bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Simon Aeschbacher[1,2,3,a], Jessica P. Selby[4], John H. Willis[4], and Graham Coop[1]

7 June 2017

[1]Department of Evolution and Ecology, University of California, Davis, CA 95616
[2]Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland
[3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[4]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 genome-wide negative correlation between recombination rate and genetic differentiation 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 that accounts for the confounding effect of background selection. Applying this method to two datasets from *Mimulus guttatus*, 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. We also 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 at least the last 250 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.

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 purely descriptive and problematic if recombination covaries with a confounding factor, e.g. gene density, that in turn affects the intensity of selection [18].

Here, we develop theory describing the pattern used by this third approach, and a way of inferring the underlying parameters. Our approach explicitly models selection against gene flow and its effect on neutral variation, estimates the strength and timing of selection and gene flow, and filters out the confounding effect of background selection.

Idea of Approach and Population-Genomic Model

Here, we exploit 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 conceptualize this relationship in terms of the effective migration rate and the expected pairwise between-population coalescence time (Fig. 1A). The latter directly connects 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, one allele is favored in population 1 over the other by an average selection coefficient s, while the deleterious allele 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. Here, we use the (MS)P and (MS)M scenarios to describe our approach. In SI Appendix A we provide extensions to more general scenarios with an intermediate *isolation* phase.

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

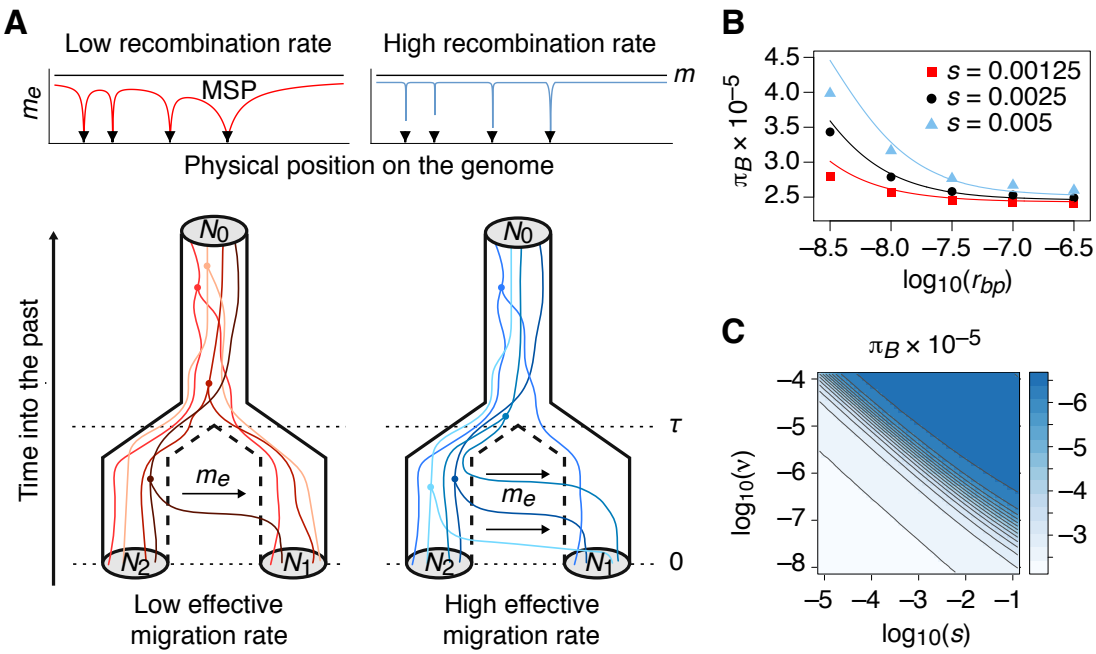

Figure 1. Divergent selection reduces gene flow and increases genetic divergence. (*A*) Selection against locally maladapted alleles at migration–selection polymorphisms (MSPs; black triangles) reduces the effective migration rate m_e. The effect is stronger in regions of low recombination (red, top left) and decreases the probability that lineages sampled in different populations migrate and coalesce. Realizations 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 higher (blue, top right), such that migration and earlier coalescences 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 scenario was used with $N_2 = 5000$, $u = 10^{-9}$, $\nu = 2.5 \times 10^{-7}$, $m = m_0 = 5 \times 10^{-4}$, and $\tau = 4N_2$. (*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, $r_{\mathrm{bp}} = 10^{-8}$ (1 cM/Mb); other parameters as in (*B*).

We denote the per-base pair recombination rate by r_{bp} and assume that the MSPs occur at a constant rate ν per base pair, such that the distance between consecutive MSPs is exponentially distributed with mean $1/\nu$ base pairs.

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 [19, 20]. This 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; Eq. S1.1 in SI Appendix A). To extrapolate from a given neutral site to the entire genome, we need to average over the possible genomic locations and selection coefficients of the MSPs. For simplicity, we assume an infinite chromosome with a linear relationship between physical and genetic map distance. Given an exponential distribution of selection coefficients, the expected effective migration rate 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 (SI Appendix A). Note that σ/r_{bp} has the meaning of a selection density

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

per genetic map unit. For instance, conditioning on two MSPs on each side of an average neutral site, we find

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

This equation is a good approximation if $\sigma/r_{\mathrm{bp}} \lesssim 0.1$, i.e. if recombination is at least ten times stronger than selection, at which point effective gene flow is reduced by about 50% (Fig. S1.2 in SI Appendix A). Equation (1) 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 $\mathbb{E}[m_e]$ (Fig. S1.2A), so that Eq. (1) captures the essential pattern if $\sigma/r_{\mathrm{bp}} \lesssim 0.1$. The exclusive dependence of $\mathbb{E}[m_e]$ on selection and recombination through the compound parameter σ/r_{bp} holds for any number of MSPs and so applies to the genome-wide average of m_e (SI Appendix A, Eq. S1.8). 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 facilitate parameter inference from population-genomic data, we phrase our theory in terms of the expected coalescence time of two lineages, one from each population. The expectation of this coalescence time under neutrality, T_B, depends on the baseline migration rate m (Table S1.2 in SI Appendix A). We incorporate the effect of selection by substituting the effective migration rate for m. Averaging over all possible numbers and genomic 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.

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 $k_f = 1/(2r_{\mathrm{bp}})$ base pairs from a neutral site are unlinked. We start by assuming that ν is so small that at most a single, nearest-neighboring MSP is linked to any focal neutral site. In the simplest case of the (MS)M scenario with $m_0 = m$ (Fig. S1.1C in SI Appendix A), the expected pairwise between-population coalescence time is approximately

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

where D and F depend on m, τ, and ν (Materials and Methods). The first two terms in Eq. (2) 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 and unlinked 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 Eq. (2) independently. Indeed, given r_{bp} and in the parameter range where Eq. (2) is a good approximation (i.e. for $\nu \ll r_{\mathrm{bp}}/s, m, \tau$), the effect of selection on $\mathbb{E}[T_B]$ is entirely captured by σ (Fig. S1.4 in SI Appendix A). For details and other demographic scenarios, see SI Appendix A. In this simplified model, the effect of all other MSPs is absorbed by m as a genome-wide reduction in gene flow that is independent of recombination.

In practice, we want to explicitly account for all MSPs possibly present in the genome, as well as for the average physical chromosome length. Finding $\mathbb{E}[T_B]$ in this more realistic setting amounts to averaging over all possible numbers and genomic locations of the MSPs. We wrote a C++ program to do this integration numerically (SI Appendix A). The result agrees well with individual-based forward simulations (Fig. 1B; Fig. S1.5 in SI Appendix A). As with a single MSP, 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, returning to the idealizing assumption of a global linear relationship between physical and genetic map distance (i.e. $r_f \to \infty$), we show that this dependence holds exactly (SI Appendix A, Eq. S1.68). This finding

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

corroborates σ as a key parameter and 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 fit the model with multiple MSPs 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. Minimizing 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 explored two alternative demographic scenarios, but focus here on the one that provided more plausible parameter estimates and tighter 95% CIs. Unless otherwise stated, we only report results obtained with genomic windows of 500 kb because results for windows of 100 and 1000 kb were very similar (SI Text 2).

Accounting for background selection

In *M. guttatus*, pericentromeric regions are gene-poor and characterized by low recombination rates, which results in a genome-wide positive correlation between gene density and recombination rate [14] (Fig. S4.4 in SI Text 2). Such a correlation could attenuate or even reverse the positive correlation between diversity and recombination rate otherwise expected under background selection (BGS) and other modes of selection at linked sites, because the strength of such selection is in turn expected to be positively correlated with gene density [18]. This effect could create a false signal of selection against gene flow, as the increased coalescence rate within populations in gene-dense regions could produce a negative correlation between r_{bp} and π_B. Comparing tests of a partial correlation between r_{bp} and diversity with and without gene density as a covariate, we found that this effect might be present in our first dataset (SI Text 2; File S4.1). Therefore, we first fit a BGS model to the genetic diversity *within* source populations by allowing the effective population size (N_2) to vary as a function of gene density and r_{bp}. We then incorporated BGS into our migration-selection inference procedure, using these predicted relationships between N_2, gene density, and r_{bp} (SI Appendix B). This procedure filters out the effect of BGS since, with unidirectional gene flow, BGS in the *source* but not the *focal* population may affect π_B (Eq. 2).

Adaptive divergence maintained in the face of gene flow

Mimulus guttatus can be found growing on serpentine soils 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 see whether there was 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 proximate off-serpentine population (SOD and TUL, respectively),

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

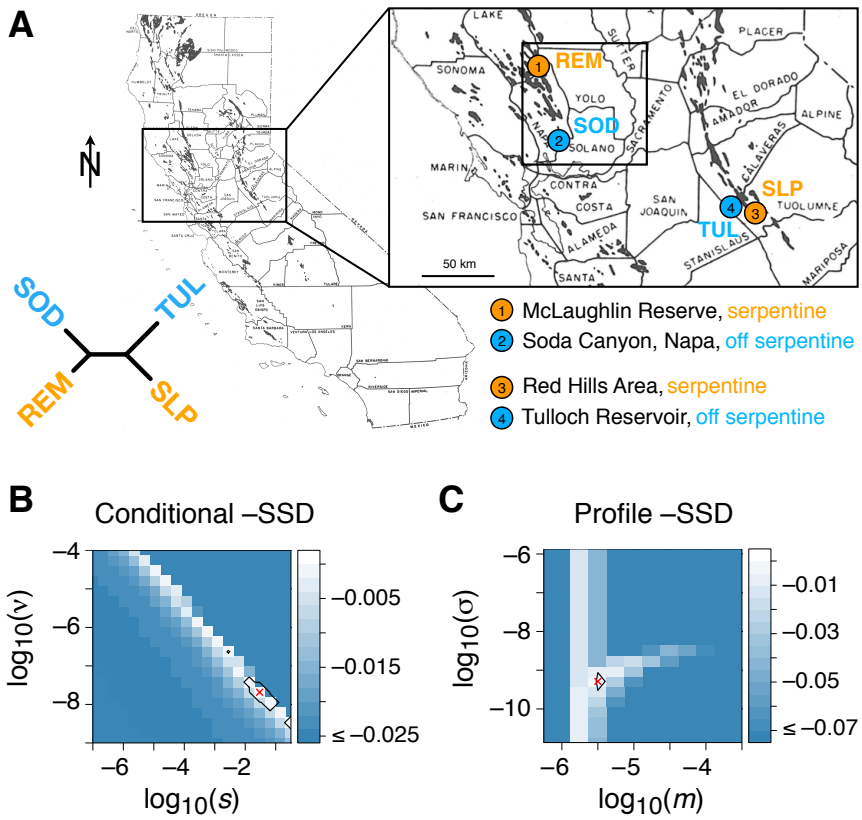

Figure 2. Geographic context of serpentine dataset and quasi-likelihood surfaces. (*A*) Sampling sites in California, USA (modified with permission from [23]), and unrooted population phylogeny based on linearized genetic divergence [24]. (*B*) The negative sum of squared deviations ($-$SSD) for the 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, maximized over τ. Results are shown for the population pair REM $\times$ SOD under the (MS)M scenario with genomic windows of 500 kb.

using the latter in each pair as a proxy for the source of gene flow. These pairs of geographically close serpentine $\times$ off-serpentine populations are genetically less diverged than any other population pair (Fig. 1A; Fig. S4.1 in SI Text 2). Since we observed a strong signal of BGS in all populations (SI Text 2), we corrected for this signal when fitting our migration–selection model to the data (SI Appendix B). 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 should 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 adjustment resulted in profile $-$SSD surfaces for σ 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 σ of about 8.3×10^{-4} and 4.8×10^{-4} per megabase (Mb) in REM $\times$ SOD and SLP $\times$ TUL, respectively, and tight 95% CIs (Fig. 3A, C; File S4.2). Given an assembled genome size of about 320 Mb for *M. guttatus*, this selection density would,

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

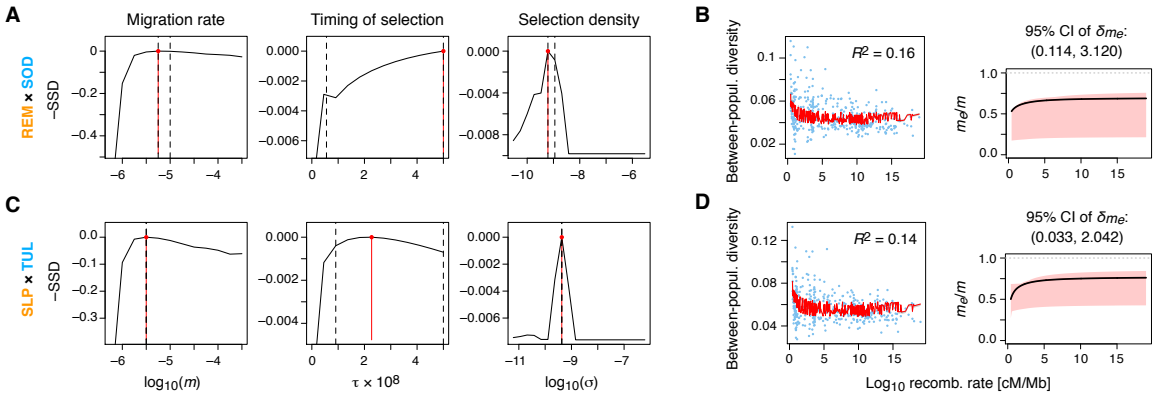

Figure 3. Parameter estimates and model fit for the serpentine dataset. (*A*, *C*) Profile curves of the quasi-likelihood (−SSD) for each parameter, maximizing over the two remaining parameters, for the serpentine × off-serpentine comparisons REM × SOD (*A*) and SLP × TUL (*C*) (Fig. 2). Vertical red and black dashed lines indicate the point estimate and 95% bootstrap confidence intervals (CIs), respectively. (*B*, *D*) Raw data (blue dots) and model fit (red curve) with 95% CI (gray shading). The corresponding ratio of the effective to the baseline migration rate is shown on the right (red shading: 95% CI). The 95% CI of the distribution of the relative difference between the maximum and minimum m_e across all bootstrap samples, δ_{m_e}, is given on top. Other details as in Fig. 2B–C. For other population pairs, see Fig. S4.17 in SI Text 2.

for instance, be consistent with about 300 MSPs, each with a selection coefficient on the order of 10^{-4} to 10^{-3}. The impact of this selection on genome-wide levels of polymorphism is reflected in an increase in the effective migration rate (m_e) with higher recombination rate (Fig. 3B, D). The 95% CI of the relative difference (δ_{m_e}) between the maximum and minimum of m_e/m clearly excludes 0 (Fig. 3B, D; SI Appendix B), indicating a partial shutdown of gene flow due to selection. According to our estimates of m, selection maintains this divergence against a baseline rate of gene flow of about 6.6×10^{-6} in REM × SOD and 3.5×10^{-6} in SLP × TUL (Fig. 3B, C). Given the estimated effective population sizes of REM and SLP (SI Text 1), these rates of gene flow imply about 3.8 and 2.1 diploid immigrants per generation, respectively.

We had little power to infer precise point estimates for τ, but lower bounds of the 95% CIs were around 10 Mya. It seems unlikely that the two ecotypes persisted for so long and so our parameter estimates should be interpreted as a long-term average over a potentially more complex scenario.

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. S4.17 in SI Text 2; File S4.2). 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. This is corroborated by the fact that, when pooling all non-focal populations to a joint source of gene flow, we observed a similar, if not even stronger, signal of selection against migrants (Fig. S4.18 in SI Text 2; File S4.2). Given the long time τ over which this selection appears to have acted, our estimates may reflect adaptive divergence between *M. guttatus* populations in response to locally varying conditions other than serpentine soil [e.g. 27–29]. Our results could also imply that adaptation to serpentine has a simple genetic basis, as our approach only has power to detect a signal that is due to polygenic divergent selection acting across the entire genome.

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Persistence of species barrier to *M. nasutus*

Where *M. guttatus* has come into secondary contact with *M. nasutus*, a selfing sister species, hybridization occurs despite strong reproductive barriers [30]. A previous genome-wide analysis identified large genomic blocks of recent introgression from *M. nasutus* into sympatric *M. guttatus* populations [14]. Using 100-kb genomic windows, this previous study also found a negative correlation between absolute divergence ($\pi_B = \pi_{\mathrm{Gut}\times\mathrm{Nas}}$) and recombination rate (r_{bp}) in sympatric but not allopatric comparisons, as would be expected if there was selection against hybrids. Reanalyzing these data (the GutNas dataset; SI Text 1) we replicate this pattern of correlation. However, if we included gene density as a covariate, all previously negative partial correlations between r_{bp} and π_B became non-significant (Fig. S4.12A in SI Text 2). This observation might indicate that the positive correlation between recombination and gene density could have camouflaged an underlying signal of BGS in the source population, as was the case with the serpentine dataset above. To test this hypothesis, we fit a model of BGS in *M. nasutus*, but found no evidence for BGS (SI Text 2) [cf. 14].

Applying our method to 100-kb windows, we indeed found a significant signal of selection against hybrids for one sympatric pair (CAC $\times$ Nas; $\hat{\sigma} \approx 7.4 \times 10^{-4}$/Mb), yet no signal in the other (DPR $\times$ Nas). We also inferred significant selection against gene flow in one of the allopatric comparisons (SLP $\times$ Nas; $\hat{\sigma} \approx 1.3 \times 10^{-4}$/Mb). The signal of selection in SLP $\times$ Nas could be due to the fact that, although allopatric, SLP is geographically close to *M. nasutus* populations [14]. We might therefore be detecting selection against ongoing gene flow over a longer distance, or against past gene flow that has stopped only recently. Since levels of recent introgression are much lower in SLP than in the sympatric populations (AHQ and CAC) [14], the second explanation is more plausible. Indeed, repeating our analyses with blocks of recent introgression excluded, we found that the signal of selection against hybrids remained for SLP $\times$ Nas, but disappeared for sympatric comparisons (Fig. S4.12B; Fig. S.4.21 in SI Text 2).

Our estimates of m for CAC and SLP imply that selection maintains the species barrier against a baseline migration rate of about 10^{-6}, i.e. 1.0 and 0.7 diploid introgressing genomes per generation, respectively (Fig. S4.19 in SI Text 2; File S4.3). With blocks of recent introgression excluded, our point estimate of m obtained with 100-kb windows dropped by a factor of 2.8 for CAC $\times$ Nas (File S4.3), consistent with the removal of a substantial part of recently introgressed DNA. In contrast to the serpentine dataset, our results for the GutNas dataset were sensitive to the choice of window size. For windows of 500 and 1000 kb, the uncertainty in parameter estimates was higher (Fig. S4.20, S4.22 in SI Text 2; File S4.3).

With 100-kb windows and blocks of recent introgression included, lower 95%-CI limits for τ were all above 250 kya. Point estimates were between about 550 kya (AHQ $\times$ Nas) and 1.6 Mya (DPR $\times$ Nas) (File S4.3). These estimates are somewhat above a previous estimate of 196 kya for the divergence time 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 [14].

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 variation in effective gene flow along the genome, created by an interaction of divergent selection and recombination [33, 34]. The recent explosion of genome-wide DNA sequencing data allows us to directly observe this mosaic, and has spurred theoretical and empirical studies aiming to better understand the mechanisms underlying local adaptation and speciation [e.g. 35–38]. Yet, an explicit,

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 is most sensitive to polygenic selection, and therefore complements existing genome scans for empirical outliers of population divergence [39–42], which tend to identify only strong barriers to gene flow. It also provides a better null model for such genome scans, as outliers could be judged against the appropriate background level of divergence.

Our approach is inspired by earlier work exploiting the genome-wide relationship between recombination rate and genetic diversity within a population for quantitative inference about genetic hitchhiking [43, 44] and background selection (BGS) [45, 46]. In fact, we have used an established model of BGS [47, 48] to filter out any confounding effect of BGS and gene density in a first step, before fitting our model of divergent selection against gene flow to the relationship between recombination rate and diversity between populations.

We have assumed that MSPs occur at a constant rate ν along the genome. This assumption could be relaxed 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 49]. We explored this solution heuristically for the serpentine dataset by setting ν proportional to the local density of exonic sites. Point estimates of σ were on the same order as before (10^{-10} to $10^{-8.25}$). Yet, the 95% CIs became much wider and the variance explained (R^2) dropped from about 15 to 5%. This reduced goodness of fit might suggest that selection against gene flow is not acting exclusively on coding (exonic) variation.

Our model currently 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 [20, 50]. 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. At the very least, our parameter estimates would indicate 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 limitation constrains the temporal resolution of our method, in particular if the duration of the migration–selection (MS) phase is short, 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 (and pairwise sequence differences), 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, 51]. We antic-

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ipate 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 (2), $D = \mathrm{Ei}[(1-g_f)m\tau] - \mathrm{Ei}[(1-g_\circ)m\tau]$ 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 as well as the inference of background selection, 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 G.C., Ross-Ibarra, and Schmitt laboratories 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 grants NIH RO1GM83098 and RO1GM107374 (to G.C.); National Science Foundation (NSF) Grant 1354688 (to J.H.W.); NSF Grant 1353380 (to G.C.); NSF DDIG Grant 1110753 (to J.P.S.); and Swiss National Science Foundation Advanced Postdoc.Mobility Fellowship P300P3_154613 (to S.A.). The computational results presented have been achieved in part using the Vienna Scientific Cluster.

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*. (Sinauer Associates Inc, Sunderland, MA) Vol. 37, 1st edition.

[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, et al. (2016) Likelihood-free inference in high-dimensional models. *Genetics* 203(2):893–904.

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[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, 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] Slotte T (2014) The impact of linked selection on plant genomic variation. *Briefings in Functional Genomics* 13(4):268–275.

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

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

[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, et al. (2008) *Mimulus* is an emerging model system for the integration of ecological and genomic studies. *Heredity* 100:220–230.

[23] 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.

[24] Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139(1):457–462.

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

[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, et al. (2015) Adaptation to heavy-metal contaminated environments proceeds via selection on pre-existing genetic variation. *bioRxiv*. 10.1101/029900.

[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, 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, 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.

bioRxiv preprint doi: https://doi.org/10.1101/072736; this version posted June 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[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] Hudson RR, Kaplan NL (1995) Deleterious background selection with recombination. *Genetics* 141(4):1605–17.

[48] Corbett-Detig RB, Hartl DL, Sackton TB (2015) Natural selection constrains neutral diversity across a wide range of species. *PLoS Biology* 13(4):e1002112.

[49] Jurić I, Aeschbacher S, Coop G (2016) The strength of selection against Neanderthal introgression. *PLoS Genetics* 12(11):e1006340.

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

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