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

Simon Aeschbacher\textsuperscript{1,\textit{a}}, Jessica P. Selby\textsuperscript{2}, John H. Willis\textsuperscript{2}, and Graham Coop\textsuperscript{1}

14 September 2016

\textsuperscript{1}Department of Evolution and Ecology, University of California, Davis, CA 95616
\textsuperscript{2}Department of Biology, Duke University, Durham, NC 27708
\textit{a}saeschbacher@mac.com

Abstract

How strong is the natural selection that maintains species and locally adapted populations in the face of gene flow? To what extent is genomic divergence limited by gene flow? Here, we use DNA polymorphism data and the genome-wide variation in recombination rate to infer the strength and timing of selection, and the baseline level of gene flow under various demographic scenarios. To achieve this, we develop theory that merges the coalescent process with the concept of effective gene flow. The latter describes the reduction in gene flow at neutral loci due to divergent selection against maladapted immigrant alleles. This effect decreases with recombinational distance from the loci under selection, such that in regions of low recombination genetic divergence among populations is on average increased compared to regions of high recombination. Our inference procedure exploits this relationship in a genome-wide aggregate manner. We validate our approach using individual-based simulations and apply it to two datasets from the yellow monkeyflower (\textit{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. We show that the genome-wide intensity of this selection is not exceptional compared to what \textit{M. guttatus} may usually experience when adapting to local conditions. Second, we quantify and date selection against introgression from the selfing sister species \textit{M. nasutus}. Our study provides a theoretical framework that explicitly links genome-wide patterns of divergence and recombination with the underlying evolutionary mechanisms.

When and how strongly divergent selection acts against gene flow to maintain species and locally adapted populations is a long-standing question in biology \cite{1-3}. Answering it is fundamental to
understanding the evolution and persistence of organismal diversity, and to assessing the prevalence of different routes to speciation [4–6]. This requires quantitative estimates of the strength and timing of selection and gene flow. Population-genomic approaches, combined with a growing amount of DNA sequencing data, offer a powerful way to obtain such estimates and identify candidate speciation genes. The genome-wide perspective of these approaches is essential: genes underlying speciation and local adaptation act as barriers to gene flow, which translates into higher genetic divergence. But this is only detectable in contrast to genomic regions void of such barriers [7, 8].

In practice, 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 [8–10].

Existing population-genomic approaches to study adaptive genomic divergence in the face of gene flow come in three broad flavours. The first set of approaches apply methods for demographic inference to scenarios of speciation [11–17]. They inform our view of speciation by dating population splits and inferring the presence or absence of gene flow, yet generally do not explicitly account for natural selection [but see 18, 19].

A second set of approaches are genome scans for peaks of genetic divergence among populations or species. Such outliers are used to identify candidate loci underlying speciation or local adaptation [20–23]. This includes the search for so-called genomic islands of divergence [24–26], i.e. extended genomic regions of elevated divergence. Methods of this type can be confounded by other modes of selection and demography, and will always propose a biased subset of candidate loci [3, 27–30].

The third set are tests for a negative correlation between absolute genetic divergence and recombination rate across the genome [31–35]. This approach is based on the prediction that divergence is increased in regions of the genome where genetic linkage to loci under divergent selection is higher on average [31, 36, 37]. Such tests are in principle very powerful, as they aggregate information across the entire genome, and because the pattern of a negative correlation is highly specific to divergent selection with gene flow [38]. However, this approach is so far purely descriptive and much underexplored.

Here, we develop novel theory describing the pattern used by this latter type of approach, and an inference procedure based on this. Unlike the first two sets of 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 sequencing 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**

The key idea behind our approach is to exploit the genome-wide variation in recombination rate and its effect on genetic divergence between populations under selection against maladaptive gene flow. Recombination and genetic divergence are linked because divergent selection reduces effective gene flow at neutral sites, and this reduction depends on the recombinational distance from the loci under selection. In the following, 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. In the second part of the paper, we therefore devise an inference procedure and show its application. We start by describing our model.
We consider two populations (1, 2) of effective size \(N_1\) and \(N_2\), with diploid individuals and non-overlapping generations. A balance between one-way gene flow at rate \(m\) per generation from population 2 and local directional selection is maintained in population 1 for \(\tau\) generations before the present. We call this the migration–selection (MS) phase (Fig. 1A). Selection against maladaptive immigrant alleles occurs at an arbitrary number of biallelic loci that we refer to as 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 interactions and no dominance.

Prior to the MS phase, we model a panmictic (P) phase in an ancestral population of effective size \(N_0\) that starts \(\tau\) 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 \(\nu\) per base pair, such that the distance between consecutive MSPs is approximately exponentially distributed with mean \(1/\nu\) base pairs. Finally, we denote the per-base pair recombination rate by \(r_{\text{bp}}\). In summary, the key parameters are the mean selection coefficient \(s\), the genomic density \(\nu\) of MSPs, the baseline migration rate \(m\), and the duration \(\tau\) of the MS phase.

### 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 \(m_e\) by a so-called gene-flow factor, \(\text{gff} = m_e/m\). The reduction increases with the strength of selection at the MSPs, and decreases with their recombinational distance from the neutral site (Fig. 1A). Considering the nearest \(I\) up- and \(J\) downstream MSPs, Aeschbacher and Bürger showed that the effective migration rate at the neutral site can be approximated as \(m_e^{(I,J)} \approx m g^{(I)} h^{(J)}\), where

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

Here, \(k_i\) is the physical distance to, and \(a_i\) the selection coefficient at the \(i\)th upstream MSP (\(h^{(J)}\) is analogous, Eq. S1.1).

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 \(s\), \(\nu\), and \(r_{\text{bp}}\) exclusively through

\[
\eta = \frac{\sigma}{r_{\text{bp}}},
\]
where $\sigma = s\nu$ is the product of the mean selection coefficient times the density of MSPs. We call $\sigma$ the ‘selection density’ (per base pair). For instance, with $I = J = 2$ we find
\[
\mathbb{E}[m_e^{(2,2)}] \approx m(1 + 2\eta \ln \eta),
\]
which is a good approximation if $\eta \lesssim 0.1$, 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 (3) shows that the average level of effective gene flow decreases with selection density and increases with recombination rate. We found that adding increasing numbers of MSPs has a diminishing effect on $\mathbb{E}[m_e^{(I,J)}]$ (Fig. S1.2A), so that (3) captures the essential pattern if $\eta \lesssim 0.1$.

As exclusive dependence of $\mathbb{E}[m_e^{(I,J)}]$ on selection and recombination through the compound parameter $\eta$ holds for any $I$ and $J$, it also applies to the genome-wide average of $m_e$ (SI Appendix A, Eq. S1.8). Note that $\eta$ is the selection density per genetic map unit. Its emergence here implies 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 $\nu$ can be inferred only jointly as $\sigma$ 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. This accounts for genetic drift at the neutral site and makes a link to the genetic divergence $\pi_B$ between the two populations, which is readily estimated from DNA sequence data. 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 $m_e$ for $m$. Averaging over all numbers and locations of the MSPs, we obtain $\mathbb{E}[T_B]$, and can predict $\pi_B$ as $2u\mathbb{E}[T_B]$, 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 either for only the 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 $r_f$ as the recombination rate that corresponds to free recombination, such that MSPs located more than $k_f = r_f/r_{bp}$ base pairs from a neutral site are unlinked to it.

#### Selection at one locus

Under the one-MSP model, we implicitly assume that $\nu$ is small, i.e. $\nu \ll r_{bp}/s, m, \tau$, 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 $\nu$, the expected pairwise between-population coalescence time at an average neutral site is
\[
\mathbb{E}[T_B] \approx 2N_2 + \frac{1}{m} + \frac{2s\nu}{m r_{bp}} (e^{-m_0 \tau} D + F) + \frac{1}{m r_f} s e^{-2\nu k_f},
\]
where $D$ and $F$ depend on $m$, $\tau$, and $\nu$, (Materials and Methods). The first two terms in (4) are the expectation without selection (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 suggests that $\eta = \sigma / r_{bp}$ strongly determines $E[T_B]$, although $s, \nu$, and $r_{bp}$ also enter \[4\] independently. Indeed, given $r_{bp}$ and in the parameter range where \[4\] is a good approximation, the effect of selection on $E[T_B]$ is entirely captured by the selection density $\sigma$ (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 $E[T_B]$ is in principle similar to the one-MSP model 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 S1.7). As for the one-MSP model, if $r_{bp}$ is given, $E[T_B]$ depends on $s$ and $\nu$ effectively only through the selection density $\sigma$ (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 $\sigma$ as a key parameter and a natural metric to quantify genome-wide divergent selection in the face of gene flow.

**Application to *Mimulus guttatus***

Based on our theory above, we developed an inference procedure and applied it to two datasets from the predominantly outcrossing yellow monkeyflower (*Mimulus guttatus*), an important model system for speciation [33, 46–55] and local adaptation [56–61] native to western North America. In the first dataset, we detected a strong signal of local adaptation in the face of gene flow. In the second, we inferred the strength and timing of selection against introgression from the selfing sister species *M. nasutus* (SI Text 1).

We fitted our model to the empirical relationship between recombination rate ($r_{bp}$, estimated from a linkage map) and putatively neutral between-population diversity ($\pi_B$, 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 inference procedure computes the sum of squared deviations (SSD) across genomic windows between these estimates of $\pi_B$ and those predicted by our model given the estimate of $r_{bp}$ for each window and for a given set of parameter values. Minimising the SSD over a large grid of parameter values, we obtained point estimates for the selection density ($\sigma$), baseline migration rate ($m$), and duration of the MS phase ($\tau$). 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). In the following, we report results obtained under the multi-MSP model 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, so that background selection is unlikely to bias our findings [62, 64].
Adaptive divergence maintained in the face of gene flow

Like many species across different plant taxa [65–67], M. guttatus has locally adapted to serpentine outcrops occurring throughout its range [68, p. 4]. Serpentine soils are characterised by high concentrations of heavy metals and a low calcium-to-magnesium ratio [69, 70], conditions that are toxic to plants [71, 72]. While the mechanism and molecular basis of serpentine tolerance in M. guttatus are unresolved [73], strong differences in survival on serpentine soil exist between serpentine and non-serpentine M. guttatus ecotypes, both in the greenhouse and in nature [68]. To investigate a population-genomic signal of local adaptation to serpentine soil, we sequenced the whole genomes of 324 M. guttatus individuals collected from two pairs of geographically close populations growing on and off serpentine soil in California in a pooled approach (the serpentine dataset; Fig. 2A, SI Text 1). We started by estimating 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 off-serpentine population 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 $-\text{SSD}$ (holding m and $\tau$ at their point estimates) showed a pronounced ridge for $s$ and $\nu$, 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 $\nu$ can only be estimated jointly as their product, the selection density $\sigma$. We therefore adjusted our inference procedure to jointly infer $m$, $\tau$, and $\sigma$ instead of $m$, $\tau$, $s$, and $\nu$ (SI Appendix B). This resulted in conditional $-\text{SSD}$ surfaces for $\sigma$ and $m$ with a unique peak and tight confidence hulls (Fig. 2C).

For both serpentine $\times$ off-serpentine pairs, we found a strong genome-wide signal of divergent selection against gene flow, with point estimates for $\sigma$ of about $10^{-3}$ and $10^{-4}$ per megabase (Mb) in REM $\times$ SOD and SLP $\times$ TUL, respectively, and tight 95 % CIs (Fig. 3A, C; SI File S4.2). 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 ($\delta_\pi$) between the maximum and minimum of this fitted curve for $\pi_B$ 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 $10^{-5.4}$ in REM $\times$ SOD and $10^{-5.7}$ in SLP $\times$ TUL (Fig. 3B, C). Given the estimated effective population sizes of REM and SLP (SI Text 1), this implies high rates of about 2.5 and 1.3 diploid immigrants per generation, respectively.

We had no power to infer precise point estimates for $\tau$, 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 source of gene flow (Figs. S4.14, S4.16), we obtained very similar results to those above. Our inference about selection and gene flow therefore seems to be robust to the unknown specifics of the demography.

To assess if the selection against gene flow we found is specific to serpentine $\times$ off-serpentine comparisons (REM $\times$ SOD, SLP $\times$ TUL), we also fitted our model for the two long-distance off-serpentine $\times$ off-serpentine configurations (SOD $\times$ TUL, TUL $\times$ SOD) as well as the long-distance serpentine $\times$ off-serpentine pairs (REM $\times$ TUL, SLP $\times$ SOD). Interestingly, we inferred selection densities, durations of the migration-selection regime, and migration rates on the same order as those estimated for the short-distance serpentine $\times$ off-serpentine comparisons (Fig. S4.13, SI File S4.2). This suggests that the signal we detect may have little to do with local adaptation to
serpentine per se, and may not be specific to the history of particular pairs of populations. Rather, given the long time $\tau$ that this selection appears to have acted over, we view our estimates as reflecting adaptive divergence in response to a range of locally varying conditions that *M. guttatus* experiences across its entire range [56–61].

### Persistence of species barrier to *M. nasutus*

Where *M. guttatus* has come into secondary contact with its selfing sister species *M. nasutus*, hybridisation occurs [33, 55] despite strong reproductive barriers [46–50, 52–55]. A genome-wide analysis found large genomic blocks of recent introgression from *M. nasutus* into *M. guttatus* [33, 55]. In the same study, absolute divergence ($\pi = \pi_{Gut \times Nas}$) was found to be negatively correlated with recombination rate ($r_{bp}$) in sympatric, but not allopatric Gut $\times$ Nas comparisons, consistent with selection against this introgression [33]. We reanalysed whole-genome sequences from single individuals and recombination data from this study (the GutNas dataset) with our new method. While we replicate the negative, albeit weak, partial correlation between $\pi_B$ and $r_{bp}$ in sympatric comparisons (Fig. S4.10, SI Text 2), our estimates of $\sigma$ were generally very low, except for the allopatric SLP $\times$ Nas pair (Fig. S4.17). Our model fit showed an uptick of $\pi_B$ at low values of $r_{bp}$ in all comparisons, yet the 95% CIs included the case of neutrality ($\delta = 0$), except for SLP $\times$ Nas (S4.17). After removal of genomic windows that fall inside the blocks of recent introgression [33], our estimates of $\sigma$ were significantly different from neutrality in both allopatric pairs (AHQ $\times$ Nas: $10^{-3.8}$ per Mb; SLP $\times$ Nas: $10^{-3}$ per Mb), but remained non-significant in sympatric ones (CAC $\times$ Nas, DPR $\times$ Nas) (Figs. S4.18).

With blocks of recent introgression excluded, we estimated $m$ to be on the order of $10^{-6}$ with fairly tight 95% CIs (Figs. S4.18). Although small in absolute value, when scaled by the effective population size, these estimates would imply an average net influx of about 0.1 to 1 diploid genomes per generation in the absence of selection. Lower confidence bounds for $\tau$ were consistently above 250 kya, with point estimates between about 500 kya (SLP $\times$ Nas) and 1.1 Mya (DPR $\times$ Nas), and no systematic difference between allopatric and sympatric comparisons (Figs. 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* [33]. Our older estimates of $\tau$ are compatible with divergent selection acting already in the ancestral, geographically structured, *M. guttatus* clade before speciation [33].

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. Our results also suggest that a signal of selection against historical gene flow [74] might be partially masked by blocks of recent introgression that have not yet converged to the pattern predicted 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 [55, 75–79]. 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 [39–41, 80, 81]. This idea has shaped the modern view of local adaptation and speciation since its inception [7, 82]. The recent explosion of genome-
Wide DNA sequencing data has created the opportunity to directly observe this mosaic. It has spurred theoretical and empirical studies aiming to better understand the mechanisms underlying local adaptation and speciation [41, 83–87], as well as identify the genetic architecture of the traits involved [88–96]. Yet, an explicit, model-based framework linking observed genome-wide patterns of divergence with the underlying mechanism has hitherto been missing [10, 97].

Here, we developed such a framework by merging the concept of effective migration rate with coalescence theory. One key insight that emerges from this work is that the empirical pattern of a genome-wide negative relationship of between-population diversity with recombination rate [e.g. 33, 38] can be described by the compound parameter ‘selection density’. Its implication is that, at a genome-wide level, very different mosaic patterns lead to the same average outcome: a large number of weak genetic barriers to gene flow (MSPs) is equivalent to a much smaller number of strong barriers. Our approach therefore complements existing genome scans for empirical outliers of population divergence that can only hope to identify loci that are strong barriers to gene flow [29, 98–100]. It also could provide a better null model for such genome scans, as outliers could be judged against an appropriate background level of divergence for their local recombination rate.

The idea behind 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 [101, 102] and background selection [36, 64]. While in Mimulus we did not observe a correlation between recombination rate and genetic diversity 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. 103] and our existing model for divergent selection against gene flow.

We have assumed that the genetic barriers to gene flow occur at a constant rate (ν) along the genome. We could improve on this 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 104]. Our model also does not account for the clustering of locally adaptive mutations arising in tight linkage to previously established MSPs [43, 105], and the synergistic sheltering effect among MSPs that helps them against being swamped by gene flow [106]. 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 empirical data. Or, at the very least, our parameter estimates would offer a guide to whether we should see such an effect, and in which genomic regions clustering might be expected to have evolved.

An inherent limitation to our approach is that the populations or species must have had sufficient time for neutral divergence to accumulate. Otherwise, there is no power to detect variation in divergence among regions. This constrains the temporal resolution of our model, in particular if the duration of the migration–selection (MS) phase is short, which would be the case 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, as subsequent neutral divergence should then 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 contrasting roles of gene flow and selection in speciation and local adaptation have a long and contentious history in evolutionary biology and population-genetics theory [4, 107]. In our view, a promising way forward is to provide quantitative, genome-wide statements about the strength and timing of selection and gene flow, so that empirical observations can be placed on a firm theoretical footing. We anticipate that the type of approach developed here, applied to the growing number of datasets with genome-wide polymorphism and recombination data, may help to resolve the role of gene flow in constraining divergent selection.

Materials and Methods

The terms $D$ and $F$ in (4) are $D = \text{Ei}[(1 - g_f)m\tau] - \text{Ei}[(1 - g_o)m\tau]$ and $F = \text{Ei}[-g_o m\tau] - \text{Ei}[-g_f m\tau] + \text{Ei}[-\nu k_f] - \text{Ei}[-\nu k_o]$, where $\text{Ei}[z] = - \int_{-\infty}^{z} e^{-t}/t \, dt$ is the exponential integral. Here, $g_f = [1 + s/r_f]^{-1}$ and $g_o = [1 + s/(k_o r_{bp})]^{-1}$ are the contributions to the gff if the MSP is unlinked ($k_1 = r_f/r_{bp}$) or if it is maximally linked ($k_1 = k_o$, $0 < k_o \lesssim 1/r_{bp}\tau$), respectively, and $k_o$ is a small positive lower limit for the physical distance to the MSP.

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


Figure 1. Divergent selection reduces gene flow and increases genetic differentiation between populations. (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. Multiple realisations of the coalescent process are shown in the bottom left for the (MS)P scenario (Fig. S.1.1). 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) Comparison of the predicted between-population diversity \( \pi_B = 2u\mathbb{E}[T_B] \) (curves) with individual-based simulations (dots) as a function of the recombination rate \( r_{bp} \) for different selection coefficients \( s \). 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 \( \pi_B \) as a function of \( \log_{10}(s) \) and \( \log_{10}(\nu) \) support the compound parameter selection density, \( \sigma = s\nu \). Here, \( r_{bp} = 10^{-8} \) (1 cM/Mb), \( r_T = 0.5 \), and the other parameters are as in (B).
Figure 2. Geographic context of serpentine dataset and quasi-likelihood surfaces for population pair SLP × TUL. (A) Map of sampling sites in California, USA (modified with permission [44]). (B) Surface of the negative sum of squared deviations (−SSD) for the selection coefficient s and the genomic density ν of MSPs, conditional on point estimates of m and τ. The ridge with slope −1 confirms the compound parameter selection density, σ = sν. The cross indicates the point estimate and black hulls comprise the area of 95% bootstrap confidence. (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 SLP × TUL pair under the multi-MSP model and the (MS)P scenario, and with genomic windows of size 500 kb.
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, δπ, 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.
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 $\delta_\pi$ 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.