PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Genomics of isolation in hybrids

Zachariah Gompert, Thomas L. Parchman and C. Alex Buerkle

Phil. Trans. R. Soc. B 2012 **367**, 439-450
doi: 10.1098/rstb.2011.0196

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 439–450
doi:10.1098/rstb.2011.0196

Research

Genomics of isolation in hybrids

Zachariah Gompert*, Thomas L. Parchman and C. Alex Buerkle

Department of Botany and Program in Ecology, University of Wyoming, Laramie, WY 82071, USA

Hybrid zones are common in nature and can offer critical insights into the dynamics and components of reproductive isolation. Hybrids between diverged lineages are particularly informative about the genetic architecture of reproductive isolation, because introgression in an admixed population is a direct measure of isolation. In this paper, we combine simulations and a new statistical model to determine the extent to which different genetic architectures of isolation leave different signatures on genome-level patterns of introgression. We found that reproductive isolation caused by one or several loci of large effect caused greater heterogeneity in patterns of introgression than architectures involving many loci with small fitness effects, particularly when isolating factors were closely linked. The same conditions that led to heterogeneous introgression often resulted in a reasonable correspondence between outlier loci and the genetic loci that contributed to isolation. However, demographic conditions affected both of these results, highlighting potential limitations to the study of the speciation genomics. Further progress in understanding the genomics of speciation will require large-scale empirical studies of introgression in hybrid zones and model-based analyses, as well as more comprehensive modelling of the expected levels of isolation with different demographies and genetic architectures of isolation.

Keywords: hybrid zone; admixture; introgression; reproductive isolation; Bayesian inference

1. INTRODUCTION

Speciation is a fundamental evolutionary process that occurs by the evolution of reproductive isolation. There has been considerable recent progress documenting the genetics of reproductive isolation, and individual genes that contribute to isolation have been identified in several groups of organisms [1–3]. However, little is known regarding the genetics of speciation at the scale of a genome, including knowledge of the number and effect of loci that contribute to isolation, how these loci are distributed across the genome and how they contribute to genome-wide patterns of genetic variation. The study of hybrid zones is a powerful approach to address these questions.

Hybrid zones are common in nature, and offer direct observations of the evolutionary process of speciation and the genetic architecture of reproductive isolation [4–8]. When species diverge in allopatry and hybridize upon secondary contact, the products of meiosis and segregation in admixed individuals produce combinations of parental genotypes (or chromosomal blocks) that are tested by natural selection [9,10]. Hybrid zones offer tractable and important settings for the study of reproductive isolation because the introgression of foreign alleles is a direct measure of reproductive isolation. In contrast, population and species divergence often reflect the consequences of additional evolutionary processes, beyond those directly associated with reproductive isolation and speciation [11].

Introgression of loci will vary with their effect on fitness, and might take distinct forms reflecting the source and pattern of selection [9,12]. This includes variation in fitness caused by intrinsic hybrid incompatibilities or extrinsic selection. The geographical extent of introgression of loci (and linked genetic regions) that contribute to reproductive isolation should be reduced relative to the rest of the genome [4,9,13,14]. Conversely, introgression of loci that do not affect hybrid fitness (i.e. neutral loci) is affected by linkage disequilibrium, but should typically be more frequent and geographically extensive [9,14].

Introgression in hybrid zones has been quantified using geographical and genomic clines. Geographical clines describe the relationship between allele frequency or expected phenotype and geographical location [4,9,15–17]. Geographical clines are used to estimate the strength of the barrier to gene flow experienced by a neutral locus because of reduced hybrid fitness or other forms of selection, and allow inference of the number of genetic regions contributing to reproductive isolation [9,15,18]. Moreover, tests of geographical concordance and coincidence among genetic or phenotypic clines can detect variable introgression and have been used to identify loci potentially associated with reproductive isolation [19–21]. *Genomic clines* are mathematical functions that describe the probability of locus-specific ancestry along a gradient in genome-wide admixture or hybrid index, which is defined as the proportion of an admixed individual's genome inherited from one of two parental populations [12,22,23], with related models in earlier studies [15,24–26].

* Author for correspondence (zgompert@uwyo.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2011.0196 or via http://rstb.royalsocietypublishing.org

One contribution of 13 to a Theme Issue 'Patterns and processes of genomic divergence during speciation'.

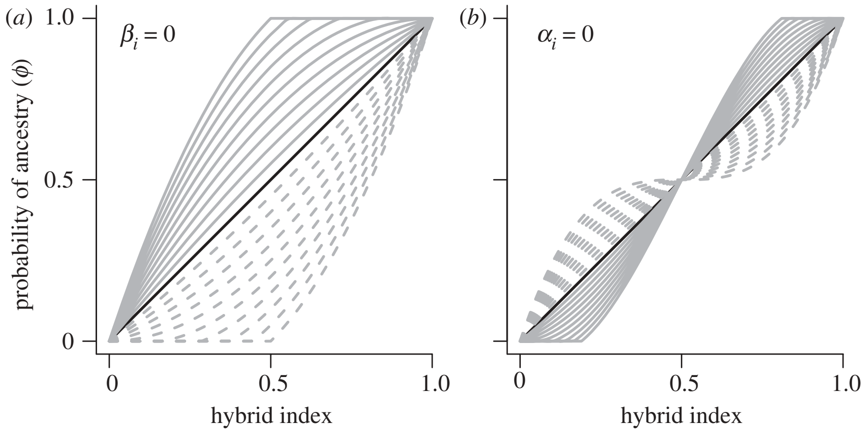

Figure 1. Plots depict hypothetical genomic clines. In (*a*) and (*b*), the solid black line depicts a genomic cline when $\alpha = \beta = 0$, which is the genome-wide mean expectation with hybrid index equal to the probability of ancestry from parental population 1. In (*a*) β_i is set to 0 and α_i is varied from 0.1 to 1 (solid grey lines) and from −0.1 to −1 (dashed grey lines). Similarly, in (*b*) α_i is set to 0 and β_i is varied from 0.1 to 1 (solid grey lines) and from −0.1 to − 1 (dashed grey lines).

This is not a spatial gradient, rather nearly pure individuals of each parental type occupy opposite ends of the gradient. Thus, whereas geographical clines measure the movement of alleles across space, genomic clines measure the movement of ancestry blocks into different genomic backgrounds. Genomic cline models provide a way to generate a null-distribution for patterns of introgression that is conditioned on genome-wide admixture using a simple parametric model, a permutation method, or a hierarchical Bayesian framework [12,23]. This model framework can be used to identify outlier loci that are characterized by extreme introgression, and might be associated with adaptation or reproductive isolation.

Empirical studies of hybrid zones using these methods have repeatedly provided evidence for variable introgression among loci [18,21,24,27–29]. The observed variation among loci supports the hypothesis that reproductive isolation is an attribute of individual genomic regions rather than of the genome as a whole [30]. For example, sex chromosomes have been found to introgress relatively infrequently, consistent with theoretical expectations that such sex-linked loci should be important in speciation [18,19,31]. Both concordant and discordant patterns of locus-specific introgression in geographically separate hybrid zones have been reported [27,29,32]. Such comparisons allow analysis of how geographical, population and environmental variation contribute to the genetic architecture of isolation. Ultimately, comparisons of isolation in different settings and hybrid zones will enhance our understanding of the dynamics of species isolation and divergence, and the efficacy and polymorphism of isolating barriers [29,33,34].

Despite considerable progress in understanding the genetics of speciation, knowledge of the genome-level consequences and genome-wide signature of reproductive isolation is still in its infancy. In this paper, we begin to fill this gap by determining whether different genetic architectures of reproductive isolation lead to different and detectable patterns of introgression at a genome scale. To address this question, we first review and extend the recently developed Bayesian genomic cline model (described in §2*a*; [23]) to account for genomic autocorrelations in patterns of introgression for linked genetic regions. We then simulate admixture under different demographic conditions and a variety of genetic architectures of isolation, including reduced hybrid fitness owing to underdominance or epistasic incompatibilities, variation in the number and effect sizes of genes contributing to reproductive isolation, and variation in the genomic distribution of genetic factors affecting fitness. We analyse these simulations in the first place to determine the genetic outcomes of different models and summarize patterns of locus-specific introgression using the revised Bayesian genomic cline model. Finally, we discuss the implications for future genomic studies of the genetic architecture of isolation.

2. MODEL

(a) *Bayesian genomic cline model*

The purpose of the genomic cline model is to quantify variable introgression of loci. With this model, we are specifically interested in the movement of parental ancestry segments into different hybrid backgrounds (measured by hybrid index) within an admixed population. Genomic clines are mathematical functions that quantify introgression of individual loci relative to a gradient in hybrid index. In other words, a genomic cline describes the probability of locus-specific ancestry given genome-wide ancestry or hybrid index. Gompert & Buerkle [23] recently described a Bayesian model for estimating genomic clines. This model contains two key genomic cline parameters, α and β, that are used to quantify variable patterns of introgression (figure 1). Additionally, these parameters form the basis for detecting outlier loci that might be associated with reproductive isolation. Genomic cline parameter α specifies an increase ($\alpha > 0$) or decrease ($\alpha < 0$) in the probability of locus-specific ancestry from parental population 1 and defines the centre of the cline (we assume two parental populations labelled as parental population 0 and 1). Genomic cline parameter β denotes the rate of change in the probability of ancestry along the genome-wide admixture gradient. β is the rate parameter and positive values of β specify a steeper cline, whereas negative values of β specify a wider cline (figure 1). Herein, we review the Bayesian genomic cline model before proposing an extension of this model for high-resolution, genome-wide genetic data. Readers interested in a more thorough description of this model should refer to Gompert & Buerkle [23].

The Bayesian genomic cline model assumes two parental populations are known *a priori*, and uses genetic data from these populations to estimate parental population allele frequencies. The model assumes Hardy–Weinberg and linkage equilibrium within parental populations (as in earlier studies [23,35] and similar models), but does not assume these populations are fixed for different alleles. Estimates of parental allele frequencies are necessary to estimate the ancestry for individuals from putatively admixed populations. Given the observed allele counts ψ_0 and an uninformative Dirichlet prior, the posterior probability distribution for allele frequencies in parental population 0 (π_0) is:

$$P(\pi_0|\psi_0) \sim \prod_i \text{Dirichlet}(\psi_{i01}+1, \psi_{i02} + 1, \ldots, \psi_{i0k}+1),$$

where the product is taken across all I loci, and each Dirichlet distribution is parametrized by a vector of length K_i (K_i is the number of unique alleles observed for locus i). The posterior probability distribution for allele frequencies in parental population 1 is equivalently specified.

The genomic cline model assumes putatively admixed individuals are sampled randomly from each of J admixed populations, but no assumption is made about the geographical distribution of these populations. We model the ancestry of each allele copy in admixed individuals. Ancestry (**z**) simply denotes whether an allele was inherited from parental population 0 ($z = 0$) or parental population 1 ($z = 1$). The likelihood of the genetic data from the admixed populations (**x**) is conditional on ancestry and parental population allele frequencies:

$$P(\mathbf{x}|\pi,\mathbf{z}) = \prod_i \prod_j \prod_n \prod_a \begin{cases} \prod_k \pi_{i0k}^{x_{ijnak}} & \text{if } z_{ijna} = 0 \\ \prod_k \pi_{i1k}^{x_{ijnak}} & \text{if } z_{ijna} = 1. \end{cases} \tag{2.1}$$

Here, x_{ijnak} is one if individual n has allelic state k for allele copy a at locus i (and diploid individuals have two allele copies at each locus), and x_{ijnak} is zero otherwise. When multiple admixed populations are included, j denotes the population from which individual n was sampled. The likelihood is a product across all loci, admixed populations (if more than one admixed population is included), individuals, and allele copies.

A genomic cline function specifies the prior probability of locus-specific ancestry from parental population 1, which is conditional on hybrid index and the locus-specific genomic cline parameters. This prior probability is denoted as ϕ. This probability is a simple function of an auxiliary variable θ, but has the imposed biologically meaningful constraints that ϕ must be bounded by zero and one and must be a monotonically increasing function of hybrid index. The mathematical details of the transformation between ϕ and θ can be found in Gompert & Buerkle [23]; here, we simply describe the polynomial function for the auxiliary variable θ. This function includes a term for hybrid index (**h**) and the genomic cline parameters α and β that were described previously (figure 1):

$$\theta_{ijn} = h_n + 2(h_n - h_n^2)(\alpha_{ij} + \beta_{ij}(2h_n - 1)). \tag{2.2}$$

Hybrid index is polarized such that an individual with $h_n = 0$ has ancestry only from parental population 0, and an individual with $h_n = 1$ has ancestry only from parental population 1.

The genomic cline parameters are allowed to vary by locus (i) and admixed population (j; if multiple admixed populations are analysed). Therefore, linear random-effect models are specified for α_{ij} and β_{ij}:

$$\alpha_{ij} = \gamma_i + \eta_{i(j)} \tag{2.3}$$

and

$$\beta_{ij} = \zeta_i + \kappa_{i(j)}. \tag{2.4}$$

Gompert & Buerkle [23] assumed hierarchical normal priors for the locus and nested population effects for the cline parameters: $\gamma \sim \text{N}(0,\tau_\alpha)$, $\zeta \sim \text{N}(0,\tau_\beta)$, $\eta_i \sim \text{N}(0,\nu_i)$ and $\kappa_i \sim \text{N}(0,\omega_i)$. Each of these hierarchical priors has a mean of zero. The precision parameters for these prior distributions describe the magnitude of variation in patterns of introgression among loci. When analysis is restricted to a single admixed population the nested population effects are dropped from the model, such that $\alpha_i = \gamma_i$ and $\beta_i = \zeta_i$. To complete the Bayesian genomic cline model, uninformative priors are specified for the random-effect precision parameters (τ_α, τ_β, ν and ω) and hybrid index (**h**).

(b) *Intrinsic conditional autoregressive ρ prior for linkage*

The choice of hierarchical normal priors for γ, ζ, η and κ assumes that each γ_i, ζ_i, $\eta_{i(j)}$ and $\kappa_{i(j)}$ is an independent sample from a genome-wide distribution (i.e. a normal distribution with mean zero and an estimated precision parameter). In other words, because these priors specify conditional independence for cline parameter random effects, values for γ_i and γ_i' are assumed to be independent given the precision parameter τ_α. Linkage creates spatial genomic autocorrelation in the evolutionary history of loci [36]. Thus, the assumption of conditional independence is likely to be violated when high-resolution genetic data are used to study genome-wide divergence during speciation, and modelling autocorrelation will be highly relevant for detecting and interpreting outlier loci.

Therefore, we now model autocorrelation owing to linkage by specifying intrinsic conditional autoregressive ρ (ICARρ) priors for the genomic cline parameters. This form of statistical model is commonly used to account for spatial autocorrelation in geographical models [37,38] and is quite similar to the approach proposed by Guo *et al.* [39] to account for correlations owing to linkage when conducting F_{ST}-outlier genome scans. For the purpose of this paper, we describe

ICARρ priors for the locus-specific random effects γ and ζ, but not the nested population effects. The extension of this prior to the nested population effects would be trivial. Accurate estimates of recombination rates between loci are required to implement the ICARρ model. The conditional forms of the ICARρ priors for γ_i and ζ_i are

$$P(\gamma_i|\gamma_{[i]}) \sim N\left(\frac{\rho\sum_{i'\neq i} w_{ii'}\gamma_{i'}}{w_{i+}}, \tau_\alpha w_{i+}\right) \qquad (2.5)$$

and

$$P(\zeta_i|\zeta_{[i]}) \sim N\left(\frac{\rho\sum_{i'\neq i} w_{ii'}\zeta_{i'}}{w_{i+}}, \tau_\beta w_{i+}\right). \qquad (2.6)$$

In these equations $\gamma_{[i]}$ is the collection of γ_i, for all $i^{'} \neq i$, $\zeta_{[i]}$ is the collection of ζ_i, for all $i^{'} \neq i$, $w_{ii}{'}$ is an entry from an $I \times I$ weight matrix (**W**), $w_{i+} = \sum_{i'\neq i} w_{ii'}$ and ρ is a spatial dependence parameter that is included in part to ensure propriety of the posterior distribution [37,38]. The weight matrix (**W**) defines the expected correlations among genetic regions owing to linkage (genomic proximity). We impose a sum-to-zero constrain on γ and ζ to ensure identifiability of the parameters. Additional details of the ICARρ model are provided in the electronic supplementary material (see §S1, ICARρ model details).

(c) *Designating outlier loci*

The Bayesian genomic cline model provides a framework to designate statistical outlier loci, which we define as loci with unlikely or extreme cline parameters given the appropriate hierarchical prior [23]. Outlier loci have aberrant patterns of introgression and might be associated with reproductive isolation [12,23]. We designate outlier loci with respect to genomic cline centre (α) or genomic cline rate (β). Following Guo *et al.* [39], these can be local (i.e. relative to nearby genetic regions) or global (i.e. relative to the entire genome) outlier loci. Locus i is a local outlier with respect to α if the posterior point estimate of α_i is not contained in the interval q_N, which is defined as the interval bounded by the N/2 and (1 − N)/2 quantiles of the ICARρ prior for α. Local outliers with respect to β are designated likewise. Designating global outliers requires a different reference distribution. Because we impose sum-to-zero constraints on the cline parameter random effects and because the mean genome-wide cline parameter should always be zero, we designate global outliers with respect to zero-centred distributions. Specifically, locus i is a global outlier with respect to α if the posterior estimate of α is not contained in the interval q_N^*, where q_N^* is the interval bounded by the N/2 and (1 − N)/2 quantiles of $N(0,\tau_\alpha\, w_{i+})$ (likewise for β).

3. METHODS

(a) *Simulations*

We simulated genetic data for admixed populations to determine whether different genetic architectures of reproductive isolation left distinct genomic signatures. We have used a related model for simulating admixed populations in other studies [12,23,40]. We were interested in three major contrasts regarding the genetic architecture of reproductive isolation: (i) the number and fitness effect of genetic loci contributing to isolation, (ii) whether isolation was the result of underdominance or epistatic incompatibilities, and (iii) the genomic location and distribution of loci associated with isolation.

We simulated underdominant selection and selection arising from pairwise epistatic interactions. For both forms of selection, we assumed that fitness was multiplicative and was the result of an individual's ancestry at N_s loci, and that each locus had an equal effect on fitness. Underdominant selection arises when the fitness of the heterozygous genotype is lower than that of either homozgyous genotype. We modelled underdominance by defining the relative fitness of an individual as $f = 1 - (1 - s)^{x_s}$, where s is the reduction in fitness associated with having heterozygous ancestry at a single locus with an effect on fitness. x_s is the number of loci (out of the N_s loci) at which an individual had heterozygous ancestry. Epistatic interactions are fundamental for reproductive isolation owing to the accumulation of Bateson–Dobzhansky–Muller (BDM) incompatibilities [41–43], and are thought to play a significant role in the genetics of speciation [42,44,45]. We assumed a simple model of BDM incompatibilities, where interactions are between pairs of loci and the ancestral state at each locus is fitter [44,46]. For each pair of interacting loci, we assumed that parental population 0 is fixed for a derived allele at the first locus (A) and parental population 1 is fixed for a derived allele at the second locus (B; the ancestral genotype is *aabb*). Combinations of derived alleles at the interacting locus pair are incompatible and cause reduced fitness in hybrids. There are three different potential incompatibilities: homozygous derived × homozygous derived (H_2), homozygous derived × heterozygous (H_1) and heterozygous × heterozygous (H_0; [44]). The fitness effect of each incompatibility depends on dominance. To achieve the greatest effect possible with BDM incompatibilities, we assumed complete dominance of derived alleles $s = s_{H_2} = s_{H_1} = s_{H_0}$, such that an individual's relative fitness was given by $f = 1 - (1 - s)^{x_{H_2}+x_{H_1}+x_{H_0}}$ (where x_{H_2} refers to the number of pairs of H_2 interactions, etc.). The simulations modelled admixture between parental populations with fixed allele differences and no demic structure within the admixed population (these are not assumptions of the genomic cline model). We simulate a variety of demographic conditions (table 1; electronic supplementary material, S2 'Simulation details').

(b) *Analyses*

To assess the genome-level signature of reproductive isolation, we first summarized hybrid index and interspecific heterozygosity (i.e. the proportion of the genome with alleles at a locus inherited from different parental populations) across all individuals for each simulated dataset, directly from the simulation output. We also quantified mean locus-specific ancestry (i.e. the proportion of allele copies inherited from parental population 1) and locus-specific interspecific heterozygosity for each combination of simulation conditions. These parameters measure the effect of different genetic architectures of isolation directly from known simulation output, independent of the genomic clines model.

Table 1. Genome-level summary of simulated datasets and results. UD, underdominance; EP, BDM epistatic incompatibility; $S_{\max}$, reduction in fitness for the least fit genotype; m, per generation rate of dispersal from parental populations; g, number of generations simulated; $\bar{h}_{\mathrm{dev}}$, mean and standard deviation of the absolute deviation of h from 0.5; $\bar{h}_{et}$, mean and standard deviation for interspecific heterozygosity; s.d.(α), mean standard deviation of the α_i cline parameters across replicates; s.d.(β), mean standard deviation of the β_i cline parameters across replicates; $\mathrm{I}_{(0\mathrm{cM},2\mathrm{cM}]}(\alpha)$ and $\mathrm{I}_{(0\mathrm{cM},2\mathrm{cM}]}(\beta)$, mean value of Moran's I at lag 2cM across replicates; $I_{1/2}(\alpha)$ and $I_{1/2}(\beta)$, mean value of lag (l) for which $\mathrm{I}_{(l-0.2\mathrm{cM},l\mathrm{cM}]} \leq (I_{(0\mathrm{cM},2\mathrm{cM}]}/2)$; other variables are defined in the main text.

type	s	N_s	N_{ch}	$S_{\max}$	m	g	$\bar{h}_{\mathrm{dev}}$	$\bar{h}_{et}$	s.d.(α)	s.d.(β)	$I_{(0\mathrm{cM},2\mathrm{cM}]}(\alpha)$	$I_{(0\mathrm{cM},2\mathrm{cM}]}(\beta)$	$I_{1/2}(\alpha)$	$I_{1/2}(\beta)$
UD	0.01	5	1	0.05	0.05	10	0.095 (0.082)	0.496 (0.081)	0.037 (0.004)	0.025 (0.009)	1.026 (0.145)	0.936 (0.101)	0.162 (0.052)	0.208 (0.033)
UD	0.01	110	11	0.67	0.05	10	0.126 (0.091)	0.480 (0.096)	0.039 (0.008)	0.035 (0.017)	0.999 (0.079)	0.977 (0.127)	0.284 (0.103)	0.220 (0.087)
UD	0.05	20	2	0.64	0.05	10	0.114 (0.088)	0.487 (0.093)	0.041 (0.010)	0.030 (0.013)	0.982 (0.121)	1.009 (0.084)	0.272 (0.118)	0.224 (0.045)
UD	0.20	5	1	0.67	0.05	10	0.132 (0.096)	0.472 (0.099)	0.042 (0.012)	0.045 (0.024)	1.044 (0.074)	1.005 (0.099)	0.212 (0.091)	0.270 (0.073)
UD	0.65	1	1	0.65	0.05	10	0.141 (0.094)	0.467 (0.096)	0.056 (0.027)	0.058 (0.022)	1.004 (0.136)	0.927 (0.091)	0.244 (0.107)	0.344 (0.079)
UD	0.01	110	11	0.67	0.20	10	0.226 (0.144)	0.434 (0.223)	0.022 (0.006)	0.020 (0.005)	0.955 (0.125)	0.998 (0.081)	0.284 (0.118)	0.260 (0.092)
UD	0.20	5	1	0.67	0.20	10	0.228 (0.145)	0.433 (0.224)	0.024 (0.002)	0.032 (0.023)	0.996 (0.124)	1.009 (0.108)	0.288 (0.115)	0.284 (0.088)
UD	0.01	110	11	0.67	0.05	25	0.110 (0.086)	0.483 (0.079)	0.067 (0.007)	0.021 (0.009)	0.966 (0.137)	1.034 (0.138)	0.184 (0.117)	0.152 (0.045)
UD	0.20	5	1	0.67	0.05	25	0.195 (0.096)	0.418 (0.100)	0.097 (0.022)	0.058 (0.043)	0.843 (0.078)	0.898 (0.116)	0.184 (0.042)	0.168 (0.065)
UD	0.01	110	22	0.67	0.05	10	0.140 (0.097)	0.468 (0.100)	0.037 (0.006)	0.039 (0.024)	1.063 (0.131)	0.991 (0.134)	0.198 (0.067)	0.250 (0.106)
UD	0.20	5	5	0.67	0.05	10	0.150 (0.098)	0.457 (0.101)	0.061 (0.038)	0.144 (0.077)	0.995 (0.121)	0.907 (0.082)	0.328 (0.136)	0.482 (0.041)
EP	0.05	10	2	0.23	0.05	10	0.096 (0.083)	0.494 (0.079)	0.041 (0.012)	0.023 (0.011)	0.979 (0.114)	1.014 (0.070)	0.190 (0.044)	0.182 (0.040)
EP	0.05	20	2	0.40	0.05	10	0.104 (0.086)	0.492 (0.087)	0.049 (0.015)	0.025 (0.012)	1.042 (0.077)	0.925 (0.094)	0.196 (0.045)	0.222 (0.052)
EP	0.20	10	2	0.67	0.05	10	0.121 (0.092)	0.477 (0.092)	0.093 (0.014)	0.032 (0.024)	1.026 (0.085)	0.963 (0.104)	0.208 (0.018)	0.210 (0.066)
EP	0.20	20	2	0.89	0.05	10	0.339 (0.140)	0.258 (0.202)	0.081 (0.032)	0.078 (0.051)	1.083 (0.087)	1.025 (0.125)	0.206 (0.037)	0.222 (0.060)
EP	0.65	2	2	0.65	0.05	10	0.123 (0.087)	0.475 (0.088)	0.092 (0.013)	0.043 (0.035)	1.036 (0.162)	1.014 (0.118)	0.194 (0.037)	0.222 (0.046)
EP	0.20	10	1	0.67	0.05	10	0.122 (0.093)	0.477 (0.097)	0.099 (0.014)	0.034 (0.009)	0.910 (0.061)	0.981 (0.135)	0.474 (0.051)	0.242 (0.109)
EP	0.65	2	1	0.65	0.05	10	0.107 (0.084)	0.482 (0.081)	0.117 (0.009)	0.033 (0.167)	0.885 (0.060)	1.011 (0.126)	0.488 (0.020)	0.226 (0.087)

In addition to direct quantification of the simulation results, we estimated genomic cline parameters to quantify genome-wide variation in patterns of introgression for different genetic architectures of reproductive isolation. For each simulated dataset, we estimated cline parameters using the Bayesian genomic cline model with ICARρ priors described in §2. Posterior probability distributions for all model parameters were estimated using Markov chain Monte Carlo (MCMC). Computer software that implements MCMC estimation for the Bayesian genomic cline model parameters was written by the authors using C++ and the GNU Scientific Library [47]. For each simulated dataset, we obtained 22 500 MCMC samples from the posterior probability distribution, which were sampled every other MCMC iteration following a 5000 iteration burnin. We ran a single MCMC chain for each dataset, and visually assessed convergence to the stationary distribution and chain mixing by haphazard inspection of sample history plots. Parameter estimates were based on the median and 95 per cent equal tail probability interval of marginal posterior probability distributions. We designated local and global outlier loci as described in §2*c*, and using $q_N = 0.05$ for local outliers and $\overset{*}{q}_N = 0.01$ for global outliers (we used a more stringent cutoff for global outliers because it was much more common for loci to have extreme parameters relative to the global, zero-centred distribution than a distribution centred on the local mean).

The scale and form of genomic autocorrelations in genome-wide patterns of introgression could provide important information about the genetic architecture of reproductive isolation and effect our interpretation of loci classified as outliers. As a measure of autocorrelation for genomic cline parameters, we calculated Moran's I [48,49] at various recombination distances (measured in centimorgans) according to the following equation:

$$I_{(k1,k2]} = \frac{n_k \sum_i \sum_{i'} r_{ii'}\{(k1,k2]\}\alpha_i\alpha_{i'}}{R_{(k1,k2]} \sum_i \alpha_i^2}. \tag{3.1}$$

In equation (3.1), the r_{ii}' are binary variables that take on a value of one if the recombination distance between locus i and i' is $k1 < r_{ii}' \leq k2$, and otherwise take on a value of 0. $R_{(k1,k2]}$ is the sum of all r_{ii}' and n_k is the total number of loci. $I_{(k1,k2]}$ was calculated for β in a similar matter. We calculated $I_{(k1,k2]}$ for α and β with $k1 = [0,50]$ with 2 cM increments, and $k2 = k1 + 2$ (or infinity for $k1 = 50$). Estimates of Moran's I were used to construct correlograms, which are plots of autocorrelation as a function of genetic distance.

4. RESULTS

The distribution of hybrid indexes for each simulated admixed population was flat or unimodal (see electronic supplementary material, figure S1). Flat hybrid index distributions were more common when dispersal from the parental populations was high ($m = 0.2$), whereas unimodal distributions were observed more frequently when dispersal was low ($m = 0.05$; compare electronic supplementary material, figure S1*a–c* with *e*,*f*). Under most conditions, the mean hybrid index for each admixed population was close to 0.5 (50% ancestry from each parental population), but underdominance with one or several genes of moderately large or very large effect caused the distribution of hybrid indexes to shift towards zero or one (e.g. electronic supplementary material, figure S1*c*; table 1). Mean interspecific heterozygosity in the admixed populations tended to be close to 0.5, but a high dispersal rate from the parental populations increased the variance in interspecific heterozygosity among individuals (electronic supplementary material, figure S1; table 1).

Mean locus-specific ancestry and interspecific heterozygosity varied among genetic regions regardless of the simulated demographic conditions or genetic architectures of isolation (figure 2 and table 1). However, variation in both metrics was most pronounced with underdominance with one or several loci of moderately large or very large effect (e.g. figure 2*c*,*d*), or BDM incompatibilities with one or many interacting locus pairs of moderately large or very large effect (e.g. figure 2*f*). With these genetic architectures, genetic regions near loci that caused fitness variation had extreme values for mean ancestry (i.e. values with a greater absolute deviation from 0.5) and reduced mean interspecific heterozygosity. Consistent with the distribution of hybrid indexes, underdominance with one or several genes of moderately large or very large effect shifted mean ancestry genome wide.

Similar to the direct observations of ancestry and interspecific heterozygosity in the simulations, estimated genomic cline parameters varied among genetic regions (figure 3). We observed the most genome-wide variation in the genomic cline centre parameter (α) with underdominance involving one locus of very large effect ($s = 0.65$) or five loci of moderately large effect ($s = 0.2$) that were located near each other on a single chromosome, with underdominance and 25 generations of admixture, or with BDM incompatibilities involving one or many interacting locus pairs with $s = 0.2$ or $s = 0.65$ (figure 3 and table 1). We obtained similar results for the genomic cline rate parameter (β), but genome-wide variation was much greater with underdominance involving five loci of moderately large effect on a single chromosome than any other genetic architecture, and genome-wide variation was modest even after 25 generations of admixture when underdominance was associated with many genes of small effect ($s = 0.01$, $N_s = 110$). Genome-wide variation in β with BDM incompatibilities was also generally low, except with 20 interacting locus pairs of moderately large effect ($s = 0.2$). In general, genome-wide variation in genomic cline parameters was lower with higher dispersal rates from parental populations. The lowest levels of genome-wide variation for both α and β were associated with weak isolation overall or a diffuse genomic architecture of isolation (i.e. genes causing reduced fitness were spread evenly throughout the genome; figure 3 and table 1).

The total number of statistical outlier loci varied considerably for simulations conducted with different demographic conditions and genetic architectures of isolation (figure 3; electronic supplementary material, table S1). For example, high dispersal from parental populations ($m = 0.2$) led to very few outliers (a total of 10 ($s = 0.01$, $N_s = 110$) or 15 ($s = 0.2$, $N_s = 5$)

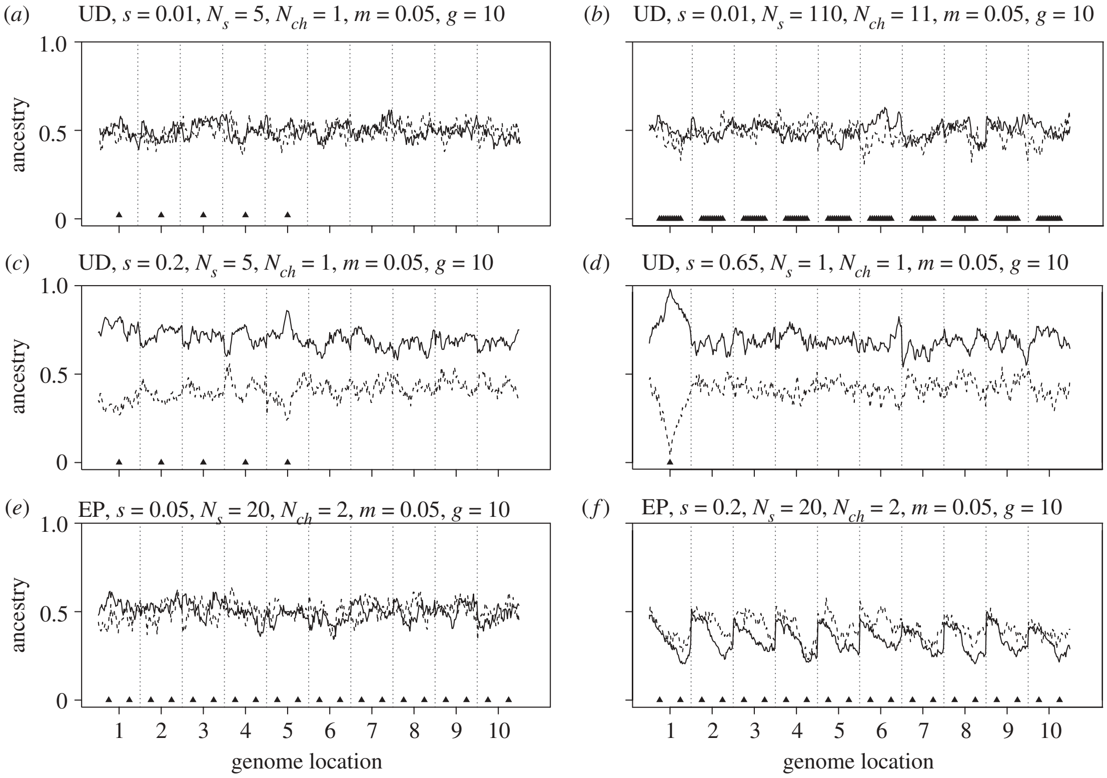

Figure 2. Plots depict the observed proportion of ancestry from population 1 (solid line) and the observed proportion of interspecific heterozygotes (dashed line) as a function of genome location. Vertical dotted grey lines denote chromosome boundaries. Triangles denote loci that contribute to fitness. Each panel gives results for a single replicate for a set of simulation conditions (UD, underdominance; EP, BDM epistatic incompatibility).

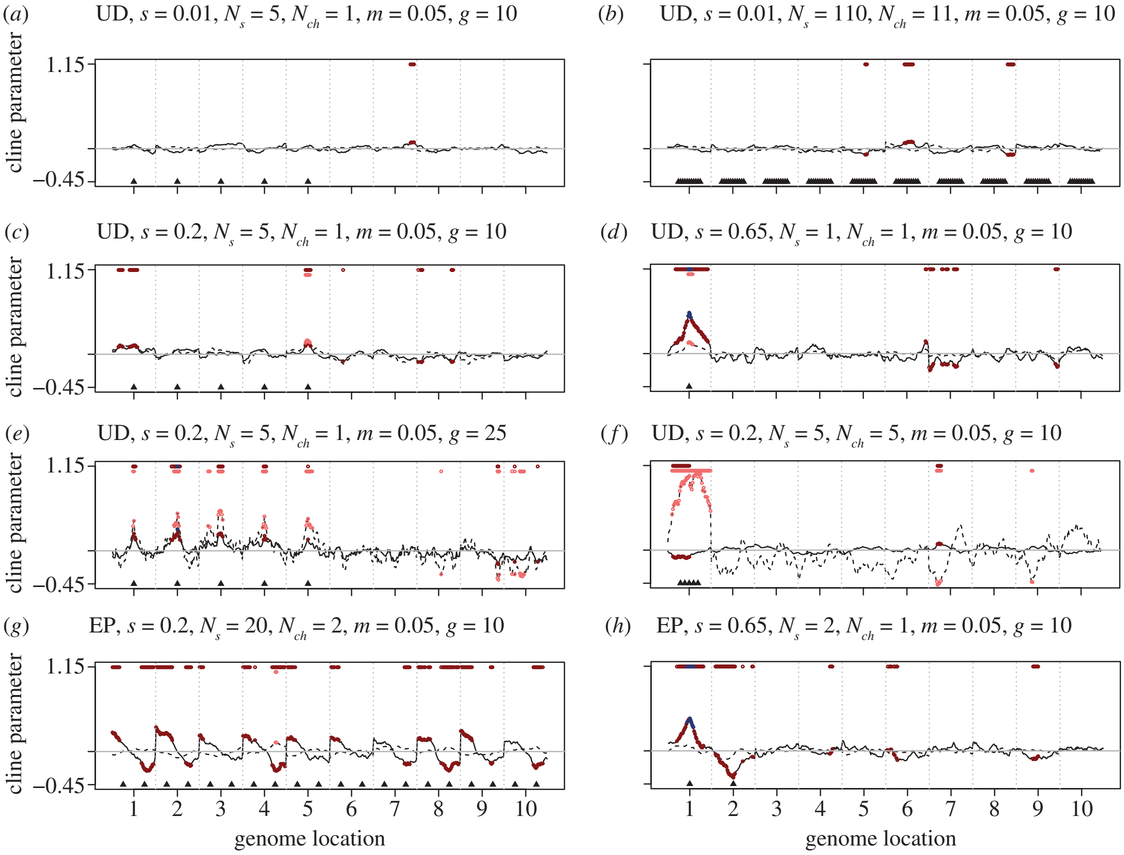

Figure 3. Plots depict estimates of genomic cline parameters α (solid black line) and β (dashed black line) as a function of genome location. Vertical dotted grey lines denote chromosome boundaries and solid grey lines correspond to a parameter value of zero. Triangles denote loci that contribute to fitness. Coloured dots denote outlier loci: red (global α), pink (global β), dark blue (local α) and light blue (local β). Each panel gives results for a single replicate for a set of model conditions (UD, underdominance; EP, BDM epistatic incompatibility).

across 10 replicate simulations), whereas over 449 outlier loci were detected for underdominance with $s = 0.01$, $N_s = 110$, $m = 0.05$ and $g = 25$. Outlier loci were more common with respect to α than β, and global outliers were much more common than local outliers (no local β outliers were detected; electronic supplementary material, table S1). The correspondence between outlier loci and genetic regions affecting fitness varied among simulation conditions and cline parameters. We define true outliers as loci within 10 cM of genetic regions contributing to isolation, however, this designation is arbitrary and is not meant as a general rule of thumb. Given this definition, global outlier loci were approximately 10 times more likely to be true outliers than false outliers with underdominance involving a single selected locus of very large effect ($s = 0.65$, $N_s = 1$, $N_{ch} = 1$ and $m = 0.05$, $g = 10$), five loci of moderately large effect and 25 generations of admixture ($s = 0.2$, $N_s = 5$, $N_{ch} = 1$, $m = 0.05$ and $g = 25$), five loci of moderately large effect that co-occurred on a single chromosome ($s = 0.2$, $N_s = 5$, $N_{ch} = 5$, $m = 0.05$ and $g = 10$), or BDM incompatibilities with two interacting loci of very large effect ($s = 0.65$, $N_s = 2$, $m = 0.05$ and $g = 10$; electronic supplementary material, table S1). However, for other simulated demographic conditions or genetic architectures global outlier loci were only about twice as likely to be true outliers rather than false outliers (e.g. underdominance with $s = 0.2$, $N_s = 5$, $N_{ch} = 1$, $m = 0.05$ and $g = 10$, or BDM incompatibilities with $s = 0.2$, $N_s = 20$, $N_{ch} = 2$, $m = 0.05$ and $g = 10$). Finally, when isolation was the product of genes with little individual effect, it was common for false and true outliers to be equally common (electronic supplementary material, table S1). Local outliers with respect to α were rare, but were nearly always true outliers (only three false outliers were recorded).

Genomic autocorrelation for cline parameters α and β was affected by demographic conditions and the genetic architecture of isolation (figure 4; electronic supplementary material, figure S2). Genomic autocorrelation (measured by Moran's I) was often greater than 0.9 and close to one for genetic regions separated by 2 cM (table 1). This result indicates that nearby genetic regions generally had similar patterns of introgression. We found considerable variation among model conditions for the rate at which genomic autocorrelation decreased with increasing recombination distance. For example, the mean genetic distance required for genomic autocorrelation to drop below half of $I_{(0\text{cM},2\text{cM}]}$ (denoted as $I_{1/2}$) was 32.8 (α) or 48.2 cM (β) for underdominance involving five linked loci with $s = 0.2$, but was only 16.2 (α) or 20.8 cM (β) for underdominance with five unlinked loci of weak effect ($s = 0.01$; table 1). Moreover, with some genetic architectures, genomic autocorrelation rapidly approached zero or took on negative values as the genetic distance between loci increased (e.g. or BDM incompatibilities with 20 loci of moderately large effect), whereas we observed values of $I > 0.5$ for all genetic distances less than 50 cM for other simulation conditions (e.g. α for underdominance involving one locus with $s = 0.65$; figure 4; electronic supplementary material, figure S2). Negative genomic autocorrelations arise because of epistatically interacting loci and because α and β parameters are constrained to sum to zero across loci.

5. DISCUSSION

The results indicate that different genetic architectures of reproductive isolation often, but not always, leave distinct genomic signatures in admixed populations. We observed increased heterogeneity in patterns of introgression among genetic regions when reproductive isolation was caused by one or several genes with moderately large to very large fitness effects, rather than many genes with small fitness effects. Similarly, genome-wide heterogeneity in the rate and form of introgression was higher when genetic regions affecting fitness were clumped together, rather than spread out across a greater number of chromosomes. For comparable selection intensities, underdominance and BDM incompatibilities caused similar levels of genomic heterogeneity in patterns of introgression, although the former had a greater effect on genomic cline rate (β). Patterns of genomic autocorrelation in cline parameters were complicated, with different genetic architectures often leading to moderately to strikingly different rates of change in autocorrelation with genetic distance. Genetic architectures of isolation based on underdominance with the fewest loci contributing to isolation, particularly those with a single locus of very large effect, or BDM epistasis with interacting locus pairs on different chromosomes caused the greatest long-range genomic autocorrelation in cline parameters, whereas BDM epistasis with interacting locus pairs on the same chromosome resulted in some of the lowest long-range genomic autocorrelations (because of selection for different parental ancestry at each locus).

BDM incompatibilities caused greater genomic heterogeneity in α (genomic cline centre) than β (genomic cline rate) and the same was often true of underdominance. This observation highlights an important distinction between geographical and genomic introgression. The geographical extent of introgression is reduced for loci causing BDM incompatibilities or underdominance [9,50]. Under a few conditions, we obtained a similar pattern whereby alleles inherited from one population did not introgress into the alternative genomic background (positive β). However, the observed genomic heterogeneity in α indicates that, within the admixed population, introgession of loci causing BDM incompatibilities or underdominance was often extreme and ancestry blocks from one of the parental populations achieved a high frequency across the hybrid index gradient at the expense of ancestry blocks from the other parental population. With underdominant selection, α for selected loci tended to deviate in a single direction from expectations based on genome-wide admixture (i.e. all positive or all negative). Whether ancestry from parental population 0 or 1 was favoured varied stochastically among simulations. This finding indicates that drift during the first few generations affected the genomic composition of the admixed population, which influenced whether alleles inherited from parental population 0 or 1 had a higher

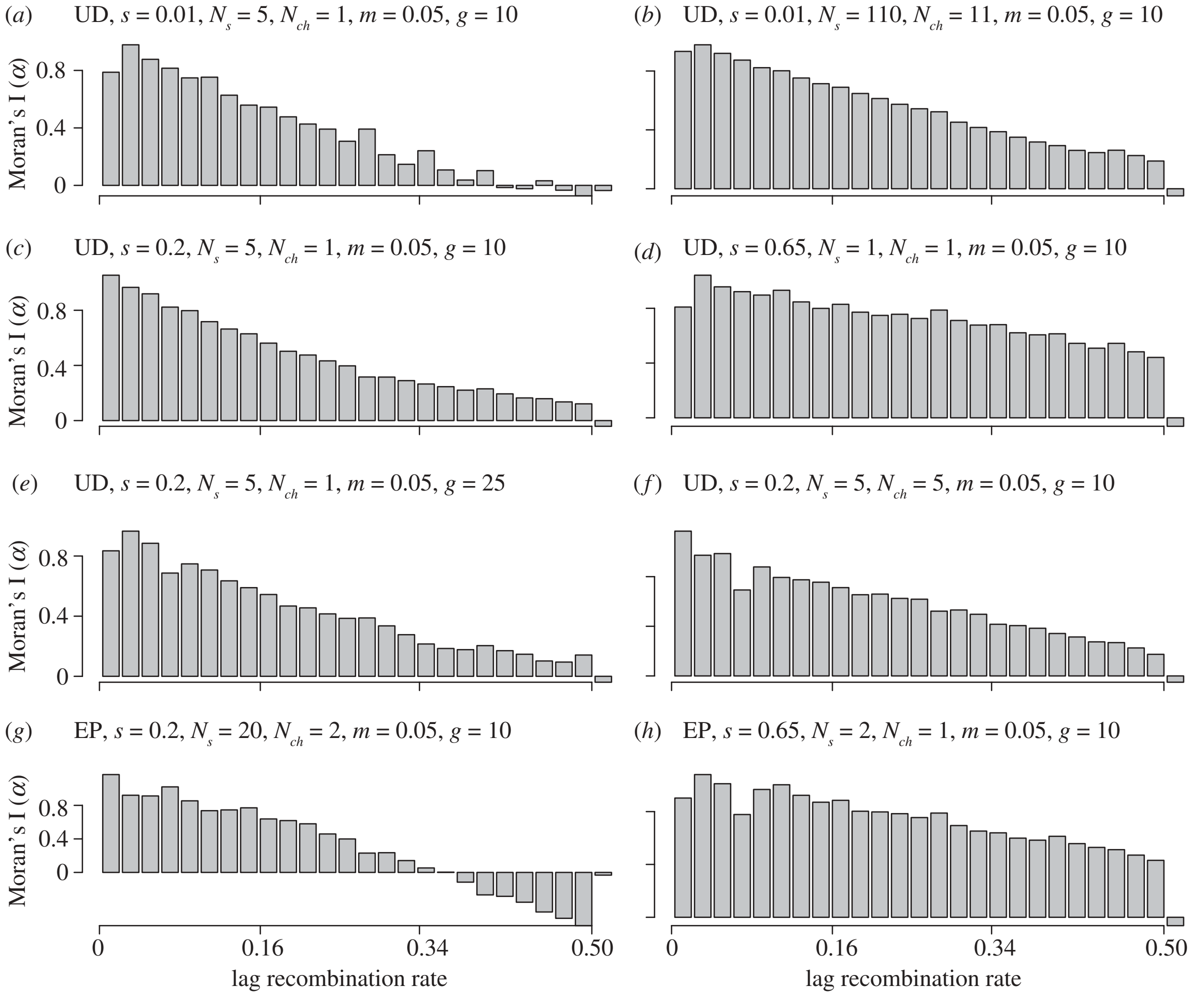

Figure 4. Correlograms depict genomic spatial autocorrelation for estimated α cline parameters. Each panel gives results for a single replicate for a set of model conditions. The model conditions shown here correspond to those described for figure 3.

marginal fitness. With BDM incompatibilities, positive values of α were often associated with the first locus of each pair and negative values of α were often associated with the second locus of each pair. This pattern suggests selection increased ancestry associated with the ancestral alleles (*a* and *b*) at each locus. This should not be surprising as the ancestral alleles had higher marginal fitness than the derived alleles (*A* and *B*), which causes a form of directional selection. When considering a single locus, underdominant selection and BDM incompatibilities can affect α in the same manner as directional selection [23], which suggests it might not be possible to discern the specific form of selection affecting a locus. Thus, previous attempts (including our own) to associate a specific form of selection (i.e. underdominance, epistasis or directional selection) with a locus based on the sign or magnitude of cline parameters were perhaps overly simplistic and naïve [12,26,27,29].

The simulation results indicate that the genomic signature of reproductive isolation is also profoundly affected by demographic conditions. For example, relative to $m = 0.05$, increased dispersal from parental populations ($m = 0.2$) caused reduced heterogeneity in patterns of introgression, but slightly increased long-range genomic autocorrelations, particularly for genomic cline rate parameter β. This is not surprising, as increased dispersal should retard shifts in ancestry owing to selection in admixed individuals, and inject parental chromosomal blocks and increase admixture linkage disequilibrium [15,40,51]. Conversely, relative to 10 generations of admixture, 25 generations of admixture caused increased heterogeneity in patterns of introgression across the genome and reduced long-range genomic autocorrelations. Again, this result could be expected, as an increased number of generations allows more time for recombination in admixed individuals to break up parental haplotype blocks. These results demonstrate that our ability to map the genetic architecture of isolation in a hybrid zone very clearly depends on its demographic history.

Demographic conditions also affected genome-wide admixture. Although the form and strength of isolation can affect the distribution of hybrid indexes in an admixed population, we found that this distribution was altered to a great extent by the dispersal rate from parental populations. With high dispersal, we generally observed a flat (uniform) distribution of hybrid indexes (presumably higher dispersal rates would have led to a bimodal distribution), whereas low-dispersal rates resulted in a unimodal distribution.

Although these results are consistent with the hypothesis that bimodal hybrid zones are associated with nearly complete (prezygotic) reproductive isolation [7], they suggest that this might be an oversimplification. Rather, the results indicated that the modality of a hybrid zone is likely a function of the strength and form of isolation, as well as the dispersal rate from (or geographical overlap of) parental populations. High dispersal rates or geographical overlap of parental populations coupled with strong isolation should yield a bimodal hybrid zone, whereas even with strong isolation a geographically isolated admixed population will often have a unimodal distribution of hybrid indexes, which might be centred well away from 0.5. Thus, it is difficult to draw inferences about the nature of isolation based on the observation of a unimodal hybrid zone. Interpretation would require an understanding of the geographical context in which the unimodal hybrid zone occurs.

The results suggest that our ability to identify genetic regions associated with reproductive isolation in admixed populations based on variable introgression depends on population demography and the genetic architecture of isolation. Specifically, when reproductive isolation is caused by a modest number of genetic regions with moderate to large effects on fitness, global outliers are enriched for genetic regions associated with isolation. The same is true for local outliers, when they occur. However, when isolation is caused by genes with small fitness effects (whether few or many genes), outliers are at best marginally enriched for genetic regions associated with isolation. An optimistic assessment of these results is that variable introgression can be used most effectively to map the genetic architecture of isolation when the identity and location of genes contributing to isolation would be most interesting (i.e. when these genes have moderate to large effects on fitness). Even under these circumstances, it will be difficult to identify the true loci underlying reproductive isolation as their effect can extend over large portions of a chromosome. When patterns of introgression are relatively homogeneous, little confidence should be placed in individual outliers. But this pattern, if coupled with independent knowledge indicating moderate to strong overall isolation between hybridizing populations, would provide preliminary support that isolation is the result of many genes with small effects on fitness. Finally, the results show that a greater number of generations of admixture can radically increase the correspondence between outlier loci and the genetic architecture of isolation, whereas high rates of dispersal from parental populations can retard evolution in the admixed population to the extent that outliers are very rare, making the genetic architecture of isolation difficult to map.

We simulated complicated genetic architectures of isolation, but the genetic architecture of isolation in many natural hybrid zones is likely more complex. For example, documented genome-wide variation in genomic clines from natural populations certainly tends to be greater than that observed from these simulated data sets [23,27,29]. Moreover, each simulation involved a single form of selection and equal fitness effects for all selected loci. Neither of these assumptions is expected to hold in natural populations. Nonetheless, we have uncovered considerable complexity in patterns of introgression that result from even these relatively simple simulations. This suggests that the study of the genomics of isolation in hybrids will benefit from additional theoretical work to further develop our expectations for the genomic consequences of different isolating barriers. The complexity also suggests that specific evolutionary causes of heterogeneous introgression among loci should be inferred with caution.

Our findings point to both promise and potential limitations for discovering the genomics of speciation by studying admixed populations. Further progress will require genome-wide studies of introgression in hybrid zones, coupled with appropriate model-based analyses. Indeed, genome-wide sequence data have recently been published for several species known to hybridize in nature [52–54]. For example, Nadeau *et al.* [55] report genetic divergence between hybridizing lineages of *Heliconius* butterflies associated with colour pattern genes. These and other systems are well suited for genome-wide surveys of introgression in hybrid zones, which will complement studies of genome-wide divergence.

With increasing sequence coverage of genomes, model-based population genomic analyses of hybrid zones will need to account for linkage among genetic regions. We have proposed one way of doing this by specifying ICARρ priors for cline parameters. Hidden Markov models (HMM), which explicitly model correlated parameter states or evolutionary models along a chromosome, provide an alternative method [36,56,57]. It would be possible to introduce a HMM for the genomic cline parameters (α and β) with a finite number of model states. Falush *et al.* [58] proposed a Markov model for ancestry along a chromosome in admixed individuals. This model could conceivably be combined with ICARρ priors for cline parameters to model correlated patterns of introgression and admixture linkage disequilibrium in a complementary manner. We intend to explore this possibility in future work.

Finally, interpreting genetic patterns in the context of the complex genomic environment in which they are embedded (physical linkage relationships, density of genes and other functional regions, etc.) will advance our understanding of speciation. For example, a study of genomic divergence and local and long-distance linkage disequilibrium in threespine stickleback populations demonstrates the utility of analysing associations among loci, which leads the authors to suggest a new model for evolution in sticklebacks [59]. Similarly, our simulation results indicate that statistical measures of association across the genome (our genomic autocorrelations) capture aspects of genome introgression that have previously received little attention in the genomics of speciation, but that might support critical future insights.

We thank Patrik Nosil and Jeff Feder for organizing this special issue and making this contribution possible. We thank two anonymous reviewers, Patrik Nosil and Jeff Feder for comments that improved this manuscript. We also thank Jeffrey Lang and the Department of Geology and Geophysics (University of Wyoming) for access to the Seismic computer cluster. This work was supported by NSF DDIG 1011173 to Z.G. and NSF DBI 0701757 to C.A.B.

REFERENCES

1 Ting, C., Tsaur, S., Wu, M. & Wu, C. 1998 A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* **182**, 1501–1504. (doi:10.1126/science.282.5393.1501)

2 Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C. & Forejt, J. 2009 A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* **323**, 373–375. (doi:10.1126/science.1163601)

3 Nosil, P. & Schluter, D. 2011 The genes underlying the process of speciation. *Trends Ecol. Evol.* **26**, 160–167. (doi:10.1016/j.tree.2011.01.001)

4 Endler, J. A. 1977 *Geographic variation, speciation, and clines*. Princeton, NJ: Princeton University Press.

5 Barton, N. H. & Hewitt, G. M. 1989 Adaptation, speciation and hybrid zones. *Nature* **341**, 497–503. (doi:10.1038/341497a0)

6 Hewitt, G. M. 1988 Hybrid zones: natural laboratories for evolution studies. *Trends Ecol. Evol.* **3**, 158–166. (doi:10.1016/0169-5347(88)90033-X)

7 Jiggins, C., Naisbit, R., Coe, R. & Mallet, J. 2001 Reproductive isolation caused by colour pattern mimicry. *Nature* **411**, 302–305. (doi:10.1038/35077075)

8 Mallet, J. 2005 Hybridization as an invasion of the genome. *Trends Ecol. Evol.* **20**, 229–237. (doi:10.1016/j.tree.2005.02.010)

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

10 Buerkle, C. A. & Lexer, C. 2008 Admixture as the basis for genetic mapping. *Trends Ecol. Evol.* **23**, 686–694. (doi:10.1016/j.tree.2008.07.008)

11 Nosil, P. & Feder, J. L. 2012 Genomic divergence during speciation: causes and consequences. *Phil. Trans. R. Soc. B* **367**, 332–342. (doi:10.1098/rstb.2011.0263)

12 Gompert, Z. & Buerkle, C. A. 2009 A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Mol. Ecol.* **18**, 1207–1224. (doi:10.1111/j.1365-294X.2009.04098.x)

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

14 Barton, N. H. 1983 Multilocus clines. *Evolution* **37**, 454–471. (doi:10.2307/2408260)

15 Szymura, J. M. & Barton, N. H. 1986 Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution* **40**, 1141–1159. (doi:10.2307/2408943)

16 Barton, N. H. 1993 The probability of fixation of a favoured allele in a subdivided population. *Genet. Res.* **62**, 149–157. (doi:10.1017/S0016672300031748)

17 Payseur, B. A. 2010 Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Mol. Ecol. Resour.* **10**, 806–820. (doi:10.1111/j.1755-0998.2010.02883.x)

18 Macholan, M., Munclinger, P., Sugerkova, M., Dufkova, P., Bimova, B., Bozikova, E., Zima, J. & Pialek, J. 2007 Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. *Evolution* **61**, 746–771. (doi:10.1111/j.1558-5646.2007.00065.x)

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

20 Teeter, K. C. *et al.* 2008 Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res.* **18**, 67–76. (doi:10.1101/gr.6757907)

21 Carling, M. D. & Brumfield, R. T. 2009 Speciation in *Passerina* buntings: introgression patterns of sex-linked loci identify a candidate gene region for reproductive isolation. *Mol. Ecol.* **18**, 834–847. (doi:10.1111/j.1365-294X.2008.04038.x)

22 Lexer, C., Buerkle, C. A., Joseph, J. A., Heinze, B. & Fay, M. F. 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. (doi:10.1038/sj.hdy.6800922)

23 Gompert, Z. & Buerkle, C. A. 2011 Bayesian estimation of genomic clines. *Mol. Ecol.* **20**, 2111–2127. (doi:10.1111/j.1365-294X.2011.05074.x)

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

25 Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E. G. & Risch, N. J. 2007 Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am. J. Hum. Genet.* **81**, 626–633. (doi:10.1086/520769)

26 Macholan, M., Baird, S. J. E., Dufkova, P., Munclinger, P., Bimova, B. V. & Pialek, J. 2011 Assessing multilocus introgression patterns: a case study on the mouse X chromosome in Central Europe. *Evolution* **65**, 1428–1446. (doi:10.1111/j.1558-5646.2011.01228.x)

27 Nolte, A. W., Gompert, Z. & Buerkle, C. A. 2009 Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. *Mol. Ecol.* **18**, 2615–2627. (doi:10.1111/j.1365-294X.2009.04208.x)

28 Gompert, Z., Lucas, L. K., Fordyce, J. A., Forister, M. L. & Nice, C. C. 2010 Secondary contact between *Lycaeides idas* and *L. melissa* in the Rocky Mountains: extensive introgression and a patchy hybrid zone. *Mol. Ecol.* **19**, 3171–3192. (doi:10.1111/j.1365-294X.2010.04727.x)

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

30 Wu, C. I. 2001 The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865. (doi:10.1046/j.1420-9101.2001.00335.x)

31 Moyle, L. C., Muir, C. D., Han, M. V. & Hahn, M. W. 2010 The contribution of gene movement to the 'two rules of speciation'. *Evolution* **64**, 1541–1557. (doi:10.1111/j.1558-5646.2010.00990.x)

32 Buerkle, C. A. & Rieseberg, L. H. 2001 Low intraspecific variation for genomic isolation between hybridizing sunflower species. *Evolution* **55**, 684–691. (doi:10.1554/0014-3820(2001)055[0684:LIVFGI]2.0.CO;2)

33 Sweigart, A. L., Mason, A. R. & Willis, J. H. 2007 Natural variation for a hybrid incompatibility between two species of *Mimulus*. *Evolution* **61**, 141–151. (doi:10.111/j.1558-5646.2007.00011.x)

34 Good, J. M., Handel, M. A. & Nachman, M. W. 2008 Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution* **62**, 50–65. (doi:10.1111/j.1558-5646.2007.00257.x)

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

36 Hahn, M. 2006 Accurate inference and estimation in population genomics. *Mol. Biol. Evol.* **23**, 911–918. (doi:10.1093/molbev/msj094)

37 Sun, D. C., Tsutakawa, R. K. & Speckman, P. L. 1999 Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika* **86**, 341–350. (doi:10.1093/biomet/86.2.341)

38 Congdon, P. 2006 *Bayesian statistical modeling. Wiley series in probability and statistics*, 2nd edn. Chichester, UK: John Wiley and Sons, Ltd.

39 Guo, F., Dey, D. K. & Holsinger, K. E. 2009 A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *J. Am. Stat. Assoc.* **104**, 142–154. (doi:10.1198/jasa.2009.0010)

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

41 Bateson, W. 1909 Heredity and variation in modern lights. In *Darwin and modern science* (ed. A. C. Seward), pp. 85–101. Cambridge, UK: Cambridge University Press.

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

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

44 Turelli, M. & Orr, H. A. 2000 Dominance, epistasis and the genetics of postzygotic isolation. *Genetics* **154**, 1663–1679.

45 Brideau, N. J., Flores, H. A., Wang, J., Maheshwari, S., Wang, X. & Barbash, D. A. 2006 Two Dobzhansky–Muller genes interact to cause hybrid lethality in *Drosophila*. *Science* **314**, 1292–1295. (doi:10.1126/science.1133953)

46 Orr, H. A. & Turelli, M. 2001 The evolution of postzygotic isolation: accumulating Dobzhansky–Muller incompatibilities. *Evolution* **55**, 1085–1094.

47 Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M. & Rossi, F. 2009 *GNU scientific library: reference manual*. Bristol, UK: Network Theory Ltd.

48 Moran, P. 1950 Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23.

49 Epperson, B. K. 2003 *Geographical genetics, monographs in population biology*. Princeton, NJ: Princeton University Press.

50 Gavrilets, S. & Cruzan, M. B. 1998 Neutral gene flow across single locus clines. *Evolution* **52**, 1277–1284. (doi:10.2307/2411297)

51 Slatkin, M. 1987 Gene flow and the geographic structure of natural populations. *Science* **236**, 787–792. (doi:10.1126/science.3576198)

52 Gompert, Z., Forister, M. L., Fordyce, J. A., Nice, C. C., Williamson, R. & Buerkle, C. A. 2010 Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Mol. Ecol.* **19**, 2455–2473. (doi:10.1111/j.1365-294X.2010.04666.x)

53 Forister, M. L., Gompert, Z., Fordyce, J. A. & Nice, C. C. 2011 After 60 years, an answer to the question: what is the Karner blue butterfly? *Biol. Lett.* **7**, 399–402. (doi:10.1098/rsbl.2010.1077)

54 Lawniczak, M. K. N. *et al.* 2010 Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* **330**, 512–514. (doi:10.1126/science.1195755)

55 Nadeau, N. J. *et al.* 2012 Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Phil. Trans. R. Soc. B* **367**, 343–353. (doi:10.1098/rstb.2011.0198).

56 Boitard, S., Schloetterer, C. & Futschik, A. 2009 Detecting selective sweeps: a new approach based on hidden Markov models. *Genetics* **181**, 1567–1578. (doi:10.1534/genetics.108.100032)

57 Kern, A. D. & Haussler, D. 2010 A population genetic hidden Markov model for detecting genomic regions under selection. *Mol. Biol. Evol.* **27**, 1673–1685. (doi:10.1093/molbev/msq053)

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

59 Hohenlohe, P. A., Bassham, S., Currey, M. & Cresko, W. A. 2012 Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Phil. Trans. R. Soc. B* **367**, 395–408. (doi:10.1098/rstb.2011.0245)