Population-genomic inference of the strength and timing of selection against gene flow

Simon Aeschbacher[1,a], Jessica P. Selby[2], John H. Willis[2], and Graham Coop[1]

14 September 2016

[1]Department of Evolution and Ecology, University of California, Davis, CA 95616
[2]Department of Biology, Duke University, Durham, NC 27708
[a]saeschbacher@mac.com

Abstract

The interplay of divergent selection and gene flow is key to understanding how populations adapt to local environments and how new species form. Here, we use DNA polymorphism data and genome-wide variation in recombination rate to jointly infer the strength and timing of selection, as well as the baseline level of gene flow under various demographic scenarios. We model how divergent selection leads to a negative correlation between recombination rate and genetic differentation among populations. Our theory shows that the selection density, i.e. the selection coefficient per base pair, is a key parameter underlying this relationship. We then develop a procedure for parameter estimation and apply it to two datasets from *Mimulus guttatus*. First, we infer a strong signal of adaptive divergence in the face of gene flow between populations growing on and off phytotoxic serpentine soils. However, the genome-wide intensity of this selection is not exceptional compared to what *M. guttatus* populations may typically experience when adapting to local conditions. Second, we find that selection against genome-wide introgression from the selfing sister species *M. nasutus* has acted to maintain a barrier between these two species over the last 250 to 500 ky. Our study provides a theoretical framework for linking genome-wide patterns of divergence and recombination with the underlying evolutionary mechanisms that drive this differentiation.

Estimating the timing and strength of divergent selection is fundamental to understanding the evolution and persistence of organismal diversity [1–3]. Genes underlying local adaptation and speciation act as barriers to gene flow, such that genetic divergence around these loci is higher compared to the rest of the genome. However, a framework that explicitly links observable patterns of DNA polymorphism with the underlying evolutionary mechanisms and allows for robust parameter inference has so far been missing [4].

One way of studying adaptive genomic divergence in the face of gene flow is to apply methods for demographic inference to scenarios of speciation [e.g. 5, 6]. This approach allows dating population splits and inferring the presence or absence of gene flow, yet generally does not explicitly account for natural selection [but see 7]. Another approach is to scan genomes for loci that are statistical outliers of divergence among populations. These scans are used to identify candidate loci underlying speciation or local adaptation [e.g. 8, 9], and include the search for so-called genomic islands of divergence [e.g. 10], i.e. extended genomic regions of elevated divergence. Methods of this type can be confounded by other modes of selection as well as demography, and will always propose a biased subset of candidate loci [11, 12].

A third approach is to test for a negative correlation between absolute genetic divergence and recombination rate across the genome [e.g. 13–15]. This approach is based on the prediction that divergence will be higher in regions of the genome where genetic linkage between neutral sites and loci under divergent selection is higher on average [16]. Testing for this pattern of a negative correlation is powerful because it aggregates information across the entire genome and because it is specific to divergent selection with gene flow [17]. However, this approach is currently purely descriptive.

Here, we develop novel theory describing the pattern used by this third approach, and a way of inferring the underlying parameters. Unlike the first two approaches, ours explicitly accounts for selection and its effect on neutral variation, allows estimation of the strength and timing of selection and gene flow, and is robust to various confounding factors. We apply it to DNA sequence and recombination data from *Mimulus guttatus*, showing evidence for adaptive divergence and maintenance of a species barrier despite gene flow.

Idea of Approach and Population-Genomic Model

Our approach exploits the genome-wide variation in recombination rate and its effect on genetic divergence. Divergent selection reduces effective gene flow at neutral sites, and this effect decreases with the recombinational distance from the loci under selection. We conceptualise this in terms of the effective migration rate and the expected pairwise between-population coalescence time (Fig. 1A). The latter directly relates to the absolute genetic diversity between populations, a quantity that is readily estimated from DNA sequence data. Our model considers two populations of diploids with effective sizes N_1 and N_2 and non-overlapping generations. In population 1, a balance between one-way gene flow from population 2 at rate m per generation and local directional selection is maintained for τ generations before the present. In this *migration–selection* (MS) phase (Fig. 1A), selection against maladaptive immigrant alleles acts at an arbitrary number of biallelic loci that we call *migration–selection polymorphisms* (MSPs). At each MSP, allele A_1 is favoured in population 1 over allele A_2 by an average selection coefficient s, while A_2 is introduced by gene flow. We assume additive fitness and no dominance.

Prior to the MS phase, we assume a *panmictic* (P) phase in an ancestral population of effective size N_0 that starts τ generations ago and extends into the past (Fig. 1A). We call this the (MS)P demographic scenario. The P phase can be exchanged for an ancestral *migration* (M) phase with gene flow at rate m_0, resulting in what we call the (MS)M scenario. Here, we use the (MS)P and (MS)M scenarios to describe our approach. In the Supporting Information (SI) we provide extensions to more general scenarios that include an intermediate *isolation* (I) phase (SI Appendix A, Fig. S1.1, Table S1.2).

We assume that the MSPs occur at a constant rate ν per base pair, such that the distance between consecutive MSPs is approximately exponentially distributed with mean $1/\nu$ base pairs. Finally, let the per-base pair recombination rate be r_{bp}. In summary, the key parameters are the mean selection coefficient s, the genomic density ν of MSPs, the baseline migration rate m, and the duration τ of the MS phase.

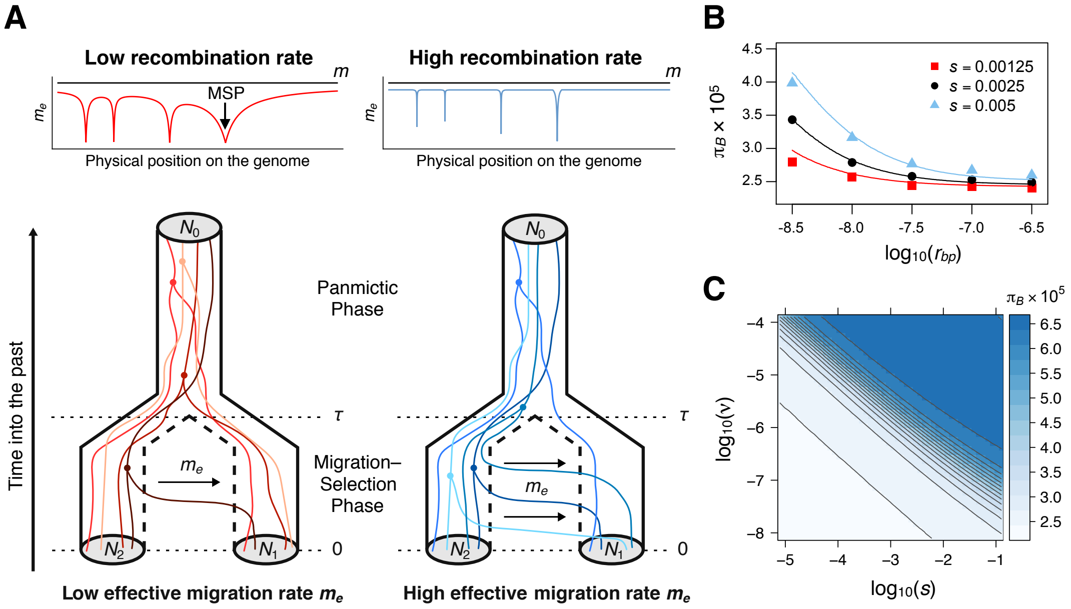

Figure 1. Divergent selection reduces gene flow and increases genetic divergence. (*A*) Selection against locally maladapted alleles at migration–selection polymorphisms (MSPs) reduces the effective migration rate m_e into population 1. The reduction is stronger in regions of low recombination (red, top left) and decreases the probability that lineages sampled in different populations migrate and coalesce. Realisations of the coalescence process are shown in the bottom left for the (MS)P scenario (Fig. S1.1). In regions of high recombination, m_e is reduced much less (blue, top right), such that migration events and earlier between-population coalescence times are more likely (bottom right). (*B*) The predicted between-population diversity $\pi_B = 2u\mathbb{E}[T_B]$ (curves) matches individual-based simulations (dots); error bars (±SE) are too short to be visible. The (MS)M multi-MSP scenario was used with $N_2 = 5000$, $u = 10^{-9}$, $\nu = 2.5 \times 10^{-7}$, $m = m_0 = 5 \times 10^{-4}$, and $\tau = 2 \times 2N_2$. (*C*) Approximately linear contour lines with slope -1 in the surface of π_B as a function of $\log_{10}(s)$ and $\log_{10}(\nu)$ support the compound parameter selection density, $\sigma = s\,\nu$. Here, $r_{\text{bp}} = 10^{-8}$ (1 cM/Mb); other parameters as in (*B*).

Average Effective Gene Flow and Selection Density

Selection against maladapted immigrant alleles acts as a barrier to gene flow in the MS phase. At a focal neutral site, the baseline migration rate m is reduced to an effective migration rate m_e [18]. Considering the nearest I up- and J downstream MSPs, the effective migration rate at the neutral site can be approximated as $m_e^{(I,J)} = m\, g^{(I)}\, h^{(J)}$, where

$$g^{(I)} \approx \left(1 + \frac{a_1}{k_1 r_{\text{bp}}}\right)^{-1} \prod_{i=2}^{I} \left(1 + \frac{a_i}{k_i r_{\text{bp}} + \sum_{n=1}^{i-1} a_n}\right)^{-1} \tag{1}$$

[19]. Here, k_i is the physical distance to the ith upstream MSP, and a_i its selection coefficient ($h^{(J)}$ is defined analogously for the downstream MSPs, Eq. S1.1). Equation (1) shows that the reduction in effective gene flow increases with the strength of selection at the MSPs, and decreases with their recombinational distance from the focal neutral site (Fig. 1A).

To understand how this effect translates from a given neutral site to the entire genome, we average over the possible genomic locations and selection coefficients of the MSPs. In doing so, we make the simplifying assumption of an infinite chromosome with a linear relationship between physical and genetic map distance. Integrating over the distances to all MSPs and assuming an exponential distribution of selection coefficients, we find that the expected effective migration rate $\mathbb{E}[m_e^{(I,J)}]$ depends on s, ν, and r_{bp} exclusively through σ/r_{bp}, where $\sigma = s\nu$ is the product of the mean selection coefficient times the density of MSPs. We call σ the 'selection density' (per base pair). For instance, with $I = J = 2$ we find

$$\mathbb{E}[m_e^{(2,2)}] \approx m\,[1 + 2\sigma/r_{\text{bp}} \ln(\sigma/r_{\text{bp}})]\,. \tag{2}$$

This is a good approximation if $\sigma/r_{\text{bp}} \lesssim 0.1$, i.e. if the recombination rate is at least ten times larger than the selection density, at which point effective gene flow is reduced by about 50 % (Fig. S1.2). Equation (2) shows that the mean effective gene flow decreases with selection density and increases with recombination rate. Adding increasing numbers of MSPs has a diminishing effect on $\mathbb{E}[m_e^{(I,J)}]$ (Fig. S1.2A), so that (2) captures the essential pattern if $\sigma/r_{\text{bp}} \lesssim 0.1$. That this exclusive dependence of $\mathbb{E}[m_e^{(I,J)}]$ on selection and recombination through the compound parameter σ/r_{bp} holds for any I and J means it also applies to the genome-wide average of m_e (SI Appendix A, Eq. S1.8). Note that σ/r_{bp} is the selection density per genetic map unit. Our results imply that doubling the number of MSPs has the same effect on average effective gene flow as doubling the mean selection coefficient. We therefore anticipate that, in practice, s and ν can be inferred only jointly as σ from population-genomic data in our framework.

Expected Pairwise Coalescence Time With Selection

To enable parameter inference from population-genomic data, we phrase our theory in terms of the expected coalescence time of two alleles, one from each population. The expected between-population coalescence time under neutrality, T_B, depends on the baseline migration rate m (Table S1.2). To incorporate the effect of selection, we substitute the effective migration rate for m. Averaging over all possible numbers and locations of MSPs, we obtain $\mathbb{E}[T_B]$, and can predict the between-population diversity π_B as $2\,u\,\mathbb{E}[T_B]$, where u is the mutation rate per base pair and generation. In the following, we obtain $\mathbb{E}[T_B]$ for two versions of our model, accounting for only the

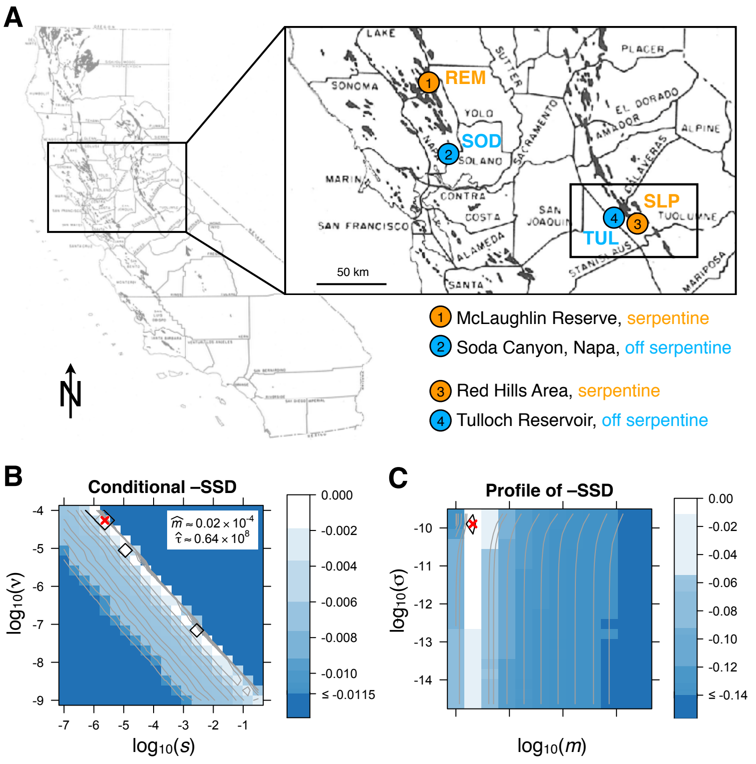

Figure 2. Geographic context of serpentine dataset and quasi-likelihood surfaces for population pair SLP × TUL. (*A*) Sampling sites in California, USA (modified with permission from [20]). (*B*) The negative sum of squared deviations (−SSD) for the selection coefficient s and the genomic density ν of MSPs, conditional on estimates of m and τ. The ridge with slope -1 confirms the compound parameter selection density, $\sigma = s\nu$. A cross denotes the point estimate and black hulls the 95% bootstrap confidence area. (*C*) Joint profile surface of the −SSD for the baseline migration rate m and the selection density σ, maximised over τ. Results are shown for the multi-MSP model and the (MS)P scenario, with genomic windows of 500 kb.

nearest-neighbouring MSP (one-MSP), or all possible numbers of MSPs (multi-MSP). To better reflect real genomes, we now assume a finite genome size and define $r_f = 0.5$ as the recombination rate that corresponds to free recombination, such that MSPs located more than $k_f = 1/(2r_{\text{bp}})$ base pairs from a neutral site are unlinked.

Selection at one locus

Under the one-MSP model, we implicitly assume that ν is small, i.e. $\nu \ll r_{\text{bp}}/s, m, \tau$, where s is now the selection coefficient at the single MSP. In the simplest case of the (MS)M scenario with $m_0 = m$ (Fig. S1.1C), we find that, for small ν, the expected pairwise between-population coalescence time at an average neutral site is

$$\mathbb{E}[T_B] \approx 2N_2 + \frac{1}{m} + \frac{1}{m}\frac{2\sigma}{r_{\text{bp}}}\left(e^{-m\tau}D + F\right) + \frac{1}{m}\frac{s}{r_f}e^{-2\nu k_f}, \tag{3}$$

where D and F depend on m, τ, and ν, (Materials and Methods). The first two terms in (3) are the expectation without selection [21] (Table S1.2). The third and fourth term reflect the increase in coalescence time if the MSP is linked ($k_1 < k_f$) and unlinked ($k_1 \geq k_f$) to the neutral site, respectively. Importantly, the term accounting for a linked MSP shows that σ/r_{bp} strongly determines $\mathbb{E}[T_B]$, although s, ν, and r_{bp} also enter (3) independently. Indeed, given r_{bp} and in the parameter range where (3) is a good approximation (i.e. for $\nu \ll r_{\text{bp}}/s, m, \tau$), the effect of selection on $\mathbb{E}[T_B]$ is entirely captured by σ (Fig. S1.4). For details and other demographic scenarios, see SI Appendix A.

Selection at multiple loci

Under the multi-MSP model, we explicitly account for all MSPs possibly present in the genome and for the average physical chromosome length. Finding $\mathbb{E}[T_B]$ is in principle similar to the one-MSP above, but amounts to averaging over all possible numbers and genomic locations of the MSPs. We wrote a Monte-Carlo integration program to do this (SI Appendix A). The result agrees very well with individual-based forward simulations (Figs. 1B and S1.5). Similar to the one-MSP model, if r_{bp} is given, $\mathbb{E}[T_B]$ depends on s and ν effectively only through the selection density σ (Fig. 1C). In fact, falling back on the idealising assumption of a global linear relationship between physical and genetic map distance (i.e. $r_f \to \infty$), we can prove that this holds exactly (SI Appendix A, Eq. S1.59). This corroborates σ as a key parameter and a natural metric to quantify genome-wide divergent selection in the face of gene flow.

Application to *Mimulus guttatus*

We developed an inference procedure based on our theory and applied it to two datasets from the predominantly outcrossing yellow monkeyflower (*Mimulus guttatus*), an important model system for speciation and local adaptation [22]. For both datasets, we fitted the multi-MSP version of our model to the empirical relationship between recombination rate (r_{bp}, estimated from a linkage map) and putatively neutral between-population diversity (π_B, estimated from 4-fold degenerate coding sites), after correcting the latter for genomic correlates and divergence to the outgroup *M. dentilobus* (SI Appendix B). Our procedure computes the sum of squared deviations (SSD) across genomic windows between these observed values of π_B and those predicted by our model, given the estimate of r_{bp} for each window and a set of parameter values. Minimising the SSD over a large grid of parameter values, we obtained point estimates for the selection density (σ), baseline migration rate (m), and duration of the MS phase (τ). We estimated 95 % non-parametric confidence intervals (CIs) for the parameters by doing a block-bootstrap over genomic windows (SI Appendix B). For both datasets, we fitted the (MS)P demographic scenario (Figs. 1A and Fig. S1.1D). We report results obtained with genomic windows of 500 kb. Results for windows of 100 and 1000 kb were very similar (SI Appendix B, SI Text 2). Genetic diversity *within* populations was not positively correlated with recombination rate, which means that background selection is unlikely to bias our findings [23].

Adaptive divergence maintained in the face of gene flow

Like many plant species [24], *M. guttatus* has locally adapted to serpentine outcrops throughout its range [25, p. 4]. While the mechanism and molecular basis of this adaptation are unresolved [26], strong differences in survival on serpentine soil exist between serpentine and non-serpentine ecotypes [25]. To investigate a population-genomic signal of local adaptation, we used whole-genome pooled-by-population sequencing of 324 individuals collected from two pairs of geographically close populations growing on and off serpentine soil in California (the serpentine dataset; Fig. 2A, SI Text 1). We inferred the strength of selection in serpentine populations (REM and SLP) against maladaptive immigrant alleles from the geographically closer off-serpentine population (SOD and TUL, respectively), assuming that the latter in each pair is a proxy for the source of gene flow. Fitting our model to the data, we found that the conditional surface of the −SSD (holding m and τ at their point estimates) showed a pronounced ridge for s and ν, with the 95 % confidence hull falling along this ridge (Fig. 2B). With parameters on a $\log_{10}$ scale, the slope of this ridge is −1, nicely confirming our theoretical result that s and ν can only be estimated jointly as their product, the selection density σ. We therefore adjusted our inference procedure to jointly infer m, τ, and σ instead of m, τ, s, and ν (SI Appendix B). This resulted in conditional −SSD surfaces for σ and m with a unique peak and tight confidence hulls (Fig. 2C).

For both serpentine × off-serpentine pairs, we found a strong genome-wide signal of divergent selection against gene flow, with point estimates for σ of about 1.3×10^{-9} and 2.8×10^{-9} per base pair, i.e. 1.3×10^{-3} and 2.8×10^{-3} per megabase (Mb), in REM × SOD and SLP × TUL, respectively, and tight 95 % CIs (Fig. 3A, C; File S4.1). Given an assembled genome size of about 320 Mb for *M. guttatus*, this would for instance be consistent with about 300 MSPs, each with a selection coefficient of about 10^{-3} to 10^{-4}. The strong impact of this selection on genome-wide levels of polymorphism is visible from the red curves in Figs. 3B, D that represent the model fit. The 95 % CI of the relative difference (δ_π) between the maximum and minimum of this fitted curve for π_B clearly excludes 0 (Fig. 3B, D; SI Appendix B). According to our estimates of m, selection maintains this divergence against an average baseline level of gene flow of about 5.3×10^{-6} in REM × SOD and 8.1×10^{-6} in SLP × TUL (Fig. 3B, C). Given the estimated effective population sizes of REM and SLP (SI Text 1), this implies high rates of about 3.1 and 4.9 diploid immigrants per generation, respectively.

We had no power to infer precise point estimates for τ, but lower bounds of the 95 % CIs were around 10 Mya. Repeating our analyses for the (MS)M demographic scenario (Fig. S4.15), as well as when considering all non-focal populations jointly as the source of gene flow (Figs. S4.14, S4.16), we obtained very similar results to those above (File S4.2). Our inference about selection and gene flow therefore seems to be robust to the unknown specifics of demography.

To assess if the selection against gene flow we found is specific to serpentine × off-serpentine comparisons (REM × SOD, SLP × TUL), we also fit our model for the two long-distance off-serpentine × off-serpentine configurations (SOD × TUL, TUL × SOD) and the long-distance serpentine × off-serpentine pairs (REM × TUL, SLP × SOD). Interestingly, we inferred selection densities, durations of the MS phase, and migration rates on the same order as those estimated for the short-distance serpentine × off-serpentine comparisons (Fig. S4.13, File S4.1). The signal we detect may therefore have little to do with local adaptation to serpentine *per se*, and not be specific to the history of particular pairs of populations. Given the long time τ over which this selection appears to have acted, our estimates may reflect adaptive divergence in response to a range of locally varying conditions that *M. guttatus* experiences across its range [e.g. 27–29]. Our results could also imply that

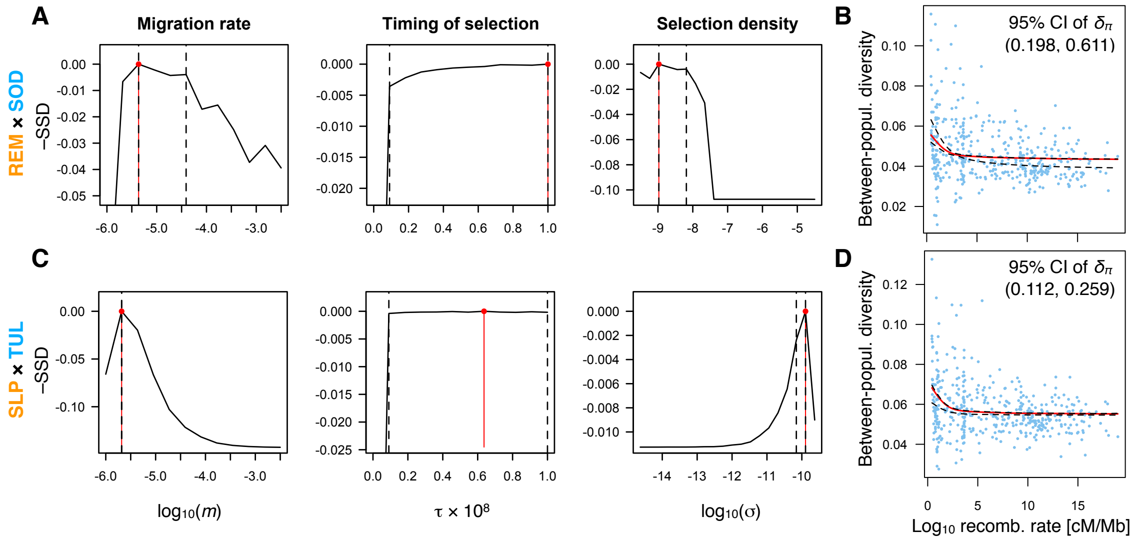

Figure 3. Parameter estimates and model fit for the serpentine dataset. (*A*, *C*) Profile curves of the quasi-likelihood (−SSD) for each parameter, maximising over the two remaining parameters, for the serpentine × off-serpentine comparisons REM × SOD (*A*) and SLP × TUL (*C*) (Fig. 2). Vertical red and black dashed lines indicate the point estimate and 95% bootstrap confidence intervals, respectively. (*B*, *D*) Raw data (blue dots) and model fit (red curve) with 95% confidence range (black dashed curves). The 95% confidence interval of the distribution of the relative difference between the maximum and minimum value of the model fit across all bootstrap samples, δ_π, is also given. Other details as in Fig. 2B–C. For other population pairs and the (MS)M scenario, see Figs. S4.13–S4.16 in SI Text 2.

adaptation to serpentine has a simple genetic basis, as our approach only detects a signal that is due to many genes under divergent selection across the entire genome.

Persistence of species barrier to *M. nasutus*

Where *M. guttatus* has come into secondary contact with *M. nasutus*, hybridisation occurs despite strong reproductive barriers [30]. A genome-wide analysis identified large genomic blocks of recent introgression from *M. nasutus* into *M. guttatus* [14]. The same study found that absolute divergence ($\pi_B = \pi_{\text{Gut}\times\text{Nas}}$) was negatively correlated with recombination rate (r_{bp}) in sympatric, but not allopatric Gut × Nas comparisons, consistent with selection against introgression. We reanalysed recombination data and whole-genome sequences from a single inbred individual per population from this study (the GutNas dataset; SI Text 1). While we replicate the negative, albeit weak, partial correlation between π_B and r_{bp} in sympatric comparisons (Fig. S4.10, SI Text 2), our estimates of σ were generally very low, except for the allopatric SLP × Nas pair (Fig. S4.17). Our model fit showed an increase of π_B at low values of r_{bp} in all comparisons, yet the 95 % CIs included the case of neutrality ($\delta_\pi = 0$), except for SLP × Nas (Fig. S4.17). After removal of genomic windows affected by recent introgression [14], estimates of σ were significantly different from neutrality in *both* allopatric pairs (AHQ × Nas: 1.7×10^{-4} per Mb; SLP × Nas: 9.4×10^{-4} per Mb), but remained non-significant in sympatric ones (CAC × Nas, DPR × Nas) (Figs. 4, S4.18).

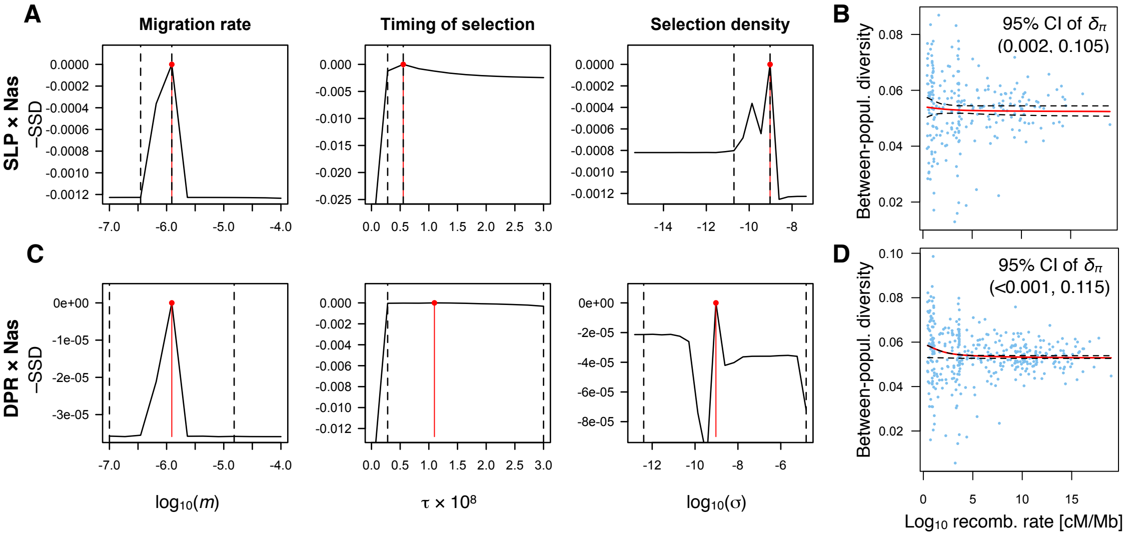

Figure 4. Parameter estimates and model fit for the Southern population pairs in the GutNas dataset. (*A*, *C*) Profile curves of the quasi-likelihood (−SSD) for each parameter, maximising over the two remaining parameters, for SLP × Nas (micro-allopatric) and DPR × Nas (sympatric), respectively. (*B*, *D*) Raw data, model fit with 95% confidence range, and the 95% confidence interval of the distribution of the relative difference δ_π between the maximum and minimum value of the model fit across bootstrap samples. Results are shown for the multi-MSP model under the (MS)P scenario after removal of blocks of recent introgression. Other details as in Fig. 3. For the Northern clade, see Fig. S4.18, and for results with blocks of recent introgression included, see Fig. S4.17.

With blocks of recent introgression excluded, we estimated m to be on the order of 10^{-6} with fairly tight 95 % CIs (Figs. 4, S4.18; File S4.2). Although small in absolute value, when scaled by the effective population size, these estimates imply an average net influx of about 0.1 to 1 diploid genomes per generation in the absence of selection. Lower confidence bounds for τ were consistently above 250 kya, with point estimates between about 550 kya (CAC × Nas, SLP × Nas) and 1.1 Mya (DPR × Nas), and no systematic difference between allopatric and sympatric comparisons (Figs. 4, S4.18). These estimates are somewhat above a previous estimate of about 196 kya for the onset of divergence between *M. guttatus* and *M. nasutus* [14]. Our older estimates of τ are compatible with divergent selection acting already in the ancestral, geographically structured, *M. guttatus* clade before speciation [14].

In summary, we found evidence for divergent selection maintaining a species barrier against gene flow over at least the last 250 to 500 ky. The current extent of range overlap with *M. nasutus* is not predictive of the strength of selection that we infer. While we do recover the previously reported negative correlation between π_B and r_{bp} for *sympatric* comparisons [14], our new method detects a signal of selection only for *allopatric* ones. This may be because we assume that gene flow has been continuous over the τ generations before the present, and that it has reached an equilibrium with selection. The many large blocks of recent introgression seen in sympatric comparisons [30] may violate this assumption and therefore lead to a poor fit by our model (Fig. S4.12; SI Text 2).

Discussion

The genomes of incompletely isolated species and locally adapted populations have long been thought of as mosaics of regions with high and low divergence [31, 32]. This pattern is due in part to varying levels of effective gene flow along the genome, created by an interaction of divergent selection and recombination-rate variation [33, 34]. The recent explosion of genome-wide DNA sequencing data allows us to directly observe this mosaic. It has spurred theoretical and empirical studies aiming to better understand the mechanisms underlying local adaptation and speciation [e.g. 35–38]. Yet, an explicit, model-based framework linking observed genome-wide patterns of divergence with the underlying mechanism has hitherto been missing.

Here, we developed such a framework by merging the concept of effective migration rate with coalescence theory. We showed that a genome-wide negative correlation of between-population diversity with recombination rate [14, 17] can be described by the compound parameter 'selection density', such that very different genomic mosaic patterns are compatible with the same aggregate effect of divergent selection and gene flow: a large number of weak genetic barriers to gene flow (MSPs) is equivalent to a much smaller number of strong barriers. Our approach complements existing genome scans for empirical outliers of population divergence [39–42], which tend to identify only strong barriers to gene flow. It also could provide a better null model for such genome scans, as outliers could be judged against the appropriate background level of divergence given local recombination rates.

Our approach is inspired by earlier work exploiting the *positive* genome-wide relationship between recombination rate and genetic diversity *within* a population for quantitative inference about genetic hitchhiking [43, 44] and background selection [45, 46]. While in *Mimulus* we did not observe such a correlation within populations, this earlier work would offer a natural way to correct for the potentially confounding effects of these types of selection at linked sites within our framework. Specifically, we could jointly fit a model of background selection or selective sweeps within populations [e.g. 47] and our existing model for divergent selection against gene flow.

We have assumed that MSPs occur at a constant rate ν along the genome. This could be improved by making ν depend on the functional annotation of genomes, e.g. exon coordinates, which might allow ν and s to be estimated separately [see 48]. Our model also does not account for the clustering of locally adaptive mutations arising in tight linkage to previously established MSPs, and the synergistic sheltering effect among MSPs that protects them from being swamped by gene flow [19, 49]. If accounted for, this clustering would lead to an even more pronounced uptick of between-population diversity in regions of low recombination. Therefore, one might be able to use deviations from our current model in regions of low recombination as a way of detecting the presence of clustering in empricial data. At the very least, our parameter estimates would offer a guide to whether and in what genomic regions one should expect clustering of MSPs to have evolved.

An inherent limitation of our approach is that enough time must have passed for between-population divergence to accumulate. Otherwise, there is no power to detect variation in divergence among genomic regions. This constrains the temporal resolution of our model, in particular if the duration of the migration–selection (MS) phase is short (e.g. if the divergence time is low), or if strong reproductive isolation evolved so quickly that gene flow was completely and rapidly reduced across the entire genome. Another potential limitation is a relatively low resolution to infer the duration of the MS phase. A genome-wide negative correlation of recombination rate with between-population diversity will persist for a long time even after gene flow has come to a complete halt, because subsequent neutral divergence will just add uniformly to the existing pattern. Our inference approach should therefore still provide good estimates of the strength of selection and gene flow even after speciation has completed, as long as these estimates are interpreted as averages over the inferred time τ. In this sense, our approach is likely robust to the specifics of the most recent demographic history of the populations or species of interest. To better resolve the timing of events, we suggest using the additional information contained in the entire distribution of pairwise coalescence times, rather than relying on their mean, as we currently do.

The opposing roles of gene flow and selection in speciation and local adaptation have a long and contentious history in evolutionary biology and population-genetics theory [3, 50]. We anticipate that the type of genome-wide quantitative inference developed here, applied to the growing amount of whole-genome polymorphism and recombination data, will help to resolve how gene flow is constraining divergent selection.

Materials and Methods

In Equation (3), $D = \mathrm{Ei}[(1-g_f)m\tau] - \mathrm{Ei}[(1-g_\circ)m\tau]$ and $F = \mathrm{Ei}[-g_\circ m\tau] - \mathrm{Ei}[-g_f m\tau] + \mathrm{Ei}[-\nu k_f] - \mathrm{Ei}[-\nu k_\circ]$, where $\mathrm{Ei}[z] = -\int_{-z}^{\infty} e^{-t}/t \,\mathrm{d}t$ is the exponential integral. Here, $g_f = [1 + s/r_f]^{-1}$ and $g_\circ = [1 + s/(k_\circ r_{\mathrm{bp}})]^{-1}$ are the contributions to the gff if the MSP is unlinked ($k_1 = r_f/r_{\mathrm{bp}}$) or fully linked ($k_1 = k_\circ$, $0 < k_\circ \lesssim 1/[r_{\mathrm{bp}}\tau]$); $k_\circ$ is a small positive lower limit for the physical distance to the MSP. For details of our model, theory, and individual-based simulations, see SI Appendix A. Statistical data analyses, bias corrections, and the inference procedure are described in detail in SI Appendix B. The *Mimulus* datasets (sampling design, DNA sequencing, quality filtering), and the linkage map are discussed in SI Text 1. For complementary results, including tests of partial correlation between diversity and recombination rate, see SI Text 2.

Acknowledgments

We thank Ben Blackman for help with collections, and Yaniv Brandvain, Lex Flagel, Amanda Kenney, Andrea Sweigart, Kevin Wright, Chenling Xu, and members of the Coop, Ross-Ibarra, and Schmitt labs at UC Davis for discussions and comments that improved our study. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under grant no. NIH RO1GM83098 and RO1GM107374 to GC, grant no. 1353380 from the U.S. National Science Foundation (NSF) to GC and JW, Ddig grant no. 1110753 from the NSF to JS, and an Advanced Postdoc.Mobility fellowship from the Swiss National Science Foundation (no. P300P3_154613) to SA. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

References

[1] Endler JA (1973) Gene flow and population differentiation. *Science* 179(4070):243–250.

[2] Mayr E (1996) What is a species, and what is not? *Philosophy of Science* 63(2):262–277.

[3] Coyne JA, Orr HA (2004) *Speciation*, vol 37 (Sinauer Associates Inc, Sunderland, MA), 1st edn.

[4] Payseur BA, Rieseberg LH (2016) A genomic perspective on hybridization and speciation. *Molecular Ecology* 25(11):2337–2360.

[5] Frantz LAF, Madsen O, Megens HJ, Groenen MAM, Lohse K (2014) Testing models of speciation from genome sequences: divergence and asymmetric admixture in island South-East Asian *Sus* species during the Plio-Pleistocene climatic fluctuations. *Molecular Ecology* 23(22):5566–5574.

[6] Filatov DA, Osborne OG, Papadopoulos AS (2016) Demographic history of speciation in a *Senecio* altitudinal hybrid zone on Mt. Etna. *Molecular Ecology* 25(11):2467–2481.

[7] Kousathanas A, Leuenberger C, Helfer J, Quinodoz M, Foll M, et al. (2016) Likelihood-free inference in high-dimensional models. *Genetics* 203(2):893–904.

[8] Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B: Biological Sciences* 263(1377):1619–1626.

[9] Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* 180(2):977–993.

[10] Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, et al. (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367(1587):343–353.

[11] Hermisson J (2009) Who believes in whole-genome scans for selection? *Heredity* 103(4):283–284.

[12] Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* 23:3133–3157.

[13] Keinan A, Reich D (2010) Human population differentiation is strongly correlated with local recombination rate. *PLoS Genetics* 6(3):e1000886.

[14] Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genetics* 10(6):e1004410 EP.

[15] Geraldes A, Basset P, Smith KL, Nachman MW (2011) Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. *Molecular Ecology* 20(22):4722–4736.

[16] Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution* 15(5):538–543.

[17] Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367(1587):409–421.

[18] Bengtsson BO (1985) The flow of genes through a genetic barrier. *Evolution – Essays in honour of John Maynard Smith*, eds Greenwood PJ, Harvey P, Slatkin M (Cambridge University Press, New York, NY), vol 1, chap 3, pp 31–42.

[19] Aeschbacher S, Bürger R (2014) The effect of linkage on establishment and survival of locally beneficial mutations. *Genetics* 197(1):317–336.

[20] Harrison S, Safford H, Wakabayashi J (2004) Does the age of exposure of serpentine explain variation in endemic plant diversity in California? *International Geology Review* 46(3):235–242.

[21] Notohara M (1990) The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology* 29(1):59–75.

[22] Wu CA, Lowry DB, Cooley AM, Wright KM, Lee YW, et al. (2008) *Mimulus* is an emerging model system for the integration of ecological and genomic studies. *Heredity* 100:220–230.

[23] Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.

[24] O'Dell RE, Rajakaruna N (2011) *Intraspecific variation, adaptation, and evolution* (University of California Press, Berkeley, CA), pp 97–137. Serpentine: A model of evolution and ecology.

[25] Selby J (2014) *The genetic basis of local adaptation to serpentine soils in* Mimulus guttatus. Doctoral dissertation, Duke University.

[26] Palm E, Brady K, Van Volkenburgh E (2012) Serpentine tolerance in *Mimulus guttatus* does not rely on exclusion of magnesium. *Functional Plant Biology* 39:679–688.

[27] Lowry DB, Hall MC, Salt DE, Willis JH (2009) Genetic and physiological basis of adaptive salt tolerance divergence between coastal and inland *Mimulus guttatus*. *New Phytologist* 183(3):776–788.

[28] Kooyers NJ, Greenlee AB, Colicchio JM, Oh M, Blackman BK (2015) Replicate altitudinal clines reveal that evolutionary flexibility underlies adaptation to drought stress in annual *Mimulus guttatus*. *New Phytologist* 206(1):152–165.

[29] Wright KM, Hellsten U, Xu C, Jeong AL, Sreedasyam A, et al. (2015) Adaptation to heavy-metal contaminated environments proceeds via selection on pre-existing genetic variation. *bioRxiv* .

[30] Kenney AM, Sweigart AL (2016) Reproductive isolation and introgression between sympatric *Mimulus* species. *Molecular Ecology* 25(11):2499–2517.

[31] Rieseberg LH, Whitton J, Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* 152(2):713–727.

[32] Harrison RG, Larson EL (2016) Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Molecular Ecology* 25(11):2454–2466.

[33] Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridising populations. *Heredity* 57(3):357–376.

[34] Harrison RG (1986) Pattern and process in a narrow hybrid zone. *Heredity* 56(3):337–349.

[35] Kronforst M, Salazar C, Linares M, Gilbert L (2007) No genomic mosaicism in a putative hybrid butterfly species. *Proceedings of the Royal Society of London Series B: Biological Sciences* 274:1255–1264.

[36] Bürger R, Akerman A (2011) The effects of linkage and gene flow on local adaptation: A two-locus continent–island model. *Theoretical Population Biology* 80(4):272–288.

[37] Via S, Conte G, Mason-Foley C, Mills K (2012) Localizing F_{ST} outliers on a QTL map reveals evidence for large genomic regions of reduced gene exchange during speciation-with-gene-flow. *Molecular Ecology* 21(22):5546–5560.

[38] Hemmer-Hansen J, Nielsen EE, Therkildsen NO, Taylor MI, Ogden R, et al. (2013) A genomic island linked to ecotype divergence in atlantic cod. *Molecular Ecology* 22(10):2653–2667.

[39] Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* 13(4):969–980.

[40] Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, et al. (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367(1587):364–373.

[41] Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology* 24(5):1031–1046.

[42] Haasl RJ, Payseur BA (2016) Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology* 25(1):5–23.

[43] Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356(6369):519–520.

[44] Wiehe TH, Stephan W (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Molecular Biology and Evolution* 10(4):842–854.

[45] Nordborg M, Charlesworth B, Charlesworth D (1996) The effect of recombination on background selection. *Genetics Research* 67(2):159–174.

[46] Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research* 70(2):155–174.

[47] Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, et al. (2016) A genomic map of the effects of linked selection in *Drosophila*. *PLoS Genetics* 12(8):e1006130.

[48] Jurić I, Aeschbacher S, Coop G (2016) The strength of selection against Neanderthal introgression. Submitted. bioRxiv preprint http://dx.doi.org/10.1101/030148.

[49] Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution* 65(7):1897–1911.

[50] Turelli M, Barton NH, Coyne JA (2001) Theory and speciation. *Trends in Ecology and Evolution* 16(7):330–343.