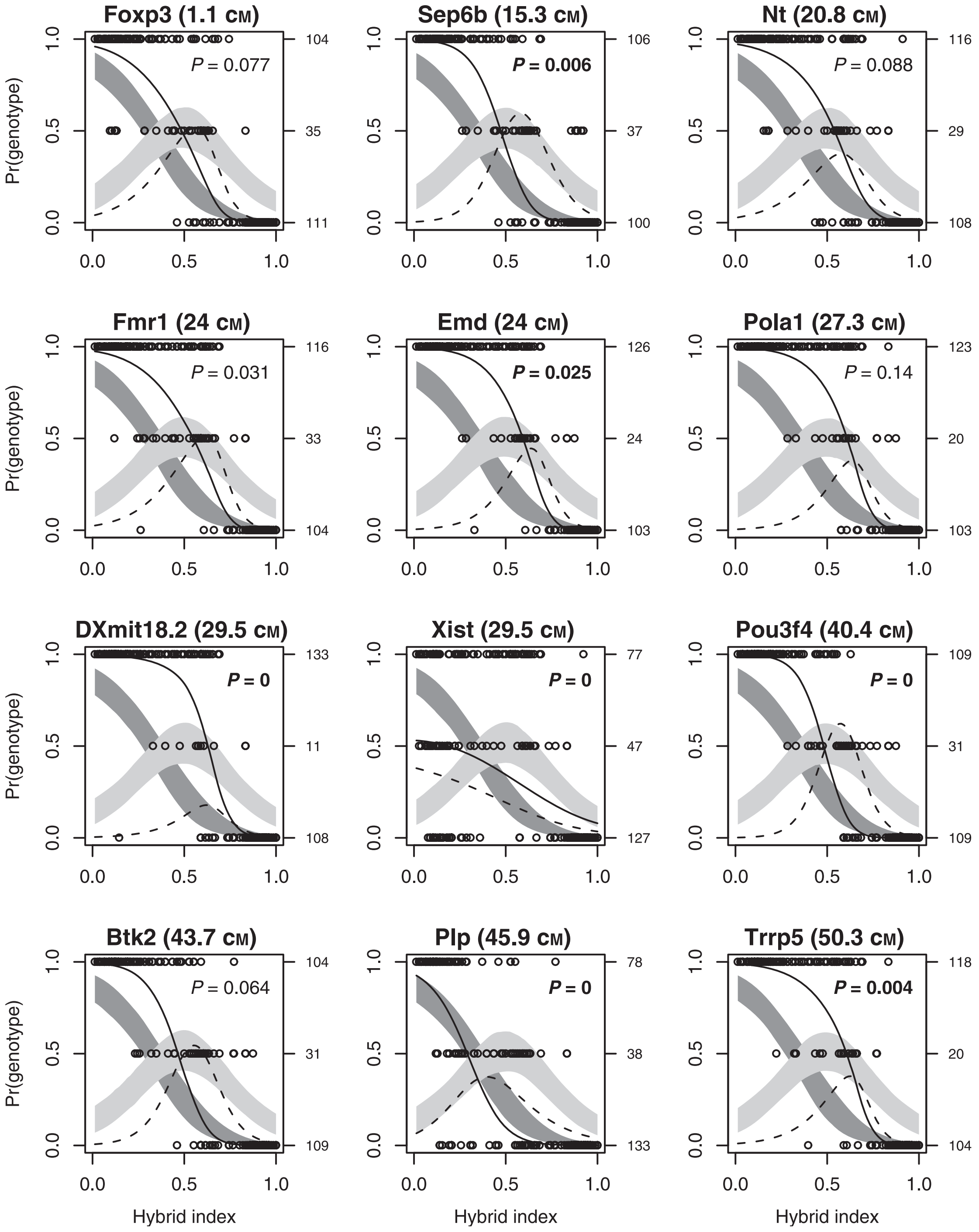

Fig. 6 Genomic clines for markers along the X-chromosome from a hybrid zone between *Mus domesticus* and *Mus musculus* (12 of the 13 loci from Payseur *et al*. 2004). The name of each locus and location on the X-chromosome is given, as is the *P* value for the test of departure from neutrality ($P < 0.031$ indicates significance after FDR correction; in bold). Solid clines represent the 95% confidence intervals for $f_{A_mA_m}$ (dark grey) and $f_{A_mA_d}$ (light grey) genomic clines given neutral introgression. The solid line and dashed lines give the estimated cline based on the observed A_mA_m and A_mA_d genotypes, respectively. Circles indicate the raw genotypic data (A_mA_m on top line, A_mA_d in centre, and A_dA_d along the bottom), with counts of each on the right vertical axis. The hybrid index quantifies the fraction of alleles derived from *Mus domesticus* across all 13 markers.

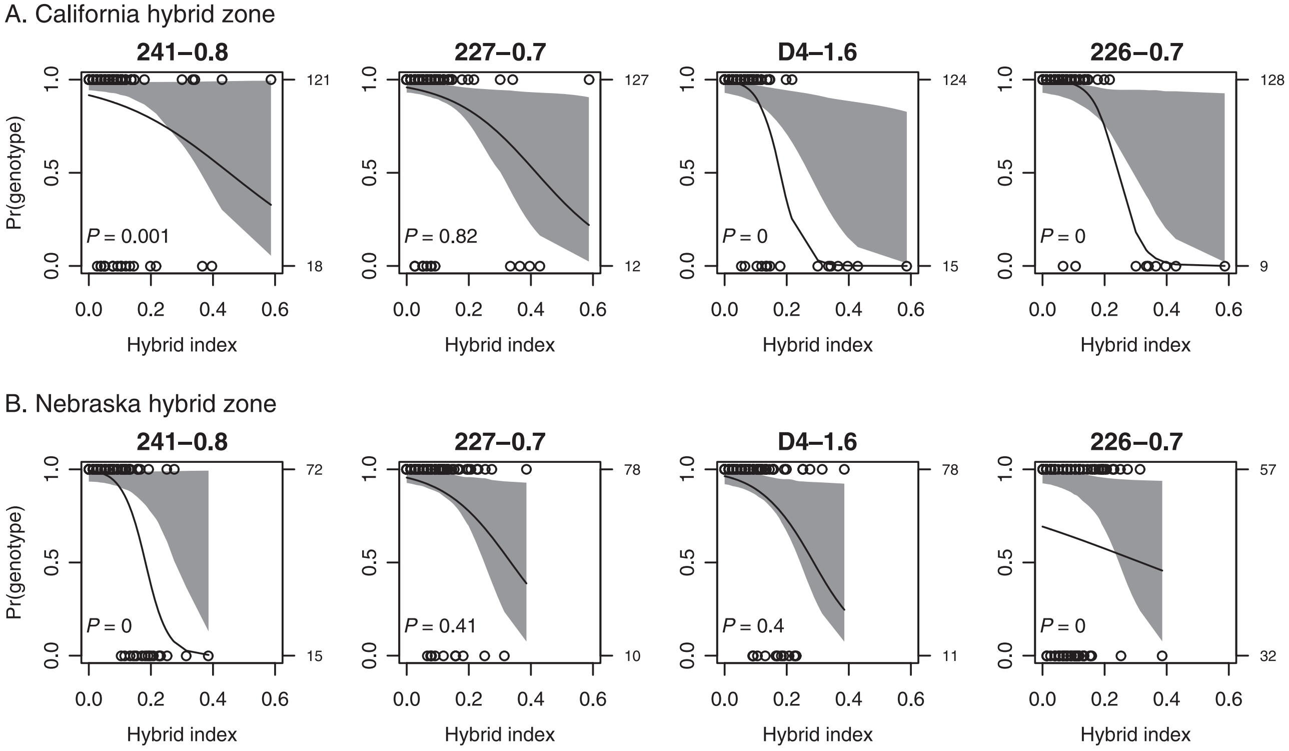

Fig. 7 Representative genomic clines for introgression of dominant markers for two *Helianthus* hybrid zones. The name of each locus and *P* value for the test of departure from neutrality are given. The solid line is the estimated cline based on individuals that lacked the dominant RAPD bands (f_{A_1}). The grey regions represent the 95% confidence intervals for f_{A_1} given neutral introgression. Circles indicate the raw marker data (RAPD 'band absent' on top line and 'band present' along the bottom), with counts of each on the right vertical axis. The similarity between hybrid zones of the observed genomic clines varies substantially among loci. The hybrid index quantifies the fraction of the genome that is derived from *H. petiolaris.*

be readily detected by this method. Additionally, crossing studies have shown that the fertility of F_1 hybrids can be reduced by 90% relative to parental individuals (e.g. Rieseberg 2000), consistent with a model of strong underdominance.

Simulations allowing for genetic drift indicate that demographic stochasticity over time can lead to false positives using our method (i.e. evidence of selection in the absence of selection). The proportion of replicates leading to false positives was similar for the permutation and parametric methods for generating distributions of genotypes under neutrality (Fig. 3). The permutation method is expected to lead to fewer false positives than the parametric method in cases where the parametric method does not adequately capture the genomic structure of the admixed population, such as populations with spatial and genetic substructure. Given that we simulated admixed populations without population structure, it is not surprising that these methods performed equally well. The rate of false positives due to genetic drift was greatest when the size of the admixed population and the migration rate were low ($n = 100$, $m = 0 - 0.05$) and more than 10 generations had passed since the initial admixture event (Fig. 3). These results are not surprising, as genetic drift, like many forms of selection, shifts allele frequencies. Moreover, genetic drift would be expected to be the most severe under precisely the conditions that led to the greatest number of false positives in our simulations. Based on these simulation results, we urge caution in the application of our proposed method to admixed populations where genetic drift is likely to be an important factor. However, simulations of admixed populations that combined the effects of genetic drift and underdominance or epistatic selection indicate that selection is more likely to lead to departures from neutrality than is drift (Fig. 4). Thus, under many conditions, one can safely assume that the majority of deviant loci detected with our method possess patterns of introgression that have been influenced by selection. Information from linked loci may be helpful if multiple loci exhibit similar deviations (Fig. 5). However, neighbouring marker loci need not covary (e.g. Fig. 6), given that the genetic architecture of isolation and the history of recombination in natural hybrid zones may be more complex than what we have modelled.

After five generations of admixture, underdominance at a focal locus was detectable, both at selected and neighbouring loci. Departures from neutrality at loci linked to the underdominant locus stem from the presence of linkage disequilibrium in our admixed populations. Beyond the selected locus, deviations from neutrality were most frequent at the neighbouring loci (10 cM from the focal locus) but could also be detected at more distantly linked loci, 50 cM from the focal locus (Fig. 5). This effect did not extend to two unlinked markers that we analysed (*c2.6* and *c3.6*). The effect of selection on neighbouring loci would decrease with time since the initial admixture event due to the decay of linkage disequilibrium over generations. Linkage disequilibrium could reach an equilibrium value determined by a balance between recombination and continued gene flow from the parental populations, but it is likely that in hybrid zones, rates of migration and the spatial context will be heterogeneous and result in variable levels of linkage disequilibrium. Nonetheless, this finding indicates that even a single underdominant locus can be sufficient to produce patterns of non-neutral introgression for markers scattered across the entire chromosome. One or a few genes under selection could be sufficient to drive or prevent significant introgression of entire chromosomes. Furthermore, if multiple genes and markers on a single chromosome are experiencing different forms of selection (e.g. directional selection vs. underdominance), patterns of observed introgression may be variable among loci or individual loci may not clearly correspond to expectations from any specific model of selection.

Consistent with the findings of Payseur *et al*. (2004) based on the analysis of geographic clines, we found substantial variation among X-linked markers in patterns of introgression for the *Mus* hybrid zone. Variation at seven of the 13 X-linked markers deviated from expectations based on the X-chromosome as a whole. This independence among loci on a single linkage group is remarkable and will not necessarily be found in other systems. In future work, one would ideally estimate genome-wide admixture on the basis of a larger number of additional, unlinked markers. Similar to Payseur *et al*. (2004), we detected increased rates of directional introgression at the *Xist* and *Plp* loci (Fig. 6). In these populations of mice, geographic position is a good predictor of hybrid index of individuals, and thus, the overlap in the results of the different analytical approaches is not surprising. However, contrary to Payseur *et al*. (2004), we found no evidence of selection at the marker *Pola1*. Furthermore, our analysis points to possible underdominance at or near *DXmit18.2* and its potential contribution to isolation, which was not revealed by the methods utilized previously (Payseur *et al*. 2004). The ability to parse genotypic effects in introgression is an important advance over earlier analytical methods that utilized alleles and their frequency in populations as the unit of study (Barton & Hewitt 1985; Barton & Gale 1993; Porter *et al*. 1997). It is noteworthy that the distribution of genotypes at X-linked markers, with substantially more introgression from *M. musculus* into the *M. domesticus* genomic background, would have been difficult to predict on the basis of analyses of male sterility in F_1 and advanced generation backcrosses (Good *et al*. 2008a, b). In both cases, decreases in male fertility and fecundity were associated with X-chromosomes from *M. musculus.*

In the two *Helianthus* hybrid zones, approximately one-half of the markers examined showed patterns of introgression consistent with neutral expectations based on genome-wide admixture. These results suggest that much of the genome may be experiencing selection in these admixed populations. Markers deviated from neutral expectations in a variety of ways, providing evidence for markers associated with reproductive isolation and adaptive introgression. The large number of loci that exhibited steeper genomic clines than expected under neutrality may be associated with incompatibilities in hybrids that are based on many genes or chromosomal rearrangements (Rieseberg *et al*. 1999b). A similarly large proportion of significant departures from neutrality was detected in a European *Populus* hybrid zone using a related method (see below, Lexer *et al*. 2007).

Our analysis of data from two *Helianthus* hybrid zones demonstrates the ability of our method to assess concordance in patterns of introgression between different hybrid zones or different classes of individuals. In the California and Nebraska sunflower hybrid zones, most loci show similar patterns of introgression, consistent with the finding of the earlier study, of high concordance between hybrid zones (Buerkle & Rieseberg 2001). However, the current analytical method revealed several loci that showed remarkably different patterns of introgression. These results indicate that, for most of the genome, there is little evidence of intraspecific variation in the genetic architecture of reproductive isolation between *H. annuus* and *H. petiolaris.* But some loci exhibit patterns of introgression that differ between the hybrid zones and may be subject to varying selection in different settings (environments and genetic backgrounds). Additional studies that test for polymorphism among multiple hybrid zones in isolating barriers will be informative about the dynamics and population genetics of speciation genes (Buerkle & Rieseberg 2001; Aldridge 2005) and our method for estimation of genomic clines is a promising approach for this purpose. Similarly, comparisons among different classes of individuals (age groups, sexes, etc.) in the same hybrid zone will reveal the genetic basis of their contributions to isolating barriers.

This method should be applicable to a variety of organisms that experience natural admixture and are current subjects of study in ecology and evolution (e.g. *Bombina*, Szymura & Barton 1991; Yanchukov *et al*. 2006; *Coregonus*, Rogers *et al*. 2001; *Cottus*, Nolte *et al*. 2005; *Gasterosteus*, Gow *et al*. 2006; Taylor *et al*. 2006; *Helianthus*, Rieseberg *et al*. 1999b; *Heliconius*, Mallet *et al*. 1990, 2007; *Mus*, Payseur *et al*. 2004; Teeter *et al*. 2008; *Populus*, Martinsen *et al*. 2001; Lexer *et al*. 2007; *Silene*, Minder *et al*. 2007). In fact, an earlier version of this method was successfully used for a coarse-scale investigation of the genetic architecture of reproductive isolation between *Populus alba* and *P. tremula* (Lexer *et al*. 2007). Lexer *et al*. (2007) found several loci that deviated from patterns of neutral introgression, including loci with genomic clines consistent with underdominance and directional selection. The application of this method and others to a diversity of hybridizing organisms with varying degrees of phylogenetic relatedness should facilitate a greater understanding of the genetic architecture of reproductive isolation, and by extension, of the speciation process.

As this method is regression-based, it is easy to incorporate additional predictors of the genotype at a locus in the model. For example, spatial or environmental variables could be added and evaluated to determine whether they improve the fit of the model. This procedure could be used to determine whether loci that deviate from patterns of neutral introgression are influenced by exogenous selection associated with identifiable aspects of the environment. This information is difficult to ascertain under traditional methods that examine genotype frequencies as a function of distance along a spatial transect (Kruuk *et al*. 1999). Phenotypic markers could also be added to the model (e.g. sex, coloration, size, etc.), which may facilitate the detection of the effects of allelic variation at individual loci in specific genetic backgrounds. Similarly, epistatic interactions may be investigated by conditioning the genotypic prediction both on the genome-wide admixture and on the genotype at a potentially interacting locus.

Despite the clear utility of this method, it is not without limitations. An accurate estimate of genome-wide admixture is necessary, requiring individuals to be assayed for a moderately large number of molecular markers. If estimates of genome-wide admixture are based on an insufficient number of molecular markers, the assumption that the estimate of genome-wide admixture is representative of neutral introgression may not hold. This is particularly true if the markers used are not widely distributed across the genome (as in the *Mus* example above). Second, this method requires accurate estimates of parental allele frequencies, which necessitates the identification of appropriate parental populations. Parental populations should be identified using detailed population genetic information and with the aid of Bayesian assignment methods (e.g. Pritchard *et al*. 2000; Falush *et al*. 2003). Finally, without a reliable linkage map, some questions will be difficult to answer, as it will be unclear to what extent loci demonstrating similar patterns of introgression are independent. Thus, without a linkage map, it will not be possible to determine the proportion of the genome experiencing different forms of selection.

Our analytical approach does not include estimates of the intensity of selection at loci that deviate from the neutral model. To do so would require a number of assumptions about the history of admixture and the nature of selection (as do methods for geographic clines, Barton & Hewitt 1985; Barton & Gale 1993; Porter *et al*. 1997) that would be difficult to prefer over alternatives. Instead, our approach identifies significant departures from neutrality and provides estimates of the lack of fit. The significant loci can be ordered on the basis of the magnitude of the departures from neutrality and on the strength of the evidence from flanking markers, to yield a list of candidate regions for additional study. This method should prove useful in the future use of hybrid zones as a context in which to dissect the genetics of isolation between divergent lineages.

Acknowledgements

Marcus Brock, Christian Lexer, Lauren Lucas, Arne Nolte, Tom Parchman, Cynthia Weinig, Adam Porter, and two anonymous reviewers provided helpful comments on earlier versions of this manuscript. This work was funded by NSF DBI award 0701757 to C.A.B. and an NSF Predoctoral Graduate Research Fellowship to Z.G.

References

Aldridge G (2005) Variation in frequency of hybrids and spatial structure among *Ipomopsis* (Polemoniaceae) contact sites. *New Phytologist*, **167**, 279–288.

Barton NH, Gale KS (1993) Genetic analysis of hybrid zones. In: *Hybrid Zones and the Evolutionary Process* (ed. Harrison RG), pp. 13–45. Oxford University Press, New York.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Beaumont M (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B – Methodological*, **57**, 289–300.

Bradshaw HD, Wilbert SM, Otto KG, Schemske DW (1995) Genetic-mapping of floral traits associated with reproductive isolation in monkeyflowers (*Mimulus*). *Nature*, **376**, 762–765.

Buerkle CA (2005) Maximum-likelihood estimation of a hybrid index based on molecular markers. *Molecular Ecology Notes*, **5**, 684–687.

Buerkle CA, Lexer C (2008) Admixture as the basis for genetic mapping. *Trends in Ecology & Evolution*, **23**, 686–694.

Buerkle CA, Rieseberg LH (2001) Low intraspecific variation for genomic isolation between hybridizing sunflower species. *Evolution*, **55**, 684–691.

Buerkle CA, Rieseberg LH (2008) The rate of genome stabilization in homoploid hybrid species. *Evolution*, **62**, 266–275.

Burch CL, Chao L (1999) Evolution by small steps and rugged landscapes in the RNA virus ϕ6. *Genetics*, **151**, 921–927.

Coyne JA, Orr HA (1998) The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **353**, 287–305.

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

Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press, Princeton, New Jersey.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fisher RA (1954) A fuller theory of 'junctions' in inbreeding. *Heredity*, **8**, 187–197.

Fitzpatrick BM (2004) Rates of evolution of hybrid inviability in birds and mammals. *Evolution*, **58**, 1865–1870.

Fitzpatrick BM, Shaffer HB (2007) Hybrid vigor between native and introduced salamanders raises new challenges for conservation. *Proceedings of the National Academy of Sciences, USA*, **104**, 15793–15798.

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

Gow JL, Peichel CL, Taylor EB (2006) Contrasting hybridization rates between sympatric three-spined sticklebacks highlight the fragility of reproductive barriers between evolutionarily young species. *Molecular Ecology*, **15**, 739–752.

Gregorius HR, Roberds JH (1986) Measurement of genetical differentiation among subpopulations. *Theoretical and Applied Genetics*, **71**, 826–834.

Harrison RG (1990) Hybrid zones: windows on evolutionary process. *Oxford Surveys in Evolutionary Biology*, **7**, 69–128.

Hewitt GM (1988) Hybrid zones—natural laboratories for evolution studies. *Trends in Ecology & Evolution*, **3**, 158–166.

Kruuk LEB, Baird SJE, Gale KS, Barton NH (1999) A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, **153**, 1959–1971.

Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2000 generations. *American Naturalist*, **138**, 1315–1341.

Lexer C, Buerkle CA, Joseph JA, Heinze B, Fay MF (2007) Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Heredity*, **98**, 74–84.

Liu J, Mercer JM, Stam LF, Gibson GC, Zeng ZB, Laurie CC (1996) Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana*. *Genetics*, **142**, 1129–1145.

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

Mackay TFC (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics*, **35**, 303–339.

Mallet J, Barton N, Lamas G, Santisteban J, Muedas M, Eeley H (1990) Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. *Genetics*, **124**, 921–936.

Mallet J, Beltran M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, **7**, 28.

Martin NH, Bouck AC, Arnold ML (2006) Detecting adaptive trait introgression between *Iris fulva* and *I. brevicaulis* in highly selective field conditions. *Genetics*, **172**, 2481–2489.

Martin NH, Sapir Y, Arnold ML (2008) The genetic architecture of reproductive isolation in Louisiana irises: pollination syndromes and pollinator preferences. *Evolution*, **62**, 740–752.

Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter introgression between species. *Evolution*, **55**, 1325–1335.

Minder AM, Rothenbuehler C, Widmer A (2007) Genetic structure of hybrid zones between *Silene latifolia* and *Silene dioica* (Caryophyllaceae): evidence for introgressive hybridization. *Molecular Ecology*, **16**, 2504–2516.

Minder AM, Widmer A (2008) A population genomic analysis of species boundaries: neutral processes, adaptive divergence and introgression between two hybridizing plant species. *Molecular Ecology*, **17**, 1552–1563.

Nolte AW, Freyhof J, Stemshorn KC, Tautz D (2005) An invasive lineage of sculpins, *Cottus* sp. (Pisces, Teleostei) in the Rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 2379–2387.

Payseur BA, Krenz JG, Nachman MW (2004) Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. *Evolution*, **58**, 2064–2078.

Porter AH, Wenger R, Geiger H, Scholl A, Shapiro AM (1997) The *Pontia daplidice-edusa* hybrid zone in northwestern Italy. *Evolution*, **51**, 1561–1573.

Presgraves DC (2002) Patterns of postzygotic isolation in Lepidoptera. *Evolution*, **56**, 1168–1183.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rieseberg LH (2000) Crossing relationships among ancient and experimental sunflower hybrid lineages. *Evolution*, **54**, 859–865.

Rieseberg LH, Archer MA, Wayne RK (1999a) Transgressive segregation, adaptation, and speciation. *Heredity*, **83**, 363–372.

Rieseberg LH, Baird S, Desrochers A (1998) Patterns of mating in wild sunflower hybrid zones. *Evolution*, **52**, 713–726.

Rieseberg LH, Buerkle CA (2002) Genetic mapping in hybrid zones. *The American Naturalist*, **159**, S36–S50.

Rieseberg LH, Wendel JF (1993) Introgression and its consequences in plants. In: *Hybrid Zones and the Evolutionary Process* (ed. Harrison RG), pp. 70–109. Oxford University Press, New York.

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

Rogers SM, Campbell D, Baird S, Danzmann RG, Bernatchez L (2001) Combining the analyses of introgressive hybridisation and linkage mapping to investigate the genetic architecture of population divergence in the lake whitefish (*Coregonus clupeaformis,* Mitchill). *Genetica*, **111**, 25–41.

Simon M, Loudet O, Durand S *et al*. (2008) Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers. *Genetics*, **178**, 2253–2264.

Szymura JM, Barton NH (1991) The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata:* comparisons between transects and between loci. *Evolution*, **45**, 237–261.

Tan Q, Christiansen L, Bathum L *et al*. (2005) Haplotype effects on human survival: logistic regression models applied to unphased genotype data. *Annals of Human Genetics*, **69**, 168–175.

Taylor EB, Boughman JW, Groenenboom M, Sniatynski M, Schluter D, Gow JL (2006) Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (*Gasterosteus aculeatus*) species pair. *Molecular Ecology*, **15**, 343–355.

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

Vallender R, Robertson RJ, Friesen VL, Lovette IJ (2007) Complex hybridization dynamics between golden-winged and blue-winged warblers (*Vermivora chrysoptera* and *Vermivora pinus*) revealed by AFLP, microsatellite, intron and mtDNA markers. *Molecular Ecology*, **16**, 2017–2029.

Venables WN, Ripley B (2002) *Modern Applied Statistics with S*, 4th edn. Springer Verlag, New York.

Vines TH, Kohler SC, Thiel A *et al*. (2003) The maintenance of reproductive isolation in a mosaic hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata. Evolution*, **57**, 1876–1888.

Whitney KD, Randell RA, Rieseberg LH (2006) Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus. The American Naturalist*, **167**, 794–807.

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Yanchukov A, Hofman S, Szymura JM, Mezhzherin SV (2006) Hybridization of *Bombina bombina* and *B. variegata* (Anura, Discoglossidae) at a sharp ecotone in western Ukraine: comparisons across transects and over time. *Evolution*, **60**, 583–600.

Zhu X, Luke A, Cooper RS *et al*. (2005) Admixture mapping for hypertension loci with genome-scan markers. *Nature Genetics*, **37**, 177–181.

The authors are broadly interested in hybridization, both as an opportunity to dissect components of reproductive isolation and for its creative role in homoploid hybrid speciation. Zach Gompert is a PhD. student and does empirical research on genomic clines and hybridization in *Lycaeides* butterflies. He also develops analytical approaches in collaboration with his adviser, Alex Buerkle, who is an assistant professor and evolutionary geneticist at the University of Wyoming.