Title

Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data

Journal

Genetics, 207(4)

ISSN

0016-6731

Authors

Lee, Kristin M

Coop, Graham

Publication Date

2017-12-01

DOI

10.1534/genetics.117.300417

Peer reviewed

Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data

Kristin M. Lee*,†,1 and Graham Coop*,†
*Center for Population Biology and †Department of Evolution and Ecology, University of California, Davis, California 95616
ORCID IDs: 0000-0003-1748-4948 (K.M.L.); 0000-0001-8431-0302 (G.C.)

ABSTRACT Geographically separated populations can convergently adapt to the same selection pressure. Convergent evolution at the level of a gene may arise via three distinct modes. The selected alleles can (1) have multiple independent mutational origins, (2) be shared due to shared ancestral standing variation, or (3) spread throughout subpopulations via gene flow. We present a model-based, statistical approach that utilizes genomic data to detect cases of convergent adaptation at the genetic level, identify the loci involved and distinguish among these modes. To understand the impact of convergent positive selection on neutral diversity at linked loci, we make use of the fact that hitchhiking can be modeled as an increase in the variance in neutral allele frequencies around a selected site within a population. We build on coalescent theory to show how shared hitchhiking events between subpopulations act to increase covariance in allele frequencies between subpopulations at loci near the selected site, and extend this theory under different models of migration and selection on the same standing variation. We incorporate this hitchhiking effect into a multivariate normal model of allele frequencies that also accounts for population structure. Based on this theory, we present a composite-likelihood-based approach that utilizes genomic data to identify loci involved in convergence, and distinguishes among alternate modes of convergent adaptation. We illustrate our method on genome-wide polymorphism data from two distinct cases of convergent adaptation. First, we investigate the adaptation for copper toxicity tolerance in two populations of the common yellow monkey flower, *Mimulus guttatus*. We show that selection has occurred on an allele that has been standing in these populations prior to the onset of copper mining in this region. Lastly, we apply our method to data from four populations of the killifish, *Fundulus heteroclitus*, that show very rapid convergent adaptation for tolerance to industrial pollutants. Here, we identify a single locus at which both independent mutation events and selection on an allele shared via gene flow, either slightly before or during selection, play a role in adaptation across the species' range.

KEYWORDS coalescent theory; composite likelihood; convergent adaptation; genetic hitchhiking; positive selection

CONVERGENT adaptive evolution, where selection independently drives the evolution of the same trait, demonstrates the impressive ability of natural selection to repeatedly shape phenotypic diversity (Losos 2011). Many studies have revealed cases of repeated adaptation resulting from changes in the same molecular mechanisms across distinct lineages (Wood *et al.* 2005; Stern 2013). Here, we use the term convergence to define all cases of repeated evolution of similar traits across independent lineages, and do not distinguish between convergent and parallel evolution (Arendt and Reznick 2008). In some cases, these convergent adaptive changes are identical at the level of the same orthologous gene or nucleotide (Martin and Orgogozo 2013), suggesting adaptation may be more predictable and constrained than previously appreciated. Studying repeated evolution has long played a key role in evolutionary biology as a set of replicated natural experiments to help build comparative arguments for traits as adaptations, and to identify and understand the ecological and molecular basis of adaptive traits (Harvey and Pagel 1991).

While we often think of convergent evolution among long-separated species, populations of the same (or closely related) species often repeatedly evolve similar traits in response to similar selective pressures (Arendt and Reznick 2008). Convergent adaptation at the genetic level among closely related populations may arise via multiple, distinct modes (see Stern 2013, for a recent review). Selected alleles present at the same loci in multiple populations can have

doi: https://doi.org/10.1534/genetics.117.300417
Manuscript received March 22, 2017; accepted for publication September 30, 2017; published Early Online October 19, 2017.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300417/-/DC1.
[1]Corresponding authors: Department of Evolution and Ecology, University of California, Davis, 2320 Storer Hall, One Shields Ave., Davis, CA 95616. E-mail: krmlee@ucdavis.edu; and gmcoop@ucdavis.edu

multiple independent mutational origins (*e.g.*, Tishkoff *et al.* 2007; Pearce *et al.* 2009; Chan *et al.* 2010). Alternatively, adaptation in different populations could proceed by means of selection on the standing variation present in their ancestor (*e.g.*, Colosimo *et al.* 2005; Roesti *et al.* 2014), or a single allele spread throughout the populations via gene flow (*e.g.*, Song *et al.* 2011; Heliconius Genome Consortium 2012). Understanding the source of convergent adaptation can aid in our understanding of fundamental questions about adaptation. Distinguishing among these modes may provide evidence for how restricted the paths adaptation can take are to pleiotropic constraints, and if adaptation is limited by mutational input (see Orr 2005, for review). Additionally, we can improve our understanding of the role of standing variation and gene flow in adaptation (Barrett and Schluter 2008; Hedrick 2013; Welch and Jiggins 2014).

With the advent of population genomic data, it is now possible to detect genomic regions putatively underlying recent convergent adaptations. A growing number of studies are sequencing population genomic data from closely related populations, in which some have potentially converged on an adaptive phenotype (*e.g.*, Turner *et al.* 2010; Jones *et al.* 2012). Population genomic studies of convergent evolution often take a paired population design, sampling multiple pairs of populations that independently differ in the key phenotype or environment. These studies are usually predicated on finding large effect loci that have rapidly increased from low frequency to identify the population genomic signal of selective sweeps shared across populations that independently share a selective pressure. Regions underlying convergent adaptations can potentially be identified by looking for genomic regions where multiple pairs of populations are strongly differentiated (*e.g.*, using F_{ST}) compared to the genomic background. Another broad set of approaches identify convergent loci by looking for genomic regions where the populations that share an environment cluster together phylogenetically in a way unpredicted by genome-wide patterns or geography (*e.g.*, Jones *et al.* 2012; Pease *et al.* 2016). While these methods have proven useful in identifying loci involved in convergent adaptation, currently there are few model-based ways to identify the signal of convergence in population genomic data or to distinguish the different modes of convergent adaptation. In the case where an allele is shared due to adaptation from standing variation or migration, chunks of the haplotype on which the selected allele arose and swept on will also be shared among the populations (Slatkin and Wiehe 1998; Bierne 2010; Kim and Maruki 2011; Roesti *et al.* 2014), providing a useful heuristic for these modes to be distinguished from convergent sweeps from independent mutations. We also note there are a variety of approaches to detect introgression (see Hedrick 2013; Racimo *et al.* 2015; and Rosenzweig *et al.* 2016 for recent reviews). However, these methods are not usually focused on detecting sweeps in both populations, but rather look for signatures of unusual amounts of shared ancestry between populations. Here, we present coalescent theory that leverages these signatures selection has on linked neutral variation in a model-based approach. We extend this to a statistical method that utilizes genomic data to identify loci involved in, and that distinguish between, modes of genotypic convergence.

Positive selection impacts neutral diversity at linked loci due to hitchhiking (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989) and can be modeled as an increase in the variance in neutral allele frequencies around their ancestral frequencies. We develop coalescent theory to show how shared hitchhiking events between subpopulations act to increase covariance in allele frequencies around their ancestral frequencies at loci near the selected site, and extend this theory under different models of migration and selection on the same standing variation. We incorporate this hitchhiking effect into a multivariate normal model of allele frequencies that also accounts for population structure, allowing for the application to data from many populations with arbitrary relationships. Based on this theory, we present a composite-likelihood-based approach (Kim and Stephan 2002; Nielsen *et al.* 2005; Chen *et al.* 2010; Racimo 2016) that utilizes genomic single-nucleotide polymorphism (SNP) data to identify loci involved in convergence, and distinguish among alternate modes of convergent adaptation. As these models are also specified by relevant parameters, it is possible to obtain estimates for parameters of interest such as the strength of selection, the minimum age and frequency of a standing variant, and the source population of the beneficial allele in cases of migration. We also present a parametric-bootstrapping approach to help with model choice and construct confidence intervals for our parameters as standard likelihood approaches are not applicable to composite likelihoods.

This method should be of wide use with the increase in population genomic samples from across the geographic range of a species. Here, we illustrate the utility of our inference method by applying it to genome-wide polymorphism data from two distinct cases of convergent adaptation. First, we investigate the basis of the convergent adaptation observed across populations of the annual wildflower *Mimulus guttatus* to copper-contaminated soils from two populations sampled near Copperopolis, CA (Wright *et al.* 2015). We find selection has been acting on standing variation shared between these populations for a tolerance allele present prior to the onset of copper mining in this region. To further exemplify the flexibility of our method, we study a more complex population scenario: the rapid adaptation of four populations of killifish (*Fundulus heteroclitus*) to high levels of pollution, sampled across the Eastern seaboard of the United States (Reid *et al.* 2016). We find that even at the level of a single gene, both convergent mutation and selection on an allele shared via gene flow, either slightly before or during selection, have played a role in adaptation in this species.

Models

In the following section, we present models for the three modes of genotypic convergent adaptation: (1) multiple independent mutations at the same locus, (2) selection on shared ancestral standing variation, and (3) migration between populations spreading a beneficial allele (Figure 2). Throughout this section, we compare our derived expectations to coalescent simulations using mssel—a modified version of ms (Hudson 2002) that allows for the incorporation of selection at a single site. This simulation program takes as input the frequency trajectory of the selected allele for each population. We simulate stochastic trajectories of the selected allele in populations following our three modes of convergence (see Appendix A.2 for simulation details). We focus on a set of four populations as shown in Figure 1 where populations 2 and 3 are adapted to a shared novel selection pressure, and populations 1 and 4 are in the ancestral environment. The average coancestry coefficient values across simulations, estimated as described in Appendix A.1, are plotted for 100 bins of recombination distance away from the selected site, which occurs at distance 0. The results for all three models are shown in dashed lines in Figure 3.

Null model

We aim to model the variances and covariances of the neutral allele frequencies within and between populations due to convergent sweeps. First, we must specify a null model that accounts for population structure. Populations will have some level of shared deviations away from an ancestral allele frequency, ε, due to shared genetic drift. Let x_i represent the present day allele frequency in population i (Figure 1). We denote the deviation of this frequency from the ancestral frequency by $\Delta x_i = x_i - \epsilon$. Genetic drift, in expectation across loci, does not change the population allele frequencies (*i.e.*, $\mathbb{E}[\Delta x_i] = 0$), as an allele increases or decreases in frequency with equal probability. Drift does, however, act to increase the variance in this deviation across loci, with this variance increasing as more time is allowed for drift. The variance in the change of neutral allele frequencies in population i is

$$\mathrm{Var}[\Delta x_i] = \mathbb{E}\left[\Delta x_i^2\right] = \epsilon(1-\epsilon)f_{ii}, \tag{1}$$

where f_{ii} can be thought of as the genetic drift branch length leading from the ancestral population to population i (Nicholson *et al.* 2002), specifying how much allele frequencies in population i deviate from their ancestral values (Figure 1). By rearranging Equation 1, f_{ii} can be interpreted as the population-specific F_{ST} for population i relative to the total population, here represented by the ancestral population (Wright 1943, 1951; Nicholson *et al.* 2002; Weir and Hill 2002).

Populations covary in their deviations from ε as some populations are more closely related due to shared genetic drift resulting from shared population history or gene flow. The covariance in this deviation between populations i and j is

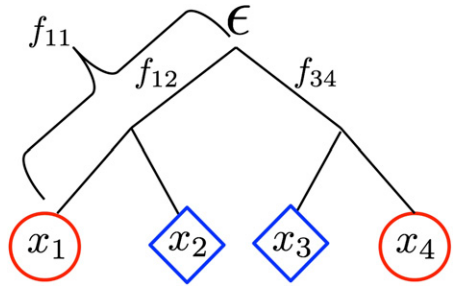

Figure 1 Present day population allele frequencies at a given neutral locus (x_1–x_4 for populations 1–4, respectively) are derived from ancestral allele frequency ε. Each population has a coancestry coefficient proportional to the amount of drift experienced since the split from the ancestral population. f_{11} is shown for population 1. Here, populations 1 and 2, and 3 and 4 share drift relative to the ancestral population, and have nonzero coancestry coefficients f_{12} and f_{34}, respectively. Blue diamonds represent the novel selective environment, and red circles the ancestral environment. Note that branch lengths are not proportional to time in generations (unless there is no migration and the amount of drift is small).

$$\mathrm{Cov}\left[\Delta x_i, \Delta x_j\right] = \mathbb{E}\left[\Delta x_i \Delta x_j\right] = \epsilon(1-\epsilon)f_{ij}, \tag{2}$$

where f_{ij} is interpreted as the coancestry coefficient between populations i and j, and can be thought of as the shared branch length connecting i and j to the ancestral population (Figure 1).

Other natural interpretations of f_{ii} and f_{ij} follow from these definitions. Specifically, these values are probabilities of a pair of lineages being identical-by-descent relative to the ancestral population, *i.e.*, the probability two sampled lineages coalesce before reaching the ancestral population (see Thompson 2013, for a recent review). We briefly review this coalescent interpretation in Appendix A.1. For f_{ii} these two lineages are sampled both from population i. For f_{ij}, one lineage is sampled from population i and the other from population j. We note that, in practice, we do not get to observe the ancestral frequency, nor may the history of our populations be well represented by a tree-like structure (for instance the history of our populations may be reticulated). However, for the sake of clarity, we proceed with these assumptions, and deal with these complications in the implementation of the method.

We define a matrix, **F**, for K populations as a $K \times K$ matrix of coancestry coefficients. For example, for the four populations shown in Figure 1, this matrix takes the following form:

$$\mathbf{F} = \begin{bmatrix} f_{11} & f_{12} & 0 & 0 \\ f_{12} & f_{22} & 0 & 0 \\ 0 & 0 & f_{33} & f_{34} \\ 0 & 0 & f_{34} & f_{44} \end{bmatrix}$$

Populations i and j that split after the ancestral population and share no additional drift (*e.g.*, populations 1 and 3) have $f_{ij} = 0$ by definition.

Incorporating selection

Positive selection impacts neutral diversity at linked loci due to hitchhiking. As the beneficial allele increases rapidly in

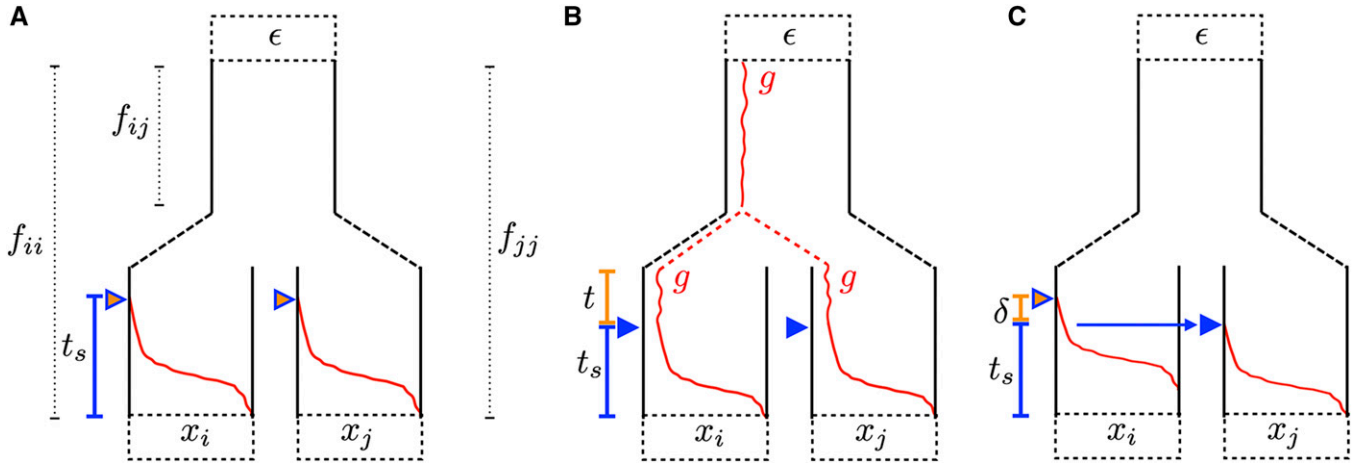

Figure 2 Trajectories of the beneficial allele (red) for the three modes of convergent adaptation. Populations *i* and *j* are under selection with present-day allele frequencies x_i and x_j at a neutral locus, derived from an ancestral population with allele frequency ε. The populations share some amount of drift proportional to f_{ij} before reaching the ancestral population. (A) Independent mutations model. Beneficial mutations, indicated by the orange triangles, occur independently in the selected populations after they have become isolated. Selection begins, indicated by the blue triangles, once the beneficial allele is present in the population. The beneficial allele sweep to fixation in t_s generations. (B) Standing variant model. The beneficial allele is standing at frequency *g* in the ancestral population. After the selected populations split, it is still standing at frequency *g* for *t* generations prior to the onset of selection. (C) Migration model. The beneficial allele arises in population *i* and begins sweeping in population *i*. Meanwhile, there is a continuous low level of migration from population *i* into population *j*. The beneficial allele establishes in *j* after δ generations, where it is swept to fixation in t_s generations.

frequency, so does the haplotype on which it arose. Neutral alleles further from the selected site may recombine off the selected background during the sweep, whose duration depends on the strength of selection (s) and weakly on the effective population size (N_e). The effect of hitchhiking on the changes of linked neutral allele frequencies is similar to that of genetic drift. Hitchhiking does not alter the expected frequency change of linked neutral alleles across loci (*i.e.*, $\mathbb{E}[\Delta x_i] = 0$) because the selected mutation arises on a random haplotypic background. Moreover, hitchhiking increases the variance in the deviation in neutral allele frequencies away from their ancestral values ($\mathrm{Var}[\Delta x_i]$) at linked sites (Gillespie 2000). Shared hitchhiking events between subpopulations will act to increase covariance in allele frequency deviations between subpopulations ($\mathrm{Cov}[\Delta x_i, \Delta x_j]$) at loci near the selected site. This effect of hitchhiking on linked diversity, within and among populations gives us a way to distinguish among alternate modes of convergent adaptation.

We define new matrices of coancestry coefficients that incorporate selection in addition to drift as $\mathbf{F}^{(S)}$. In the following section, we use a coalescent approach to derive coancestry coefficients within and between populations, $f_{ii}^{(S)}$ and $f_{ij}^{(S)}$, for the three modes of genotypic convergent adaptation (Figure 2). In Supplemental Material S2 in File S1, we derive some of the same results forward in time to help guide the reader's intuition. For all models, we assume the beneficial allele has gone to fixation in all selected populations recently. Note that all our models of selection are phrased in terms of distortions to the neutral matrix $\mathbf{F}$; therefore, the precise source of the neutral population structure (*e.g.*, whether its due to shared population history or migration) is relatively unimportant to our approach. A deeper knowledge of the basis of this structure does add to the interpretation of the results, as we explain in the *Discussion*.

Independent mutation model: We first consider the case when a beneficial allele arises independently via *de novo* mutations at the same locus, or tightly linked loci, in both of the selected populations. We expect hitchhiking to increase the variance in neutral allele frequency deviations around the selected site in both populations. However, as the sweeps are independent and there is no gene flow between populations during or after the sweep, we expect no covariance in the neutral allele frequency deviations between these populations, beyond that expected under neutrality due to shared population history prior to the introduction of the beneficial allele.

Moving backward in time, sampled neutral lineages linked to the selected site will be forced to coalesce if both lineages do not recombine off the sweep. We define the probability that a single neutral allele fails to recombine off the background of the beneficial allele during the sweep phase as *y*, which we can approximate as

$$y \approx e^{-rt_s/2} \tag{3}$$

(Kim and Stephan 2002; Durrett and Schweinsberg 2004; Nielsen *et al.* 2005), where r is the recombination rate between the neutral locus and selected site, and t_s is the amount of time the sweep phase takes (Figure 2a). When the beneficial allele arises from a new mutation and selection is additive, $t_s \approx 2\log(4N_e s)/s$, where s is the selection coefficient for the heterozygote, such that heterozygotes

experience a selective advantage of s and homozygotes $2s$ (Barton 1998; Gillespie 2000). The factor of $4N_e s$ is due to the fact that our new mutation, if it is to establish in the population, rapidly reaches frequency $1/(4N_e s)$ in the population, and then increases deterministically from that frequency (Maynard Smith 1971; Barton 1998; Kim and Stephan 2002; Kim and Nielsen 2004).

The coancestry coefficient in population i that experiences a sweep, $f_{ii}^{(S)}$, is defined as the probability that two lineages sampled from population i coalesce either due to the sweep phase, or neutrally before reaching the ancestral population. With probability y^2, both lineages fail to recombine off the beneficial background during the sweep, and they will be forced to coalesce. If one or both lineages recombines off the sweep (with probability $1-y^2$), they can coalesce before reaching the ancestral population with probability f_{ii}. Combining these, we find

$$f_{ii}^{(S)} = y^2 + (1-y^2) f_{ii} \tag{4}$$

For convenience, in our inference procedure, we assume the same strength of selection between our selected populations and thus duration of the sweep is the same. So, $f_{jj}^{(S)}$ takes the same form as Equation 4, with its own neutral probability (f_{jj}) of coalescing. Given that we assume the sweeps complete recently and have the same duration, the mutational events occur at approximately the same time in each selected population. If we assume there is no neutral migration among populations, Equation 4 will hold regardless of where the sweep occurs on the branch leading to i (but when migration occurs we need the sweep to be recent so that lineages sampled from population i are found in population i when the sweep occurs).

For the coancestry coefficient between two selected populations i and j, we can calculate the probability two lineages, one sampled from population i, and the other from population j, coalesce. When the sweeps are independent, the lineages can only coalesce with probability f_{ij} before reaching the ancestral population, as they have no probability of coalescing during the sweep phases which have independent origins. Thus,

$$f_{ij}^{(S)} = f_{ij} \tag{5}$$

Comparison to simulated data: In Figure 3a, we show the case of convergence due to independent origins of the beneficial allele. As we predicted, there is no additional coancestry between the selected populations. Additionally, we show how the coancestry within a selected population decays with distance from the selected site for a range of values for the strength of selection. These coancestry values decay to the neutral expectation at other regions of the genome. With larger s, this decay is slower as the sweep occurs more rapidly, and there are fewer chances for recombination to occur during this time.

Standing variant model: We turn now to the case of a sweep shared between populations i and j due to selection acting on shared ancestral variation (Figure 2b). Our model is appropriate for cases where the standing variation from which the sweep arises was previously neutral, or was maintained in the population at some low frequency by balancing selection. Let the beneficial allele be standing at frequency g in the ancestral population. We assume that the beneficial allele frequency does not deviate much from that of the ancestral population such that it is still g in the daughter populations prior to selection. Selection favoring the beneficial alleles begins t generations after the populations split, and the beneficial allele reaches fixation in both populations after t_s generations (see Figure 2b). We assume t, g, and s are the same for all of our selected populations. More work is needed to allow population-specific parameters to relax these assumptions. We acknowledge all selected populations starting from the same beneficial allele frequency may be unrealistic in many cases, particularly if t is long or if the populations experience bottlenecks at the time of the split.

We first consider the coalescent process of two lineages within a single selected population. Again, y is the probability that a neutral lineage fails to recombine off the background of the beneficial allele during the sweep phase. Given that the beneficial allele is increasing from frequency g, y takes the same form as Equation 3, where now $t_s \approx 2\log(1/g)/s$. If both lineages fail to recombine off the beneficial background during the sweep, there is a probability of coalescing during the standing phase that is higher than the probability of two neutral lineages randomly sampled from the population coalescing. Following from our assumptions during the standing phase, the rate at which two lineages coalesce within a population is $1/(2N_e g)$ per generation. Alternatively, a lineage can recombine off in the standing phase onto the other background with probability $r(1-g) \approx r$ per generation. As these are two competing exponential processes, the probability two lineages coalesce before either recombines off the beneficial background can be simplified to

$$\mathrm{P}(\text{coalesce in standing phase}) = \frac{1}{1+4N_e rg}, \tag{6}$$

as described by Berg and Coop (2015). If either neutral lineage recombines off the beneficial background before they coalesce, the probability of coalescing with the other lineage before reaching the ancestral population can be treated as the coancestry coefficient associated with that particular portion of the population tree.

Taking these approximations into account, we derive a coancestry coefficient for a neutral allele in population i that experiences selection from standing variation as

$$f_{ii}^{(S)} = y^2\left(\frac{1}{1+4N_e rg} + \frac{4N_e rg}{1+4N_e rg} f_{ii}\right) + (1-y^2) f_{ii} \tag{7}$$

The first term corresponds to both lineages failing to recombine off the beneficial background during the sweep phase, which puts them both on the same background as the beneficial allele in the standing phase. Now, the two lineages can

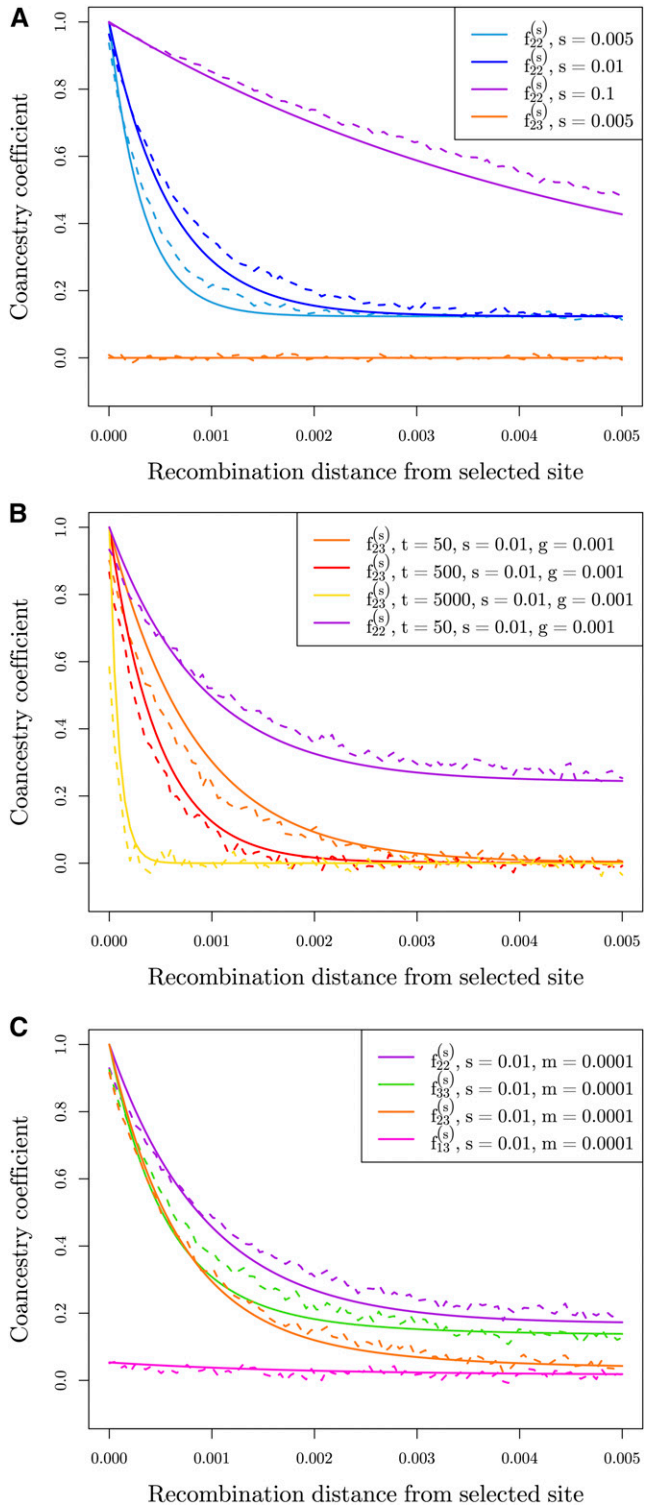

Figure 3 We calculated the average coancestry coefficient values across 1000 runs of simulations for each of 100 bins of distance away from the selected site to compare our simulation results (dashed lines) to our theoretical expectations (solid lines). (A) Average coancestry coefficients under the independent mutations model ($N_e = 100,000$) within a selected population (population 2) with varying s. Also shown is the coancestry coefficient between selected populations which in this case is 0, the neutral expectation. (B) Coancestry coefficients under the standing variation model between selected populations with varying amount of time beneficial allele has been independently standing in populations (t). The coancestry coefficient within a single population is also shown for $t = 50$. For all, $N_e = 10,000$, $g = 0.001$, and $s = 0.01$. (C) Coancestry coefficients under the migration model, within both selected populations (source population 2 and recipient population 3) as well as between source and recipient (2,3) and between recipient and a nonselected population (1,3). Here, we show one set of parameters ($s = 0.01$, $m = 0.001$,and $N_e = 10,000$), as estimates do not vary dramatically with changing m (see Figure S2 in File S1).

either coalesce in the standing phase or recombine off of the background of the beneficial allele where they can coalesce neutrally before they reach the ancestral population. Alternatively, one or both lineages can recombine off during the sweep phase, and again they can coalesce neutrally.

Populations that share a sweep due to shared standing ancestral variation will have increased covariance in the deviations of neutral allele frequencies around their ancestral means around the selected site since they will have a shared segment of the swept haplotype. From a coalescent perspective, this occurs because two lineages sampled from each population have a higher probability of coalescing if they stay on the beneficial background during the sweep and standing phases than two lineages sampled randomly between the populations.

The probability that a single lineage does not recombine off onto the nonbeneficial background during the standing phase for t generations can be approximated as

$$(1 - r_t) = (1-r(1-g))^t \approx e^{-rt} \tag{8}$$

The coancestry coefficient between populations i and j is now

$$f_{ij}^{(S)} = y^2\left((1-r_t)^2\left(\frac{1}{1+4N_erg} + \frac{4N_erg}{1+4N_erg}f_{ij}\right) + \left(1-(1-r_t)^2\right)f_{ij}\right) + (1-y^2)f_{ij}. \tag{9}$$

This derivation follows from that of $f_{ii}^{(S)}$ in Equation 7, but now incorporates the additional probability $(1-r_t)^2$ of both lineages failing to recombine off the beneficial background during their independent standing phases for time t.

This standing variation case represents a simple model of selection on standing variation. However, we expect in many cases that the beneficial allele has not been standing since the ancestral population of the convergent population, but rather has been moved among populations by migration before becoming adaptive at some later time point. In these cases we invoke a model where the standing allele spreading by migration from some source population to recipient populations t generations in the past before the allele became favored. See Appendix A.4 for details. This model differs from the migration model presented in the next section, in which we assume a continuous rate of migration throughout the duration of the sweep and that the variants sweep as soon as they are established in the population. In this standing case

with a source of the standing variant, moving backward in time, we assume that the allele is standing for t generations in a population after the sweep and before the beneficial lineage migrates back instantly into a specified source population (see Figure 11). Biologically, it naturally captures the case where the allele is shared between the populations due to migration, but is standing for sometime before it sweeps. For data analysis, we default to using this more complex model, where sampled selected populations are evaluated as possible sources of the standing variant.

Extending this models to allow for the source to be a nonsampled population would be useful in studying the so-called "the transporter hypothesis" (Schluter and Conte 2009; Bierne *et al.* 2013; Welch and Jiggins 2014), where adaptive gene flow is acting to introduction variation standing in another population. Here, more work is needed to address issues related to estimating coancestry coefficients for unsampled populations (see Appendix A.4 for more information).

Comparison to simulated data: In Figure 3b we show comparisons of simulations to show the fit of our predictions to simulations with adaptation from standing variation in the classic sense. As the duration of the independent standing phases, t, increases, the coancestry at linked neutral alleles between selected populations decreases. Forward in time, this has the interpretation that the longer the beneficial allele is standing in the populations, the shorter the shared haplotype between the populations will be due to independent recombination events before selection begins. In the case that the beneficial allele has been standing for a very long time ($t \to \infty$) before selection occurs, this additional covariance will reduce to zero, as in the independent sweeps case (Equation 5). We acknowledge this scenario is biologically unrealistic. For large values of t at small g, we expect it is likely that the allele would get either be lost or there may be allelic turnover due to recurrent mutations of the beneficial allele. However, it is useful here to gain intuition about when our models overlap. Conversely, if the standing variant is very young ($t \to 0$), the decay in covariance between populations takes the form of the variance within populations (Equation 7), which, as we will see in the next section, looks similar to the pattern generated under the migration model.

Migration model: We now consider the case where the selected allele is spread across subpopulations by migration. This scenario has been studied by a number of authors (Slatkin and Wiehe 1998; Santiago and Caballero 2005; Kim and Maruki 2011; note, these all assume that the allele sweeps in all of the populations), and our approach here follows lines similar to those of Kim and Maruki (2011). Let there be a single origin of the beneficial allele, which occurs in population i. We assume a low, continuous level of migration during the sweep, with a proportion m of individuals in population j coming from population i each generation. Here, we are considering only unidirectional migration from population i into population j. We say the sweep began in population j at time t_s generations in the past, and at time $t_s + \delta$ for population i (Figure 2c). Kim and Maruki (2011) found that the mean delay time, δ, between the two sweeps can be approximated by

$$\delta \approx \frac{1}{s}\log\left(1+\frac{s}{m}\right). \tag{10}$$

The coancestry coefficient of the source population, $f_{ii}^{(S)}$, follows that of a population experiencing an independent sweep from new mutation (Equation 4). To derive the coancestry coefficient of the recipient population, $f_{jj}^{(S)}$, we first need to consider the fate of two lineages sampled in population j at the selected site. Two events can occur if we trace the lineages of two beneficial alleles back in time: either the two lineages coalesce in population j, and a single lineage migrates back into population i, or the two lineages independently migrate back into the source population and coalesce there. We define the probability of these two events as Q and $1-Q$, respectively. We use the approximation

$$Q \approx \frac{1}{1+4Nm} \tag{11}$$

(see Pennings and Hermisson 2006). Assuming m is small, such that a beneficial allele sampled at present day in population j migrates back into population i approximately t_s generations in the past, the probability of a linked neutral allele recombining off during the sweep phase in population j can be approximated by y. If the lineage migrates back into population i before it recombines off the beneficial background, there is an additional time δ in population i for recombination to happen. So, there is an additional probability, $e^{-r\delta}$, of recombination of our linked neutral allele off the beneficial background.

Thus, the coancestry coefficient for the recipient population is now

$$\begin{aligned} f_{jj}^{(S)} = Q\left(y^2 + (1-y)^2 f_{jj} + 2y(1-y)f_{ij}\right) + (1-Q)\left(y^2 e^{-2r\delta}\right. \\ \left. + y^2\left(1-e^{-2r\delta}\right)f_{ii} + 2y(1-y)f_{ij} + (1-y)^2 f_{jj}\right) \end{aligned} \tag{12}$$

The terms in this approximation correspond to the following coalescent scenarios: first, if two lineages sampled in population j coalesce before migrating (with probability Q), then linked neutral alleles can coalesce either during the sweep if neither lineage recombines off the beneficial background, neutrally in population j if both lineages recombine off, or neutrally shared drift phase of populations i and j if just one lineage recombines off. Alternatively, if the two lineages fail to coalesce before one or both migrates (w.p. $1-Q$), there are four ways linked neutral alleles can coalesce:

1. Both lineages fail to recombine off the beneficial background during the sweep, and are forced to coalesce during the sweep in population i. The factor $e^{-2r\delta}$ represents the additional opportunity for recombination when both lineages have migrated back into population i.

2. Both lineages stay on the beneficial background in population j (w.p. y^2), but one or both lineages recombines off in population i (w.p. $1-e^{-2r\delta}$) and they coalesce neutrally in the source population with probability f_{ii} before reaching the ancestral population.
3. Either lineage recombines off the beneficial background while it is still in population j, and the two lineages coalesce neutrally in the shared drift phase of populations i and j, with probability f_{ij} before reaching the ancestral population.
4. Both lineages recombine off during the sweep phase while they are still in population j, and they coalesce neutrally with probability f_{jj}.

When a beneficial allele is shared between populations i and j via migration, there will be additional covariance in the deviations of linked neutral allele frequencies from their ancestral means. In this case, there are three ways a lineage sampled from population i and a lineage sampled from population j can coalesce. They are forced to coalesce during the sweep if both lineages fail to recombine off the background of the sweep, which occurs with probability $y^2e^{-r\delta}$. Alternatively, the lineage sampled in population j can recombine off the beneficial background before it migrates back to source population i, in which case the lineages can coalesce neutrally before reaching the ancestral population in their shared drift phase, with probability f_{ij}. Lastly, if the lineage sampled in population j migrates back into population i, then the two sampled neutral lineages can coalesce neutrally in population i with probability f_{ii} if the lineages do not coalesce due to the sweep (*i.e.*, either recombines off in time t_S or δ). Thus, in the case of continuous migration, the coancestry coefficient between the source and recipient population is

$$f_{ij}^{(S)} = y^2e^{-r\delta} + (1-y)f_{ij} + y\left(1-ye^{-r\delta}\right)f_{ii}. \tag{13}$$

To fully specify the coancestry matrix with selection, we need to take into account the effect migration has on nonselected populations. Specifically, the coancestry coefficients between recipient and nonselected populations are impacted since there is some probability linked neutral lineages will migrate from the recipient population into the source population backward in time. Let population k be a nonselected population. Now, the coancestry coefficient between populations j and k can be expressed as

$$f_{jk}^{(S)} = (1-y)f_{jk} + yf_{ik} \tag{14}$$

This is informative about the direction of migration. First, there is no impact of selection on the relationship between the source and nonselected populations. Additionally, the sweep shared via migration will induce additional coancestry between j and k if k is more closely related to our source population (*e.g.*, population 1 in Figure 1 if population 2 is the source). The opposite is true if k is more closely related to our recipient population (*e.g.*, population 4). Now, there is a deficit in the background level of coancestry between populations j and k near the selected site.

Comparison to simulated data: In Figure 3c, we show our results above compared to simulations with continuous migration during the sweep phase, for a single set of parameters ($s = 0.01,\ m = 0.001$). Here, we have migration occurring from population 2 into population 3. We show the four relevant coancestries as a function of distance from the selected site: the covariance within source ($f_{22}^{(S)}$), within recipient ($f_{33}^{(S)}$), between source and recipient ($f_{23}^{(S)}$), within recipient and a nonselected population ($f_{13}^{(S)}$). We see the coancestry within the recipient population decays more rapidly than coancestry within the source population. This fits our expectations as there is some probability a lineage will, backward in time, migrate back to the source population, decreasing the probability of coalescing before reaching the ancestral population when m is small. As m increases, this relationship changes (Figure S2 in File S1). We also see increased coancestry near the selected site between the selected populations. The pattern of decay varies from that observed in our standing variation model, except for when t is small. Additionally, we see increased coancestry between the recipient population and a nonselected population that decays with recombinational distance to their neutral expectation. Note, the reverse, coancestry recovering to the neutral expectation with recombinational distance is observed for populations that initially are more related to the recipient population (*i.e.*, population 4), is also seen (Figure S3a in File S1). The coancestries between the source population and nonselected populations are unaffected (Figure S3b in File S1). Together, these observations using information from nonselected populations help distinguish possible source populations.

Inference

We have described how selection at linked loci affects the matrix of coancestry coefficients, allowing us to parameterize the variance and covariance in neutral allele frequency deviations within and between populations. To estimate the likelihood of our data under convergent adaptation models, we need a probability model for how allele frequencies depend on these variances and covariances. Neutral allele frequencies across K populations can approximately be modeled jointly as a multivariate normal distribution around the ancestral allele frequency, ε, with covariance proportional to the coancestry coefficients (Nicholson *et al.* 2002; Weir and Hill 2002; Samanta *et al.* 2009; Coop *et al.* 2010). Specifically,

$$\vec{x} \sim \mathcal{N}\left(\epsilon\vec{1}, \epsilon(1-\epsilon)\mathbf{F}\right) \tag{15}$$

where $\vec{x}$ is a vector of population frequencies, and $\mathbf{F}$ is the K by K matrix of coancestry coefficients without selection.

Above, we demonstrated that we can generate coancestry matrices $\mathbf{F}^{(S)}$ to explain the coancestry between multiple populations due to neutral processes and various modes of convergent adaptation. $\mathbf{F}^{(S)}$ is a function of the neutral coancestry,

(**F**) the model of convergence (M) and its parameters (Θ_M), and the recombination distance a neutral site is away from a selected site (r_l). Thus, modeling neutral allele frequencies as multivariate normal with covariance proportional to this new coancestry matrix, we can calculate the likelihood of observed data a given distance away from the selected site under a specific model of convergence as

$$P\left(\overrightarrow{x_l}\middle| r_l, \mathbf{F}, M, \Theta_M\right) \approx \mathcal{N}\left(\overrightarrow{x_l}\middle| \epsilon_l \overrightarrow{1}, \epsilon_l(1-\epsilon_l)\mathbf{F}^{(S)}(r_l, \mathbf{F}, M, \Theta_M)\right) \tag{16}$$

In practice, we do not know the true ancestral mean at a given locus, ϵ_l, so we use the mean of the present-day population allele frequencies and calculate likelihoods of mean-centered allele frequencies and coancestry matrices (we account for this mean centering in Appendix A.2.6). We also do not know the true neutral coancestry matrix, **F**, but estimate it from deviations of allele frequencies from sample means across the entire genome. We also incorporate the effects of sampling into this variance-covariance matrix. See Appendix A.1 for details.

Composite-likelihood framework

We calculate the likelihood of all data (D_ℓ) in a large window around the selected site (ℓ) under a given model of convergent adaptation (M), with its associated parameters (Θ_M), as the product of the marginal likelihoods for sites all distances away from the selected site. This composite likelihood is used as an approximation to the total likelihood of all sites, but is not a proper likelihood as neighboring sites are correlated due to shared histories. Moving L_{left} sites to the left of the proposed selected site and L_{right} sites to the right,

$$\begin{aligned}\mathcal{L}_C(M, \Theta_M; D_\ell) = &\prod_{i=1}^{L_{\text{left}}} P\left(\overrightarrow{x}_i\middle| M, \mathbf{F}_M^{(S)}(r_i, \mathbf{F}, M, \Theta_M)\right) \\ &\times \prod_{j=1}^{L_{\text{right}}} P\left(\overrightarrow{x}_j\middle| \mathbf{F}_M^{(S)}(r_j, \mathbf{F}, M, \Theta_M)\right)\end{aligned} \tag{17}$$

where r_i is the genetic distance from site i to ℓ, and similarly for r_j. We can also obtain a composite likelihood of our data under a neutral model (N), $\mathcal{L}_C(N; D_\ell)$, which is only parameterized by **F**. This framework enables us to:

1. Identify the maximum likelihood location of the selected locus in a region by varying the location of the proposed selected site. For a given region and model of convergent adaptation we vary the location of the selected site, taking the maximum composite likelihood over a grid of parameters. We take as our best estimate of the location under a given model of convergence, the maximum composite-likelihood location of the selected site $\left(\hat{\ell} = \underset{\ell, \Theta_M}{arg\,max}\ \ \mathcal{L}_C(M, \Theta_M; D_\ell)\right)$.
2. Determine the parameter(s) that maximize our composite-likelihood estimates under a given model at a given location of the selected site (ℓ). We obtain these maximum composite-likelihood estimate (MCLE) parameters by evaluating the composite likelihood across a grid of parameters for a given location of the selected site $\left(\widehat{\Theta_M} = \underset{\Theta_M}{arg\,max}\ \ \mathcal{L}_C(M, \Theta_M; D_\ell)\right)$.
3. Distinguish between modes of convergence, and neutrality, in a genomic region by comparing the maximum likelihood under various models of convergent evolution. At a given location of the selected site (ℓ), we compare the maximum composite likelihood of each model to the neutral model $\left(\log\left(\mathcal{L}_C\left(M, \widehat{\Theta_M}; D_\ell\right)/\mathcal{L}_C(N; D_\ell)\right)\right)$.

This composite likelihood ignores the correlation in allele frequencies (linkage disequilibrium) between neutral sites so the composite-likelihood surface will be too peaked. A number of authors have taken composite-likelihood approaches to inferring a range of population genetic parameters [*e.g.*, Hudson (2001); see Larribe and Fearnhead (2011) and Varin *et al.* (2011) for a broader statistical views on composite likelihood]. In the setting of inferring genome-wide parameters, *e.g.*, parameters of neutral demographic models, the MCLEs are known to be consistent in the limit of many unlinked genomic regions (Wiuf 2006). While, in general, composite-likelihood methods perform well, in all of these settings typical measures of uncertainty of parameters (confidence intervals) and model choice methods [*e.g.*, Akaike information criterion (AIC)] are undermined due to the over-peakiness of the likelihood.

Composite-likelihood approaches have also been used in the context of selective sweeps, starting with Kim and Stephan (2002), who take a composite likelihood formed like Equation 17 of the product of marginal probabilities of allele frequencies within a single population moving away from a proposed selected site (an approach expanded on in Kim and Nielsen 2004; Nielsen *et al.* 2005; Chen *et al.* 2010; DeGiorgio *et al.* 2014; Racimo 2016). Our method is most closely related to that of Chen *et al.* (2010) and Racimo (2016), who look at allele frequencies across two or three populations, respectively, and look for the signal of a sweep in one of the populations [or, in the case of Racimo (2016), in the ancestor of a pair of populations]. We note that we have a further layer of abstraction over these previous composite-likelihood methods. Extending Kim and Stephan (2002), previous methods have calculated the likelihood of the sample frequency considering a binomial draw from some underlying population frequency, which is naturally modeled as being bounded between 0 and 1. We, however, use a multivariate normal likelihood to model our sample frequencies, which does not bound allele frequencies between 0 and 1. This further abstraction is justified by the fact that, by using the multivariate normal approach, we are able to handle arbitrarily large number of populations with arbitrary population structure, and to flexibly model different forms of selection into an easily extendable form to the covariance matrix. Future work could potentially concentrate on hybrid approaches, combining the flexibility of our approach with the realism of previous approaches.

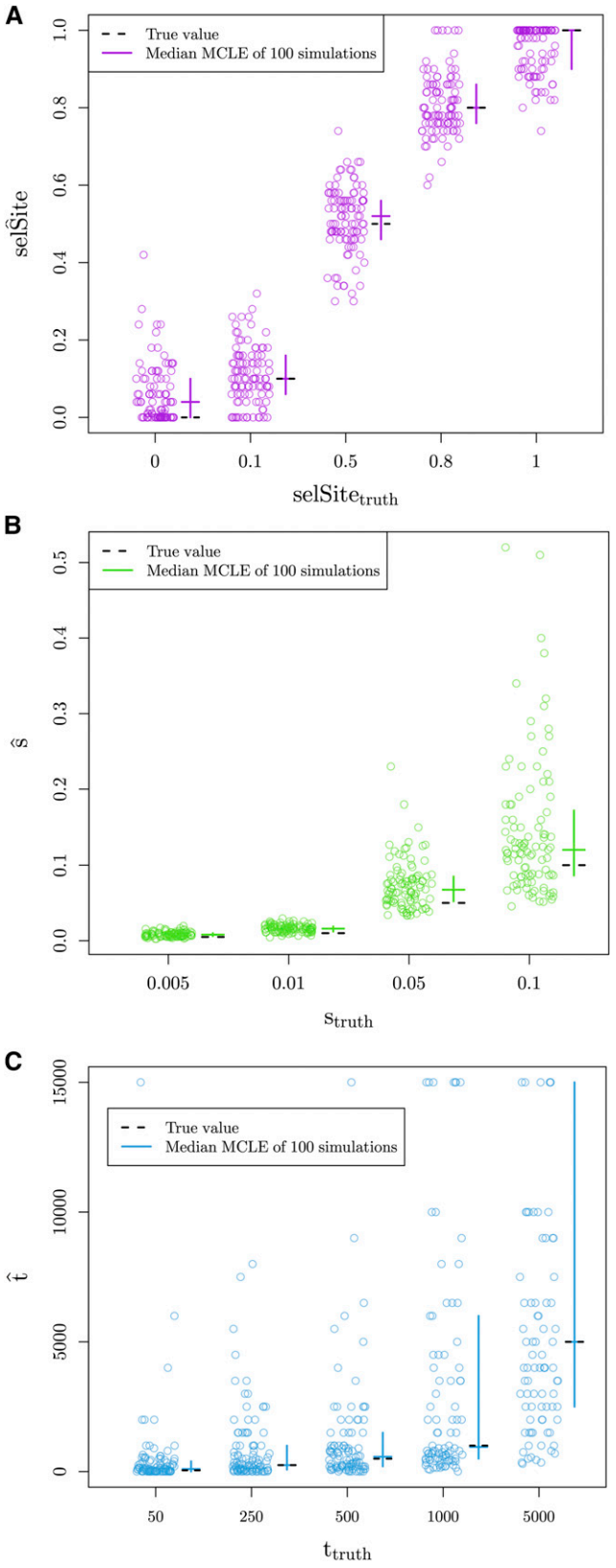

Figure 4 MCLEs calculated under model used for simulation. We vary the true value of the parameter used for simulations along the *x*-axis, and show the MCLE for each of 100 simulations (points). Crossbars indicate first and third quartiles with second quartiles (medians) as the horizontal line. The true values of the parameters are marked with dashed, black lines. (A) MCLE of the location of selected site for 100 simulations under the independent mutation model (10 chromosomes per population, N_e = 100,000, and s = 0.05). (B) MCLE of the strength of selection (s) for 100 simulations under the independent mutation model (10 chromosomes per population, N_e = 100,000). (C) MCLE of the standing time (t) for 100 simulations under the standing variant model (10 chromosomes per population, N_e = 10,000, s = 0.01, and g = 0.001). For scale, we left out estimates of t >15,000 (2, 9, and 21 data points when t_{truth} = 500, 1000, and 5000, respectively.)

Inference method on simulated data

To test our method, we utilized the datasets generated using mssel (as discussed above with details in Appendix A.2) to see if we could recover the parameters and convergent mode used for simulation. The neutral coancestry matrix **F** was estimated using data from 1000 runs with no selection (as described in Appendix A.1). We assume that the model parameters N_e and r are known, and we set these at the values used to generate the simulations. We calculated the composite log-likelihoods for each of the simulated datasets under the following four models: neutral (no selection), independent sweep model, standing variation model, and migration model with the beneficial allele originating in population 2. We calculate the likelihoods under a dense grid of selection coefficients (s), migration rates (m), and standing times (t). In the standing variation model, the standing frequency (g) is held at 0.001. See Appendices A.2.4 and A.2.5 for details. We repeat this procedure for each of 100 runs of all simulated datasets. To compare between models, we calculate the composite log-likelihood differences between the true model and all other models including the neutral model, at the MCLE obtained under each model.

Parameter estimation: *Location of selected site:* To explore the ability of our method to localize the selected site, we vary the true location of the selected site simulating under the independent mutation model. We estimate the maximum composite-likelihood location under the independent sweep model over a fine grid of locations and selection coefficients. The method is able to correctly identify the location of selection (Figure 4a), with higher accuracy when the true location of the site is in the middle of the window. The method does show an edge effect when the true location of the selected site is at the edge of the region of interest perhaps because we do not get to see the decay of coancestry on both sides of the selected site. Additionally, we are able to correctly estimate the strength of selection while allowing the location of the selected site to vary (Figure S1a in File S1), and there is no correlation between these joint parameter MLCEs (Figure S1b in File S1).

Independent mutations model: To verify our ability to recover the selection coefficient, we simulated under the independent mutation model for a range of values for s, holding the location of the selected site at its true value. We are able

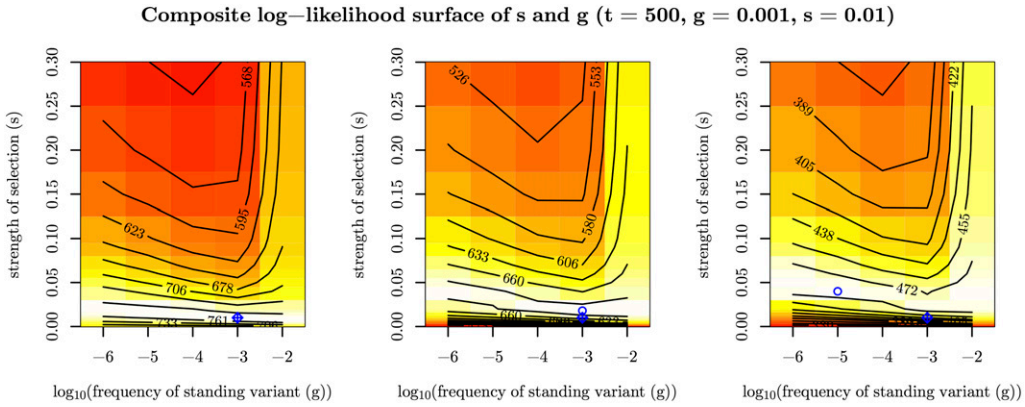

Figure 5 Composite log-likelihood surface of the strength of selection (s) and the frequency of standing variant (g) for three simulations (with N_e = 10,000, t = 500, g = 0.001, and $s = 0.01$) to exemplify confounding of s and g under the standing variant model. Blue diamond pluses represent the true location of the parameters used for simulation. Blue circles represent MCLE.

to recover the parameters used for simulation (Figure 4b). The ability to correctly estimate s breaks down for large enough s, given a fixed window-size around the selected site and r_{BP}, since we will not observe the full decay in coancestry.

Standing variant model: To explore our inference using the standing variant model, we hold the location of the selected site at its true location and take as our estimate of s and t their values at the joint maximum composite likelihood. Under the standing variant model, we are again able to accurately estimate s (Figure S6 in File S1). The inference of s and g simultaneously is somewhat more confounded (Figure 5). How the signal of the sweep within populations decays, as we move away from the selected site, is primarily determined by s and g (see Equation 7). While a higher frequency of the standing variant (g) can lead to a quicker decay, this can be partially compensated for the strength of the sweep being stronger (higher s and lower t_s). This explains the J-shaped ridge in the likelihood surfaces for s and g, seen in Figure 5. Therefore, in practice, we can often infer a lower bound s and an upper bound for g, but not find the precise values of each when inference is performed under the standing variation model. We are able to accurately estimate the time the beneficial allele has been standing in the independent populations prior to selection, t, as shown in Figure 4c. Our inference of t is relatively free of confounding with s and g, as t primarily governs the decays in coancestry between populations, making it separable from the scale of the sweep within populations.

Migration model: We explored our inference under the migration model of parameters m and s, again fixing the location of the selected site and taking the joint MCLE. We are able to correctly estimate s (Figure S4b in File S1). However, we obtain poor estimates of the rate of migration, m (Figure S4a in File S1). This is perhaps unsurprising as the coancestry coefficients under the migration model depend only weakly on m. We obtain fairly bimodal estimates of m that are usually either very low (10^{-5}–10^{-3}) or high (1). As the true value of m increases, we see fewer estimates of small m and more estimates of $m = 1$. These estimates of m seem to be a true reflection of the patterns in the simulated datasets. Specifically, this effect is mostly observed in the variance within the recipient population, as Equation 12 depends on m in both Q and δ. High m estimates correspond to datasets with lower empirical levels of coancestry within the recipient than datasets where low estimates of m were obtained (Figure S5 in File S1). We believe that the bimodality results from stochasticity in how many lineages ancestral to the sample migrate before they recombine off the sweep in the recipient population. While our estimates of m are noisy, the migration model does capture key features of the spread of adaptive alleles by migration, allowing it potentially to be distinguished from other modes of convergence. We now turn to the performance of the method in distinguishing modes of convergence.

Model comparison: To test the ability of our method to distinguish between modes of convergence, we calculated the maximum composite log-likelihood of 100 simulations for each dataset generated under both the true model and all other models with a fixed, fine-grid of parameter values. The location of the selected site is fixed at its true location. The results are summarized in Figure 6, which shows histograms of the difference in maximum composite log-likelihoods calculated under a given model relative to the true model used for simulation. For example, in evaluating the independent mutations model, we present the difference in the composite log-likelihoods calculated for data simulated under the independent mutations model for all other models and the composite log-likelihood calculated for the true independent mutations model. Thus, values <0 indicate that the correct model has a higher maximum composite log-likelihood than the true model. Conversely, values >0 indicate the incorrect model of convergence has a higher composite log-likelihood than the true model. For inference under the migration model, we fix the source to be the true source of the selected allele when simulating under the migration model, and to an arbitrary one of the two selected populations when performing inference on simulations under other models.

Neutral model: We first compare the composite likelihoods calculated for data generated with no selection. For the selection models, we fix the location of the selected site. The distributions of the resulting composite log-likelihood ratios are shown in Figure 6a. As expected for a composite likelihood, the composite log-likelihood ratio between a convergent selection model and

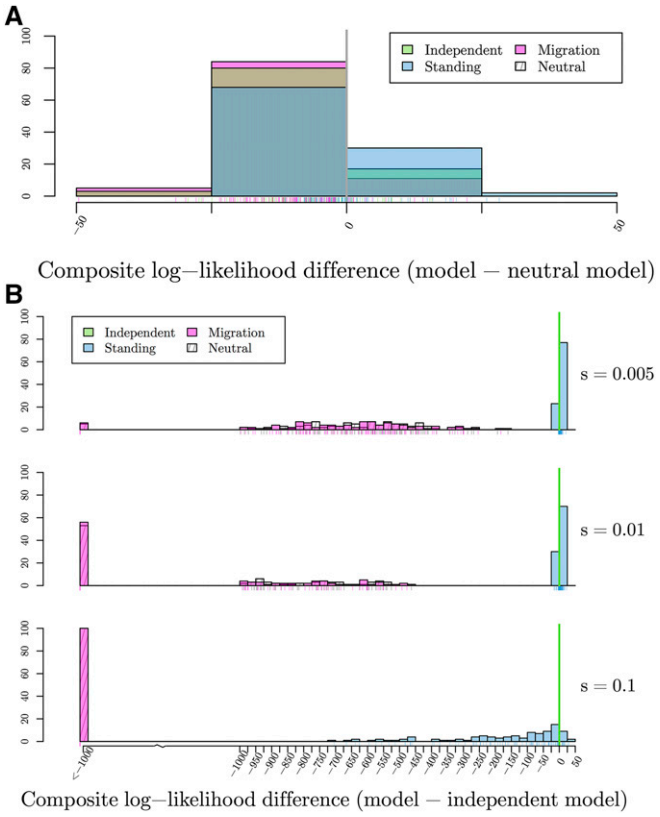

Figure 6 Histograms of the differences in maximum composite log-likelihoods calculated under a given model relative to the true model used for 100 simulations. Parameter values used to simulate are noted, varying along the vertical dimension. Values <0, marked with solid line, indicate the true model has a higher maximum composite likelihood than alternative model. Conversely, values >0 indicate the alternative, incorrect model of convergence has a higher composite log-likelihood than the true model. True models: (A) Differences in maximum composite log-likelihoods under models relative to neutral model. (B) Differences in maximum composite log-likelihoods under models relative to independent mutations model with $N_e = 100,000$. (C) Differences in maximum composite log-likelihoods under models relative to standing variation model with $N_e = 10,000$, $s = 0.01$, and $g = 0.001$. (D) Differences in maximum composite log-likelihoods under models relative to migration model with $N_e = 10,000$ and $s = 0.01$.

the neutral model with no selection are inflated compared to those expected under the usual asymptotic χ^2 distribution. However, these likelihood ratio differences are relatively small compared to those we observed when simulating under alternative models. This is because, when $s \to 0$, in all models with selection, the coancestries converge to our neutral expectations. Indeed, when we look at the MCLE for the strength of selection ($\hat{s}$) under the incorrect models with selection, we see that, for nearly all simulations, $\hat{s}$ is close to zero 0 (Figure 7a). Overall, this suggests that our null model is reasonably well calibrated, given the limitations of composite-likelihood schemes.

Independent mutations model: As shown in Figure 6b, we are able to correctly distinguish between a neutral model of no selection and the true independent mutation model by at least 160 composite log-likelihood units, even for relatively weak selection ($s = 0.005$). This difference increases as the true value of s increases. This same relationship is true when comparing the migration model to the true independent mutation model. Therefore, we have good ability to distinguish the independent sweeps model from neutral and migration model over a range of selection coefficients.

Our ability to distinguish between the standing variation model and the true independent mutation model is less clear. When the true s is small, the two models have comparable composite log-likelihoods, with differences ranging from -3 to 20. This difference decreases, with higher likelihood for the true independent mutation model more frequently, as s increases. This result makes sense when we look into the maximum likelihood estimate of the parameter t (Figure 7b). We obtain estimates of t approaching our highest value on the grid (10^6). Thus, we may not be able to distinguish between the cases where the origins of the beneficial allele are truly independent or whether selection has been on a single variant that has been standing independently for a long time as these two models converge for large t.

Standing variant model: Simulating under the standing variation model, the picture is more complicated. Like the other models, we can exclude the neutral model, although note that this would become challenging when the allele has been standing at high frequencies, $g \gg 0$ (Berg and Coop 2015). When the independent standing time, t, is small, we see little difference in the composite log-likelihoods between the true standing model and the migration model. As t increases, we see a larger difference between these two models. However, as t increases, the composite log-likelihood difference between the independent mutation model and standing variation model tightens around 0. These results fit our expectations as we know the models look similar in the extreme values of t, the migration model when the standing time is small and independent mutation model when the standing time is large, respectively.

Migration model: We are able to distinguish the migration model from the neutral and independent sweeps model. However, the standing variation and true migration model are again somewhat confounded. The values of the composite log-likelihood differences range from -44 to 123 when $m = 10^{-4}$, and this range narrows closer to 0 as m increases. These results fit our understanding when we again look at the MCLEs of t in the standing model. Now, the estimates are at or close to 0 (Figure 7c), indicating it is hard to distinguish between convergence that is due to migration or selection on a shared standing variant that has only been standing for a very short time, as they result in similar patterns in decay of coancestries.

Summary: We can clearly distinguish the outcomes of the migration and independent sweeps models from each other. Both models are hard to distinguish from the standing variation case, but in very different regimes of the standing variation model. The estimated time the variant has been standing ($\hat{t}$) for is a helpful indicator of the mode of convergence, with very low estimates meaning that the standing model is indistinguishable from the migration model, while

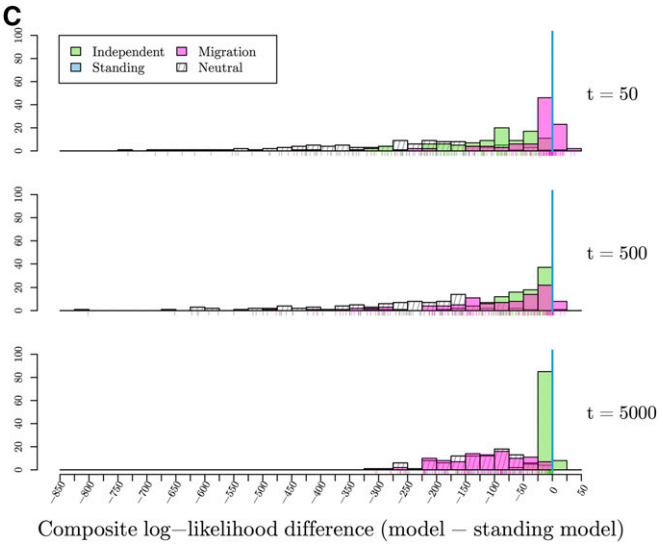

Figure 6 Continued

very high estimates mean that the standing model is indistinguishable from the independent sweeps model. When data are simulated under the standing model with intermediate values of t, we can distinguish this from both independent sweeps and recent migration models. This is because an intermediate value of t generates a covariance pattern not well explained by either other model. Therefore, while comparing the maximum composite likelihoods between models is useful, the estimated value of t is useful in judging the different models.

Evaluating properties of the estimators and models for real datasets: Our use of a composite likelihood means that we cannot rely on standard asymptotic properties of likelihood estimators to construct confidence intervals or help with model choice (*e.g.*, AIC). Therefore, we take a parametric-bootstrapping approach, simulating datasets under the MCLEs of various models, matched for sample sizes and number of segregating sites and other qualities (recombination rate and size of the region, N_e, and neutral **F** matrix), as the original data (see Appendix A.3 for more details). From these simulations, we generate a distribution of composite-likelihood ratios. Specifically, we wish to understand if we have support for a model (j) as compared to a seemingly less likely model (i); this could be a model with selection to one without, or a model with, standing variation compared to one with independent mutations. We simulate datasets under one model (i), using the MCLE of that model applied to the real data; we then estimate the maximum composite log-likelihood of dataset k under this model (L_{ki}), and the maximum composite log-likelihood under a second model j (L_{kj}), and form the distribution over our simulations of the difference $L_{kj} - L_{ki}$. We can then compare the value of the composite log-likelihood ratio ($L_{Dj} - L_{Di}$) obtained for our true dataset D to this distribution to obtain the parametric-bootstrap P-value for the comparison the alternative model (j) compared to the null model (i). Additionally, we generate parametric-bootstrap confidence interval for parameters of interest, particularly t, the minimum age of the standing variant, as this parameter is informative about the overlap of models as shown above.

Applications

Copper tolerance in M. guttatus

The study of adaptation to toxic mine tailings is a classic case of rapid local adaptation to human altered environments (MacNair *et al.* 1993). We apply our inference method to investigate the basis of the convergent adaptation seen between populations of the annual wildflower *M. guttatus* to copper-contaminated soils near Copperopolis, CA. Wright *et al.* (2015) sequenced pooled samples from 20 to 31 individuals from two mine and two off-mine populations from two distinct copper mines in close geographic proximity (all populations within 15 km of each other) to 34–72× genome-wide coverage for each population. They observed elevated genome-wide estimates of genetic differentiation between mine and off-mine populations (F_{ST} M/OM = 0.07 and 0.14), with similar levels of differentiation between the mine populations (F_{ST} MM = 0.13). Only a small number of regions had high levels of differentiation. Here, we focus on the region with the strongest signature of differentiation between the two mine/off-mine pairs found on Scaffold8 by Wright *et al.* (2015). They observed low genetic diversity within each mine population in this region compared to off-mine populations. When the mine populations are compared to each other, they have elevated differentiation in this region, except for in the center, where they share a nearly identical core haplotype. This pattern suggests the sweeps may not have been independent within each mine population, and that the sweep is possibly shared either due to migration or selection of shared standing variation.

We estimate the **F** matrix using SNPs from 12 scaffolds that showed no strong signals of selection (shown in Table S6 in File S1). Using all SNPs in the 169.3 kb Scaffold8, we apply our inference framework to both identify the locus under selection and distinguish between modes of convergence between the two mine populations. We move the proposed selected site along this scaffold, and calculate the composite likelihood under our three modes of convergent adaptation: (1) both mine populations have had independent mutations at the same locus; (2) the beneficial allele was standing in one of the mine populations, and was spread via migration into the other mine population, where it is still standing prior to the onset of selection (as detailed in Appendix A.4); and (3) the beneficial allele arose in one of the mine populations and spread to the other via migration. We estimate the maximum composite likelihood over a dense grid of parameters used to specify these models (Table S7 in File S1). For the migration model, we allow both adapted populations to be possible sources. We use an $N_e = 7.5 \times 10^5$, calculated from

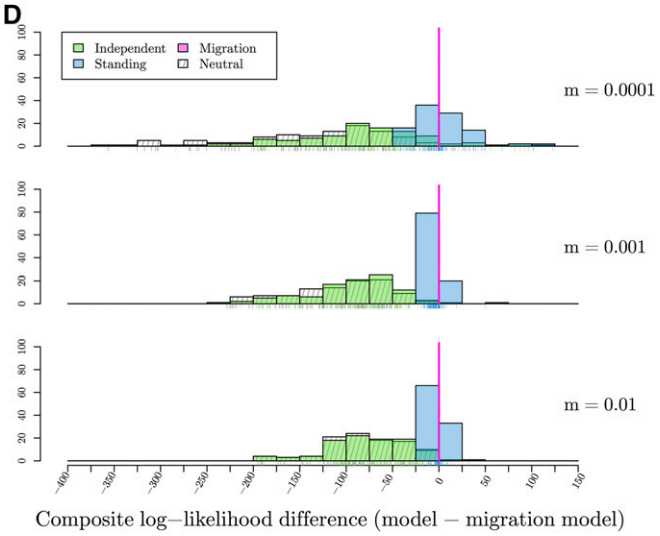

Figure 6 Continued

the observed pairwise diversity $\pi = 4N_e\mu$ using a mutation rate of $\mu = 1.5 \times 10^{-8}$ and $r_{BP} = 4.72 \times 10^{-8}$ (Lee 2009).

In Figure 8a, we summarize the results, showing the difference in maximum composite log-likelihoods between a given model of convergence, and the neutral model of no selection as a function of the proposed selected sites along the scaffold. We see the three likelihoods peaking when the selected site is approximately at position 303–308 kb, and that the model with the highest likelihood is selection on shared ancestral standing variation.

To judge the significance of differences in the composite log-likelihood between the standing-source model and the other models, we used our parametric-bootstrap procedure. We simulated 100 datasets under the independent and migration modes of convergent adaptation at their MCLE as well as a neutral model with no selection (see Appendix A.3 for details). For each simulated dataset, we calculate the composite log-likelihood ratio comparing the standing source model to the likelihood of each of the other models (for their respective simulations), under the same parameter grid as the original data (Table S7 in File S1), but holding the location of the selected site, and, where relevant, the source population constant at their respective MCLEs used for simulation. Our observed composite log-likelihood ratio, comparing the standing source model to each of the others, was well outside the range those obtained by simulation (implying a parametric-bootstrap *P*-value of $< 1/100$). The smallest difference is under the migration model where the range of out 100 composite log-likelihood ratios is [4.12, 749.45], while the observed ratio is 945.95 (see Table S8 in File S1 for all results). These results suggest that the nonstanding source models offer a significantly worse fit to the data.

Focusing on the standing-source model at the most likely selected site, we can obtain parameter estimates for the strength of selection (s), standing frequency of the beneficial allele (g), and the amount of time that the beneficial allele has been standing in both mine populations after they have been isolated but prior to selection (t). The strength of selection and starting frequency of the allele are confounded (Figure 8c) as expected. Our maximum composite log-likelihood parameter estimates suggest selection was relatively strong (> 0.02), and the allele was not standing at very high frequencies ($< 10^{-4}$) when selection began. We see the maximum composite log-likelihood is obtained when the standing time (t) is ~646 generations (Figure 8b). As the Copperopolis *Mimulus* are annual, this corresponds to 646 years. We obtained 95% parametric-bootstrap confidence interval of [364, 9525] generations (years), by simulating under the standing-source at our MCLE (see Appendix A.3). This time also has the interpretation of the minimum age of the standing variant, as it has been standing for at least this amount of time, and potentially longer in the source population. As copper mining started in 1861 in this region (Aubury 1902), this suggests the tolerance allele was present prior to the onset of mining, again consistent with the variant being a standing variant when selection began.

There is little information about the source population of the standing variant (we obtain identical likelihood surfaces for either copper population as the source, see Figure S7a in File S1). This is perhaps unsurprising, as there is relatively little hierarchical structure among the populations. Additionally, we tested the standing variant model with no source and saw no difference in the likelihood surfaces over the proposed selected sites (Figure S7a in File S1). The MCLE of t is higher for the models of standing variation with a source than the simple model of standing variation (see Figure S7b in File S1). This is likely because making one of the populations a source of the standing variant increases the covariance around the selected site among the selected populations, as described in Appendix A.4, and so the model compensates by increasing the rate of decay of this covariance.

Industrial pollutant tolerance in F. heteroclitus

We demonstrate how our method can be extended to more complex population scenarios. Populations of the Atlantic killifish, *F. heteroclitus*, have repeatedly adapted to typically lethal levels of industrial pollutants (Nacci *et al.* 1999, 2010). Reid *et al.* (2016) have sequenced 43–50 individuals from four pairs of pollutant-tolerant and sensitive populations along the US Atlantic coast (see Figure 9a), sequencing each individual to 0.6–7× depth. The southern pair of populations form a distinct clade relative to the northern populations, consistent with a phylogeographic break centered on New Jersey (Duvernell *et al.* 2008).

Reid *et al.* (2016) found that a number of the strongest signals of recent selection are shared between all tolerant populations, suggesting genotypic convergent adaptation. We focus our method on their strongest signal of selection, Scaffold9893 [the scaffold containing the aryl hydrocarbon receptor interacting protein (*AIP*) gene], where all four pairs of tolerant/sensitive populations sampled show high levels of

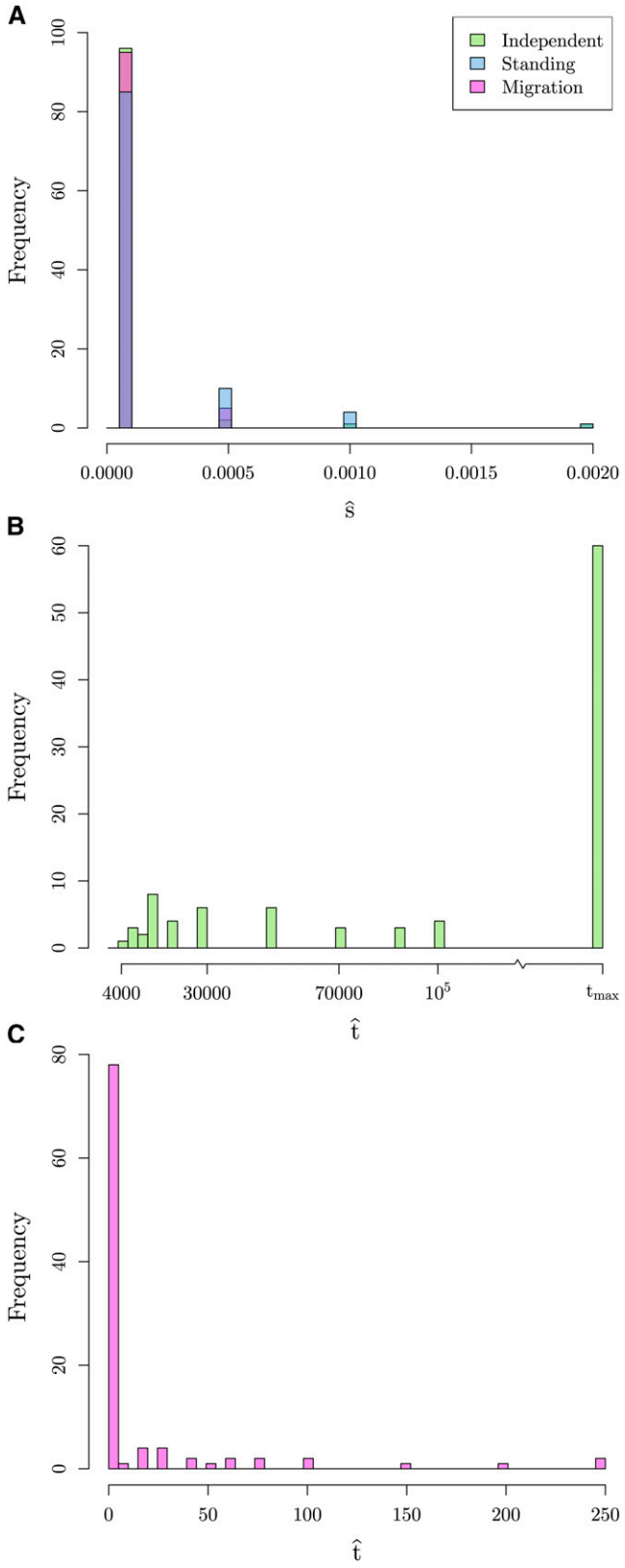

Figure 7 Histograms of MCLE for parameters estimated under incorrect models. (A) Histogram of MCLE of the strength of selection (s) under all convergent models where the neutral model is true model used for simulations. (B) Histogram of MCLE of the standing time (t) under standing variant model where the independent mutation model is true model used for simulations (s = 0.01 and N_e = 100,000). (C) Histogram of MCLE of the standing time (t) under standing variant model where the migration model is true model used for simulations (m = 0.001, s = 0.01, and N_e = 10,000).

differentiation. Here, we test the hypotheses that all four tolerant populations show convergent adaptation due to our three previous modes of independent mutation, migration, or selection on shared ancestral variation. For our standing variation model, we specified the source of the standing variant (as described in Appendix A.4). We also test the hypotheses that there is an independent mutation in the southern tolerant population while the three northern populations share a sweep at this locus, either due to migration between populations or selection on variation present in the ancestor of the Northern populations. This latter set of hypotheses is consistent with the fact that Reid *et al.* (2016) detect a shared haplotype in the three northern tolerant populations, while a different haplotype appears to have swept in the southern tolerant population. We estimated the **F** matrix from four scaffolds that show no strong signal of selection, as shown in Table S9 in File S1. We use $N_e = 8.3 \times 10^6$ and $r_{BP} = 2.17 \times 10^{-8}$ (N. Reid, personal communication).

The results are summarized in Figure 9b. For all models with migration or selection on standing variation, we plot the maximum composite log-likelihood for the most likely source at each location of the selected site (to reduce the number of lines plotted, see Figure S9 in File S1 for the full figure). We see the model with the highest composite log-likelihood is when convergence is due to selection on shared standing variation in the North, and an independent mutation in the southern tolerant population. This occurs when the selected site is at position ~1.96 Mbp on the scaffold.

To assess the significance in the composite log-likelihoods of this model, and the other models tested, we simulate 100 datasets under each model at their MCLE (see Appendix A.3 for details). We calculate the composite log-likelihood ratio for each simulated dataset to compare the standing variation in the North with an independent mutation in the South model to the others models used for simulation. We calculate the composite likelihoods under the same parameter space as used for the original data (Table S10 in File S1), holding the location of the selected site and the source population constant at their MCLEs used for simulation. For the neutral model, and the three models where all four tolerant populations have the same mode of convergence, the observed composite log-likelihood ratio was far outside the range of values obtained from the simulations (see Table S11 in File S1 for all results), suggesting these models offer a significantly worse fit to the data (parametric-bootstrap P-value $< 1/100$). However, this is not true for the model where migration is occurring in the three Northern selected populations, while there is an independent mutation at the same locus in the Southern tolerant population. Here, the range of the difference in maximum composite log-likelihood for

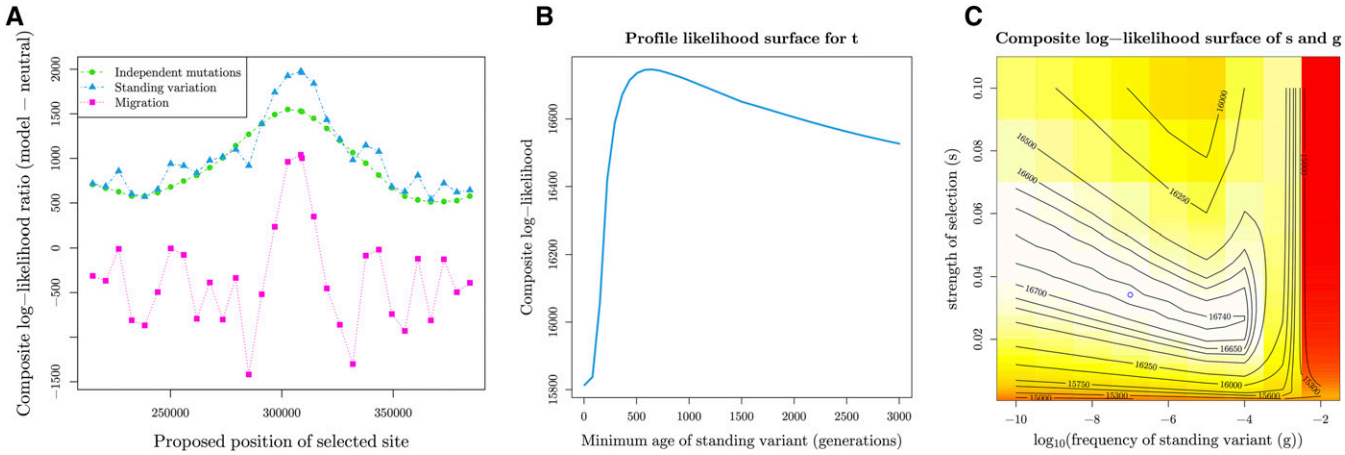

Figure 8 Inference results for *M. guttatus* copper tolerance adaptation on Scaffold8. (A) Composite log-likelihood ratio of given model relative to neutral model of no selection as a function of the proposed selected site. We show likelihoods for the standing-source model maximizing over possible sources, but all results can be seen in Figure S7a in File S1. (B and C) MCLE of parameters in standing variation model with position 308,503 as selected site. (B) Profile composite log-likelihood surface for minimum age of standing variant, maximizing over other parameters, with peak at 646 generations (C) Composite log-likelihood surface for strength of selection *vs.* frequency of standing variant. Blue circle represents point estimate of joint MCLE ($\hat{s}$ = 0.034 and $\hat{g} = 10^{-7}$). t is held constant at MCLE of 646 generations.

100 simulations is [−24,675, 38,997], while the observed difference is 8121 (parametric-bootstrap P-value = 0.58; Figure S10 in File S1). Thus, we are unable to discern these models at their MCLEs.

Under the highest likelihood model of standing variation in the North, and an independent mutation at the same locus in the South, we obtain the maximum composite log-likelihood estimate of the minimum age of the standing variant, t, of eight generations (Figure 10a). From simulating under this model at the MCLE, we obtain a 95% parametric-bootstrap confidence interval for t of [5, 310] generations. Thus, under the standing-source model, the allele has only been standing for a very short time independently in the northern populations prior to selection. This is consistent with our observed overlap for the standing variant model and migration model. The confidence interval for t does not include 0, but that is also consistent with simulations under the migration model, where inferred standing times are often slightly above zero (Figure 7c and Figure S12 in File S1). Together, these results again suggest we are unable to differentiate between the models where the southernmost tolerant population has an independent mutation, and the three northern populations are sharing the beneficial allele, either via migration or selection on the same young standing variant.

We see partial confounding of the strength of selection and the frequency of the standing variant (Figure 10b), but our results indicate selection has been very strong (> 0.3), and the allele was initially at a very low frequency ($< 10^{-6}$). For the migration in the North model, we obtain similar MCLE of s of 0.4. Lastly, both the standing variation or migration in the North models has the highest composite log-likelihood when the source population of the standing variant is T3, the southernmost population sampled in the North (standing variation composite log-likelihood = 547,060, migration composite log-likelihood = 537,744), but this model may not be distinguishable from that where the source is T2 (standing variation composite log-likelihood = 545,580, migration composite log-likelihood = 533,426).

Discussion

In this paper, we have presented a novel approach to identify the loci involved in convergent adaptation, and to distinguish among the three ways genotypic convergence can arise: selection on (1) independent mutations, (2) a variant standing independently in the selected populations, and (3) beneficial alleles introduced via migration. We leverage the effects selection has on linked neutral sites via a coalescent-based model approach that captures many of the heuristics that have been used in previous studies. This approach also allow us to potentially distinguish between more subtle models, such as the origin and the direction of gene flow of a beneficial allele, since they are explicitly modeled in our framework. Our approach takes advantage of information among all of the population samples simultaneously, while accounting for population structure. Therefore, it naturally accommodates information from across multiple samples, rather than just pairs of populations, and thus offers a number of advantages in identifying the mode of convergence over other approaches. We provide the relevant R code for our approach in https://github.com/kristinmlee/dmc.

Distinguishing among models

We have demonstrated that our method is able to accurately distinguish among modes of convergent adaptation, across a relatively wide parameter space, in simulated data. However, we do see some confounding of models in particular regions of parameter space. In particular, we see the patterns generated

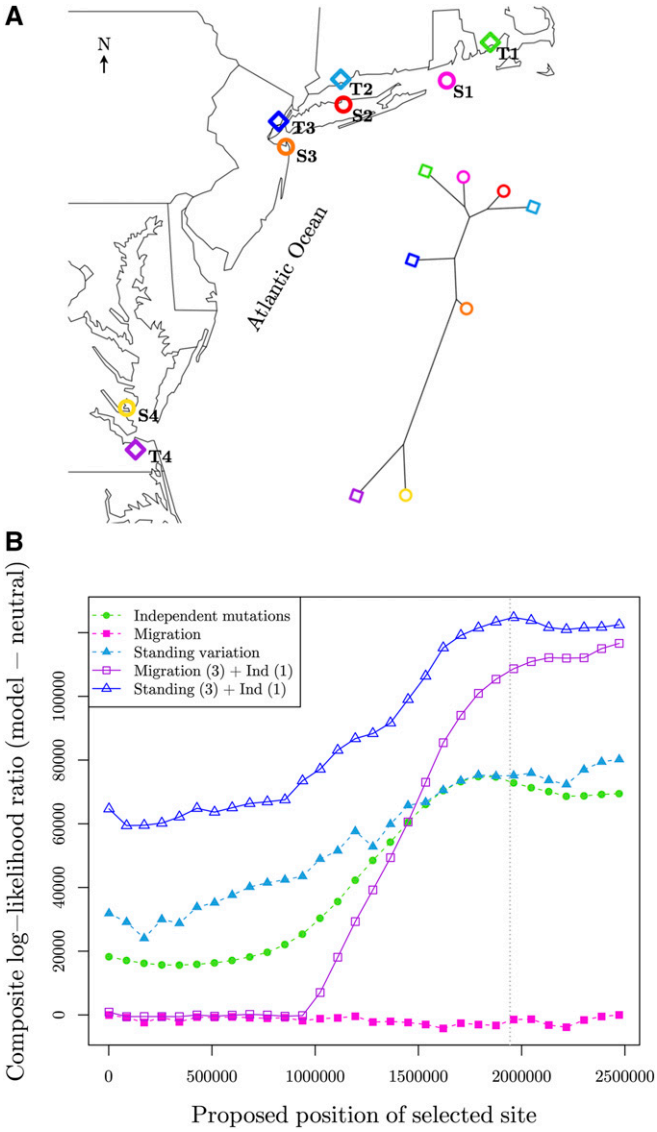

Figure 9 (A) Map of sampled killifish populations with phylogenetic tree, showing that the southern pair (T4, S4) are more distant than other populations. Tree is estimated from genome-wide biallelic SNP frequencies using Phylogeny Inference Package (PHYLIP) Gene Frequencies and Continuous Characters Maximum Likelihood (CONTML) module [see Reid *et al.* (2016) for more information]. (B) Inference results for *Fundulus heteroclitus* pollutant tolerance adaptation on Scaffold9893. Composite log-likelihood ratio of given model relative to neutral model of no selection as a function of the proposed selected site. Closed points represent models where all four populations have same convergent mode, while open points represent Southern population (T4) having an independent mutation at the proposed selected site. We show likelihoods maximizing over possible sources, but all results can be seen in Figure S9 in File S1. The AIP locus position is marked by the vertical, dashed gray lines.

from a model of selection on ancestral standing variation can look like our expectations for the other two modes of convergent adaptation for extreme values of the parameter t, the time the beneficial allele has been standing time independent in the selected populations.

When t is small, we see confounding between the standing model and a model of convergence due to gene flow. The two models are very similar since in our standing variation model, as $t \to 0$, the covariance in the deviations of a neutral allele between selected populations approaches the variance within a selected population. The strong overlap in models is especially true when we have a source for the standing variant. Intuitively, this indicates that the beneficial allele is on a haplotype that is mostly shared among the selected populations. This can be due to a very young standing variant shared among very closely related populations from an ancestral population, a standing variant that was shared by gene flow before selection, or by the selected haplotype quickly moving across populations by gene flow after selection began [which are all closely related models, see Welch and Jiggins (2014), for additional discussion].

To illustrate distinguishing between these possibilities, we now briefly revisit our applications. The Northern tolerant killifish populations, under a standing variation model with gene flow prior to selection, have a very low estimate of the standing time t (eight generations with 95% CI [5, 310] generations). However, given this very low estimate of t, the allele cannot have been standing since the common ancestral population of T1, T2, and T3 (which we estimate to coalesce >800,000 generations ago, assuming no migration, using the estimation procedure outline in Appendix A.3.1). Therefore, the allele must be shared by gene flow among the three populations, and it seems likely that the migration of the allele occurred either after selection began in one of the populations, or very shortly before, with our parametric-bootstrapping approach suggesting we are not able to discern these two models. Interestingly, Reid *et al.* (2016) find no clear signals of admixture from migration elsewhere in the genome between Northern tolerant populations, suggesting that the migration of this allele might be a rare event, although we note that this may reflect a lack of power to detect gene flow.

The case for adaptation from ancestral standing variation is more clear for the *Mimulus* copper tolerance example. Here, the estimate of t is much >0 (646 generations with 95% CI [364, 9525] generations), and, indeed, older than the putative selection pressure (~150 generations ago). Additionally, the standing variant model considerably outperforms the other models, and the results of our parametric-bootstrapping approach support this. In this case, we again favor the model that incorporates gene flow prior to selection on standing variation. The level of neutral differentiation of the mine populations very likely reflects much >646 generations of drift (see Appendix A.3.1); thus, it seems likely that this allele is shared between the mine populations by gene flow, but that the allele was standing in both populations for some time before selection began. Together, these applications show distinguishing among models of convergence is possible in some cases, but may require extra knowledge of population history to aid our inference and understanding.

Conversely, when t is large, we see a collapse of our standing model onto a model of convergence due to independent mutations in our selected populations. This intuition holds

forward in time since as $t \to \infty$ generations, recombination in our isolated populations independently breaks down the similarity of the haplotypes carrying the beneficial mutation. Thus, when selection for the standing variant begins, even tightly linked, hitchhiking neutral alleles will not be shared between populations more than expected by chance. This is also the case when beneficial alleles arise multiple times independently. For example, in the case of the killifish, it is formally possible that the signal of independent selection in the Southern tolerant population is actually due to a very old standing variant shared with the Northern populations, where there is almost no overlap between the Southern and Northern tolerant populations in the haplotype, the selected allele is present on, even close to the selected site. As the precise functional variant(s) in this swept region are currently unknown (Reid *et al.* 2016), it is hard to totally rule out this very old standing variant hypothesis. In other cases, it may be possible to rule out the standing variant hypothesis with very large parameter estimates of t if we know more about the population histories (*i.e.*, our selected populations split more recently than the standing time). Additionally, it may be possible to totally rule out the standing variant hypothesis in cases where if the functional variants can be tracked down to clearly independent genetic changes (*e.g.*, Tishkoff *et al.* 2007). However, that degree of certainty may be difficult to achieve in many cases.

Extendibility and flexibility of our approach

We show the applicability of our method on two empirical examples of convergent adaptation: the evolution of copper tolerance in *M. guttatus* and of pollutant tolerance in *F. heteroclitus*. The latter exemplifies the extendibility and flexibility of our approach. As the number of selected populations increase, our potential number of hypotheses grows, since any grouping of two or more populations could share selection due to migration or standing variation. Additionally, with more populations, we have more potential sources of the beneficial allele in the migration model. Our model could also be extended to have selection occurring in some of the adapted populations and the neutral model in others, to identify genomic regions that are not experiencing convergent adaptation among all populations sharing the selected environment. These models are all relatively easy to implement into our framework; however, the sheer number of possible hypotheses as the number of populations grows will likely call for some more systematic way of implementing these models and exploring their relationships.

Caveats and possible extensions

Studying repeated evolution has long played a key role in evolutionary biology as a tool to help identify the ecological and molecular basis of adaptation. It is worth noting that, with this approach, we are able to identify sweeps in the same region, and whether they appear to be shared or independent.

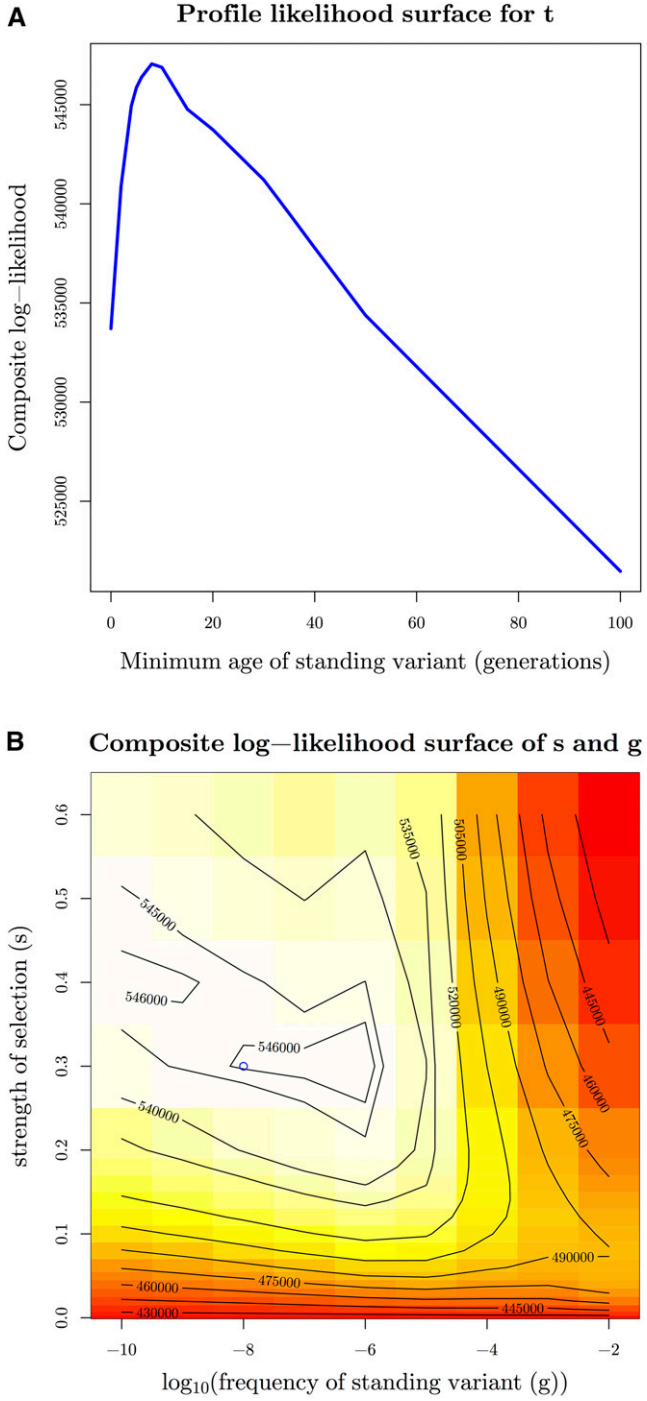

Figure 10 The composite log-likelihood surfaces for the parameters for *F. heteroclitus* convergent data in combined standing variation, and independent sweep model with position 1,961,198 on Scaffold9893 as selected site and population T3 as source. (A) Profile composite log-likelihood surface for minimum age of standing variant, maximizing over other parameters, showing the beneficial allele has been standing for a very short amount of time in our three northern populations (eight generations). (B) Composite log-likelihood surface for strength of selection *vs.* frequency of standing variant. Blue circle represents point estimate of joint MCLE ($\hat{s} = 0.3$, $\hat{g} = 10^{-8}$). t is held at MCLE of eight generations.

However, in the scale of an entire genome, it may be possible for two, functionally unrelated sweeps to overlap. In the case of adaptation via independent mutations across multiple populations, it is especially hard to determine whether selection at the same site was acting on the same phenotype. It is potentially more plausible to claim that the phenotype and selection pressure are shared among populations in cases where the swept haplotype is shared. Ultimately, in demonstrating convergence, we will have to rely on a range of evidence. Shared sweeps can offer one substantial piece of evidence, particularly when we are studying recent adaptation to a strong selective pressure that is distinct to the adapted populations.

In addition to assuming that the same locus is under selection in all adapted populations, we assume a single selected change underlies the sweep within a population, and that recombination is free to break down associations between neutral alleles and this selected variant. If, for instance, selection acts on an epistatic, haplotypic, combination of allele that sweeps, a long haplotype could be shared between populations not due to recent migration but because selection acts against recombinants breaking up the haplotype (Kelly and Wade 2000). Convergent adaptations due to shared inversions also violate the assumptions of our method. Inversions can repress recombination across the entire inversion [see Kirkpatrick (2010), for a recent review]. Inversions significantly alter both neutral and selective model expectations (*e.g.*, Guerrero *et al.* 2012) and could lead to long shared haplotypes among populations even if the shared inversion is old. It may be possible to use our approach to model the decay in coancestries outside of the inverted region, but this requires knowledge of the inversion and its break points *a priori* and a detailed knowledge of recombination rates surrounding the inversion.

Throughout this paper, we assume that the sweeps have fixed recently, and it will be important to relax this assumption. In these cases, models of migration that include selection against maladaptive migrants (Barton and Bengtsson 1986; Charlesworth *et al.* 1997; Roesti *et al.* 2014) will be important to consider. Long-term selection against migrant alleles (*i.e.*, due to local adaptation) lowers the effective migration rate at linked neutral sites, and so will distort the covariance relationships among populations (and may, in some cases, confound the signal of the mode of convergence). These deviations could be incorporated into our models, allowing us to perform inference under these models. However, in practice, we would likely be underpowered, as we only model segregating sites we cannot (in the current framework) fully account for selection that deepens the absolute divergence among particular populations.

Additionally, our framework could be extended both to leverage more information and to model other biologically relevant scenarios. Here, we analyze genomic regions that we *a priori* assume to be under convergent selection. These regions were identified on the basis of the phylogenetic relationships among the populations, with convergent populations being non-sister. This approach, however, does not take advantage of the flexibility of our framework. We are working on extensions to efficiently scan genome-wide data for genomic regions exhibiting convergence. In this case, we can potentially identify regions shared among populations that we may not have been able to previously identify via traditional approaches. Additionally, if these populations are sister to one another, our method can be extended to test whether this is convergent or whether the sister populations share an allele due to selection in their ancestor (Racimo 2016).

Acknowledgments

We wish to thank members of the Coop lab for helpful discussion and feedback on earlier drafts. We'd also like to gratefully acknowledge Noah Reid, Andrew Whitehead, John Willis, and Kevin Wright sharing their data and thoughtful comments. We thank Nicolas Bierne, Joachim Hermisson, and an anonymous reviewer for valuable suggestions on an earlier draft. This work was supported by the National Science Foundation Graduate Research Fellowship awarded to K.M.L. (1148897) and by grants from the National Science Foundation under grant no. 1353380 to John Willis and G.C., and the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award numbers NIH R01 GM108779 awarded to G.C.

Literature Cited

Arendt, J., and D. Reznick, 2008 Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol. Evol. 23: 26–32.

Aubury, L. E., 1902 *The Copper Resources of California (No. 23)*. Superintendent State Printing, Sacramento, CA.

Barrett, R. D., and D. Schluter, 2008 Adaptation from standing genetic variation. Trends Ecol. Evol. 23: 38–44.

Barton, N., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. 72: 123–133.

Barton, N., and B. O. Bengtsson, 1986 The barrier to genetic exchange between hybridising populations. Heredity 57: 357–376.

Berg, J. J., and G. Coop, 2015 A coalescent model for a sweep of a unique standing variant. Genetics 201: 707–725.

Bierne, N., 2010 The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. Evolution 64: 3254–3272.

Bierne, N., P.-A. Gagnaire, and P. David, 2013 The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Curr. Zool. 59: 72–86.

Chan, Y. F., M. E. Marks, F. C. Jones, G. Villarreal, Jr., M. D. Shapiro *et al.*, 2010 Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitx1 enhancer. Science 327: 302–305.

Charlesworth, B., M. Nordborg, and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. 70: 155–174.

Chen, H., N. Patterson, and D. Reich, 2010 Population differentiation as a test for selective sweeps. Genome Res. 20: 393–402.

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villarreal, Jr., M. Dickson *et al.*, 2005 Widespread parallel evolution in

sticklebacks by repeated fixation of ectodysplasin alleles. Science 307: 1928–1933.
Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics 185: 1411–1423.
DeGiorgio, M., K. E. Lohmueller, and R. Nielsen, 2014 A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet. 10: e1004561.
Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor Popul Biol. 66: 129–138.
Duvernell, D. D., J. B. Lindmeier, K. E. Faust, and A. Whitehead, 2008 Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, Fundulus heteroclitus. Mol. Ecol. 17: 1344–1360.
Ewens, W., 2004 *Mathematical Population Genetics 1: Theoretical Introduction*, *Interdisciplinary Applied Mathematics*. Springer, New York.
Gillespie, J. H., 2000 Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155: 909–919.
Guerrero, R. F., F. Rousset, and M. Kirkpatrick, 2012 Coalescent patterns for chromosomal inversions in divergent populations. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367: 430–438.
Harvey, P. H., and M. D. Pagel, 1991 *The Comparative Method in Evolutionary Biology*, Vol. 239. Oxford University Press, Oxford.
Hedrick, P. W., 2013 Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol. Ecol. 22: 4606–4618.
Heliconius Genome Consortium, 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487: 94–98.
Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics 159: 1805–1817.
Hudson, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics 120: 831–840.
Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli *et al.*, 2012 The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55–61.
Kaplan, N., R. R. Hudson, and M. Iizuka, 1991 The coalescent process in models with selection, recombination and geographic subdivision. Genet. Res. 57: 83–91.
Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics 123: 887–899.
Kelly, J. K., and M. J. Wade, 2000 Molecular evolution near a two-locus balanced polymorphism. J. Theor. Biol. 204: 83–101.
Kim, Y., and T. Maruki, 2011 Hitchhiking effect of a beneficial mutation spreading in a subdivided population. Genetics 189: 213–226.
Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513–1524.
Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.
Kirkpatrick, M., 2010 How and why chromosome inversions evolve. PLoS Biol. 8: e1000501.
Larribe, F., and P. Fearnhead, 2011 On composite likelihoods in statistical genetics. Stat. Sin. 21: 43–69.
Lee, Y. W., 2009 Genetics analysis of standing variation for floral morphology and fitness components. Ph.D. Thesis, Duke University, Durham, NC.
Lipson, M., P.-R. Loh, A. Levin, D. Reich, N. Patterson *et al.*, 2013 Efficient moment-based inference of admixture parameters and sources of gene flow. Mol. Biol. Evol. 30: 1788–1802.
Losos, J. B., 2011 Convergence, adaptation, and constraint. Evolution 65: 1827–1840.
MacNair, M. R., S. E. Smith, and Q. J. Cumbes, 1993 Heritability and distribution of variation in degree of copper tolerance in Mimulus guttatus at Copperopolis, California. Heredity 71: 445–455.
Martin, A., and V. Orgogozo, 2013 The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution 67: 1235–1250.
Maynard Smith, J., 1971 What use is sex? J. Theor. Biol. 30: 319–335.
Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.
Nacci, D., L. Coiro, D. Champlin, S. Jayaraman, R. McKinney *et al.*, 1999 Adaptations of wild populations of the estuarine fish Fundulus heteroclitus to persistent environmental contaminants. Mar. Biol. 134: 9–17.
Nacci, D. E., D. Champlin, and S. Jayaraman, 2010 Adaptation of the estuarine fish Fundulus heteroclitus (Atlantic killifish) to polychlorinated biphenyls (PCBs). Estuaries Coasts 33: 853–864.
Nicholson, G., A. V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson *et al.*, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. R. Stat. Soc. Series B Stat. Methodol. 64: 695–715.
Nielsen, R., S. Williamson, Y. Kim, M. Hubisz, A. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. 15: 1566–1575.
Orr, H. A., 2005 The probability of parallel evolution. Evolution 59: 216–220.
Pearce, R. J., H. Pota, M.-S. B. Evehe, E.-H. Bâ, G. Mombo-Ngoma *et al.*, 2009 Multiple origins and regional dispersal of resistant dhps in African Plasmodium falciparum malaria. PLoS Med 6: e1000055.
Pease, J. B., D. C. Haak, M. W. Hahn, and L. C. Moyle, 2016 Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14: e1002379.
Pennings, P. S., and J. Hermisson, 2006 Soft sweeps ii—molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. 23: 1076–1084.
Przeworski, M., G. Coop, and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. Evolution 59: 2312–2323.
Racimo, F., 2016 Testing for ancient selection using cross-population allele frequency differentiation. Genetics 202: 733–750.
Racimo, F., S. Sankararaman, R. Nielsen, and E. Huerta-Sánchez, 2015 Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. 16: 359–371.
Reid, N. M., D. A. Proestou, B. W. Clark, W. C. Warren, J. K. Colbourne *et al.*, 2016 The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science 354: 1305–1308.
Roesti, M., S. Gavrilets, A. P. Hendry, W. Salzburger, and D. Berner, 2014 The genomic signature of parallel adaptation from shared genetic variation. Mol. Ecol. 23: 3944–3956.
Rosenzweig, B. K., J. B. Pease, N. J. Besansky, and M. W. Hahn, 2016 Powerful methods for detecting introgressed regions from population genomic data. Mol. Ecol. 25: 2387–2397.
Samanta, S., Y.-J. Li, and B. S. Weir, 2009 Drawing inferences about the coancestry coefficient. Theor. Popul. Biol. 75: 312–319.
Santiago, E., and A. Caballero, 2005 Variation after a selective sweep in a subdivided population. Genetics 169: 475–483.
Schluter, D., and G. L. Conte, 2009 Genetics and ecological speciation. Proc. Natl. Acad. Sci. USA 106: 9955–9962.
Slatkin, M., and T. Wiehe, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. 71: 155–160.
Song, Y., S. Endepols, N. Klemann, D. Richter, F.-R. Matuschka *et al.*, 2011 Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Curr. Biol. 21: 1296–1301.

Stern, D. L., 2013 The genetic causes of convergent evolution. Nat. Rev. Genet. 14: 751–764.
Thompson, E. A., 2013 Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194: 301–326.
Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt *et al.*, 2007 Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39: 31–40.
Turner, T., E. Bourne, E. V. Wettberg, T. Hu, and S. Nuzhdin, 2010 Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat. Genet. 42: 260–263.
Varin, C., N. Reid, and D. Firth, 2011 An overview of composite likelihood methods. Stat. Sin. 21: 5–42.
Weir, B. S., and W. G. Hill, 2002 Estimating F-statistics. Annu. Rev. Genet. 36: 721–750.
Welch, J. J., and C. D. Jiggins, 2014 Standing and flowing: the complex origins of adaptive variation. Mol. Ecol. 23: 3935–3937.
Wiuf, C., 2006 Consistency of estimators of population scaled parameters using composite likelihood. J. Math. Biol. 53: 821–841.
Wood, T. E., J. M. Burke, and L. H. Rieseberg, 2005 Parallel genotypic adaptation: when evolution repeats itself. Genetica 123: 157–170.
Wright, K. M., U. Hellsten, C. Xu, A. L. Jeong, A. Sreedasyam *et al.*, 2015 Adaptation to heavy-metal contaminated environments proceeds via selection on pre-existing genetic variation. bioRxiv. doi: https://doi.org/10.1101/029900.
Wright, S., 1943 Isolation by distance. Genetics 28: 114.
Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354.

Communicating editor: J. Hermisson

Appendix A

A.1 Coalescent Interpretation of Covariances and F-Matrix Estimation

Let x_{il} be the allele frequency of allele 1 in population i at locus l, and that the frequency of this allele in the ancestral population is ϵ_l. Consider the covariance $\mathrm{Cov}(\Delta x_{il}, \Delta x_{jl})$ over replicates of the drift processes at locus l. We can write

$$\mathrm{Cov}\Big[(x_{il}-\epsilon_l),\Big(x_{jl}-\epsilon_l\Big)\Big] = \mathbb{E}\Big[(x_{il}-\epsilon_l)\Big(x_{jl}-\epsilon_l\Big)\Big] \tag{A.1}$$

$$= \mathbb{E}\Big[x_{il}x_{jl}\Big] - \epsilon_l^2, \tag{A.2}$$

which follows from the fact that $\mathbb{E}[x_{il}] = \mathbb{E}[x_{jl}] = \epsilon_l$. We can interpret $\mathbb{E}[x_{il}x_{jl}]$ as the probability that we sample a single allele in i and an allele in j, and that they both are of type 1. Taking that interpretation, assuming that there is no mutation, $\mathbb{E}[x_{il}x_{jl}]$ is the probability that, tracing back a coalescent lineage from i and a lineage from j, both lineages trace back to type 1 alleles in the ancestral population. Let our pair of lineages drawn from i and j coalesce with probability f_{ij}. If our lineages coalesce before reaching the ancestral population, then they will be identical by descent, and share the ancestral choice of allele. Therefore, we can write

$$\mathbb{E}\Big[x_{il}x_{jl}\Big] = \Big(1-f_{ij}\Big)\epsilon_l^2 + f_{ij}\epsilon_l \tag{A.3}$$

Then, we can rewrite the covariance

$$\mathrm{Cov}\big(\Delta x_{il}, \Delta x_{jl}\big) = f_{ij}\epsilon_l(1-\epsilon_l), \tag{A.4}$$

and, for the variance, we set $i = j$. Thus, under a model of genetic drift alone, we can interpret the entries of our covariance matrix as expressions of the underlying coalescent probabilities.

Estimating F

In the main text, we assume that we have estimates of our neutral coancestry matrix **F**. We now describe how we obtain these. From above, Equation A.3, the expectation of $x_{il}x_{jl}$ across loci is

$$\mathbb{E}_l\Big[x_{il}x_{jl}\Big] = \mathbb{E}_l\Big[\Big(1-f_{ij}\Big)\epsilon_l^2 + f_{ij}\epsilon_l\Big] \tag{A.5}$$

Therefore, we can write estimate f_{ij} as

$$f_{ij} = \frac{\mathbb{E}_l\big[x_{il}x_{jl}\big] - \mathbb{E}_l\big[\epsilon_l^2\big]}{\mathbb{E}_l[\epsilon_l(1-\epsilon_l)]} \tag{A.6}$$

We can obtain an unbiased estimate of $\mathbb{E}_l[\epsilon_l^2]$ and $\mathbb{E}_l[\epsilon_l(1-\epsilon_l)]$ using the sample allele frequencies from two populations on either side of the root of the population phylogeny (see Supplement of Lipson *et al.* 2013). Let i' and j' be a pair of populations that span the root of the population tree, then we can use the estimate

$$\mathbb{E}_l[\epsilon_l(1-\epsilon_l)] = \mathbb{E}_l\left[\frac{1}{2}x_{i'l}\Big(1-x_{j'l}\Big) + \frac{1}{2}(1-x_{i'l})\Big(x_{j'l}\Big)\right] \tag{A.7}$$

Likewise, we use the estimate

$$\mathbb{E}_l\big[\epsilon_l^2\big] = \mathbb{E}_l\left[\frac{1}{2}x_{i'l}\Big(x_{j'l}\Big) + \frac{1}{2}(1-x_{i'l})\Big(1-x_{j'l}\Big)\right] \tag{A.8}$$

An estimate of the term $\mathbb{E}_l[x_{il}x_{jl}]$ can be obtained by using the sample frequency of allele 1 in populations i and j. However, as we only have a sample from the population frequency, we need to account for the finite sampling bias within populations ($i = j$). Let n be the sample size in population i, then

$$f_{ii} = \frac{\mathbb{E}_l\left[x_{il}^2\right]\frac{n}{n-1} - \mathbb{E}_l[x_{il}]\frac{1}{n-1} - \mathbb{E}_l\left[\epsilon_l^2\right]}{\mathbb{E}_l[\epsilon_l(1-\epsilon_l)]} \tag{A.9}$$

where our x are now sample frequencies. There is no finite-sample size correction for f_{ij}, $i \neq j$, and Equation A.6 can be used directly.

In our simulations to show the effect of selection on the coancestry coefficients (Figure 3), we estimate f_{ij} in bins of fixed recombination distance moving away from the selected site. We do this by approximating the expectations in the numerator and denominators in Equations A.6 and A.9 by the average of the expression over all of the SNPs that fall in a given genetic distance bin over all of the relevant simulations. To account for biases induced by defining the allele of interest, we randomize the reference allele at each SNP.

A.2 Simulation Implementation Details

We perform coalescent simulations using mssel, a modified version of ms (Hudson 2002) that allows for the incorporation of selection at single site (the code for this is provided in https://github.com/kristinmlee/dmc). The program allows the user to specify the frequency trajectory of the selected allele through time across populations; this trajectory is then used to simulate genetic data under the coalescent model conditioning on this trajectory [using the subdivided coalescent model (Hudson and Kaplan 1988; Kaplan *et al.* 1991)]. We generate stochastic trajectories for the selected allele across populations, and describe the simulation process below. We simulate multiple instances of the stochastic trajectories, and average our results across datasets generated for these trajectories. We focus on a set of four populations with relationships as shown in Figure 1. Populations 2 and 3 are adapted to a shared novel selection pressure, and populations 1 and 4 are in the ancestral environment.

The original implementation of mssel assumes only a single origin of the selected allele, which occurs moving backward in time when the frequency of the derived allele goes to zero in the final population it segregates in. We modified the mssel source code directly to accommodate multiple origins of the selected allele as is necessary in the independent sweep model. We do so by allowing an independent origin of the selected allele in any population where the frequency of the derived selected allele goes to zero, if that population currently has a migration rate of zero to any other population containing the selected allele.

A.2.1 Generating stochastic trajectories for the selected allele

We generate stochastic trajectories for the selected allele to be used as input for mssel to generate sequence data for given convergent adaptation scenarios. We simulate the allele frequency trajectory for the selected allele forward in time using a normal deviate approximation to the simulation the Wright-Fisher diffusion. Specifically, given the frequency of the beneficial allele at time t, $X(t)$, we simulate its frequency at time $t + \Delta t$ according to

$$X(t+\Delta t) \sim N\Big(\mu_S\Big(X(t)\Big)\Delta t, \sigma^2\Big(X(t)\Big)\Delta t\Big), \tag{A.10}$$

where $\mu_S(\,)$ and $\sigma^2(\,)$ are the infinitesimal mean and variance of the Wright-Fisher diffusion. We set $\Delta t = 1/(2N)$, representing one Wright-Fisher generation on the diffusion time-scale ($2N$ generations). We set $X(0) = g$, the initial frequency of the beneficial allele. When selection starts from a new mutation, $g = 1/(2N)$.

For all our models, the infinitesimal variance is

$$\sigma^2(X(t)) = X(t)(1 - X(t)), \tag{A.11}$$

representing the effect of genetic drift.

For populations not impacted by migration, we condition our trajectory on the beneficial allele going to fixation forward in time. To do this, we use the conditional infinitesimal mean

$$\mu_S(X(t)) = \frac{2NsX(t)(1-X(t))}{tanh(2NsX(t))} \tag{A.12}$$

(see Przeworski *et al.* 2005; Berg and Coop 2015, for previous applications). We simulate this process forward in time till fixation is reached. Given that we are assuming the sweeps completely recently, we have fixation occur at time zero, so that the time of a new mutation is determined by the time of the sweep.

Migration model: In the case of our migration model, there is one way migration from population i into j. The trajectory of X_i is simulated first forwards in time, conditioning on fixation, using the above approach. We then simulate the frequency in population j starting from $X_j(0) = 0$, with the infinitesimal mean

$$\mu_S\big(X_j(t)\big) = 2NsX_j(t)\big(1 - X_j(t)\big) + 2Nm\big(X_i(t) - X_j(t)\big) \tag{A.13}$$

(expanded from Ewens 2004). We simulate the process forward in time until the selected allele reaches fixation in both populations. The first population to reach fixation is held at frequency 1 until the other population fixes for the beneficial allele.

Standing variation model: We define the standing variation trajectory as having three phases: the neutral phase, the standing phase, and the selected phase. To specify a trajectory in which the beneficial allele has been standing at frequency g for time t, we simply hold the allele frequency constant for this amount of time. We simulate a stochastic neutral trajectory of our beneficial allele from frequency g to 0 backward in time according to

$$X(t - \Delta t) \sim N(\mu_N(X(t))\Delta t, \sigma(X(t))\Delta t) \tag{A.14}$$

using the infinitesimal mean conditional of the neutral allele going to loss

$$\mu_N(X(t)) = -X(t) \tag{A.15}$$

(see Przeworski *et al.* 2005; Berg and Coop 2015, for previous applications). We simulate the selection phase forward in time for $2\log(1/g)/s$ generations. If the beneficial allele has reached fixation before this time, it is held constant at frequency 1 for the remaining time. If not, the trajectory is simply stopped at this time. This allows for the interpretation of the standing time and the time of the onset of selection to be the same throughout simulations. For the whole trajectory of a beneficial allele, we paste together these three components: neutral increase of allele from frequency 0 to g, the standing phase at frequency g for time t generations, and the selective phase. For populations not experiencing selection, the beneficial allele is kept at frequency g for the entire length of the trajectory. We acknowledge this is an untested approximation, but think it has little impact on our results. The frequency of the standing variant matters mostly for estimating the duration of the sweep within populations, so its frequency during this standing phase is not as important as the frequency at the onset of selection. Additionally, we assume that g is small, such that the probability of recombining off onto the other background during this phase is simply r. The frequency of the variant during the standing phase does impact the probability of coalescing before recombination (or vice versa) during this phase, but only weakly.

A.2.2 Details of coalescent simulations

In this section, we give the details of the coalescent simulations. The mssel command lines can be found in Supplement S3 in File S1. The mssel input can be interpreted as follows:

```
./mssel nsam_tot nreps nsam_anc nsam_der trajFile locSelSite -t θ -r ρ nsites -I
npops nAnc_pop1 nDerv_pop1 ... nAnc_popi nDerv_popi
```

For all of the simulations, we generate neutral allele frequency data for 10 samples from each of four populations. The populations are related to each other as shown in Figure 1. Note, we did 1000 replications of the simulations for parameters used to generate comparisons of average simulations coancestry coefficients compared to theoretical expectations; 100 replications were done for simulations used for parameter estimates and model comparisons. For simulations used for both, the first 100 runs were used.

Independent sweep model: We generated beneficial allele frequency trajectories under four different selection coefficients: $s = [0.005, 0.01, 0.05, 0.1]$ under the independent sweep model, with $N_e = 100,000$. We set r, the per generation probability of cross-over between ends of the simulated locus, to 0.005. The neutral mutation rate, μ, for the entire locus is the same as r. We also simulate, with ms the same population structure with no selection to generate data to estimate the neutral coancestry matrix, **F**.

Standing variation model: With $s = 0.01$ and $g = 0.001$, we generated beneficial allele frequency trajectories for standing times $t = [50, 250, 500, 1000,$ and $5000]$ generations under the standing variation model with $N_e = 10,000$. Our t references the time that the populations have been independent. Therefore, we adjusted the split times to ensure that the t of interest

corresponded to the duration of time that the selected populations had the standing variant prior the populations joining in the ancestral population. The population split times were determined to ensure selection started after the populations were completely isolated, and to maintain a similar ratio of time for four independent populations to two ancestral populations. We again set $r = \mu = 0.005$. Again, neutral regions were simulated in ms using the same population structure (*i.e.*, each parameter set had its own neutral data generated).

Migration model: Lastly, we simulated under the migration model with $m = [0.0001, 0.001, 0.01, \text{ and } 0.1]$, holding $s = 0.01$ for $N_e = 10,000$. Again, we simulated 10 samples from four populations related to each other as specified in Figure 1. Now, in mssel, we specify migration to start just prior to origin of the beneficial allele in the source population, and to continue until the sweep has reached fixation (time zero in the past since we fix sweeps to complete at the end). We set population 2 to be the source, and have $4N_em$ migrants from population 2 into population 3 each generation. We again set $r = \mu = 0.005$. Neutral regions were again simulated using ms. Each set of parameters has its own neutral data generated as the migration rate impacts neutral coancestry as well.

A.2.3 Interpreting mssel output

The output from mssel and ms is in the form of haplotypes for each of the sampled chromosomes at polymorphic sites in addition to their positions on a scale of $(0, 1)$. We use this to calculate sample allele frequencies at each site for each population. Prior to performing further estimations or analyses with these neutral allele frequencies, we randomize the reference allele so that there is no bias resulting from which allele was called ancestral or derived. We exclude sites where the average allele frequencies across populations are $<5\%$ or $>95\%$.

A.2.4 Composite likelihoods of simulated data under all models details

We calculated the composite log-likelihoods of each the simulated datasets under all models, including the neutral model, with the same parameter space shown in Table S1 in File S1.

A.2.5 Maximum likelihood estimate of parameters from simulated data under correct model

We also calculated the composite log-likelihoods of each the simulated datasets under the correct model used to generate the data, now with a more dense grid of parameters to obtain better estimates of the MCLE of each parameter. We allowed g to vary in the calculations of the MCLEs under the standing variation model. See Table S2, Table S4, and Table S5 in File S1.

A.2.6 Inference details: mean-centering allele frequencies and covariances, sample size correction, and speed-ups

Given that we do not know the true ancestral mean at locus l, ϵ_l, we use the mean of the present-day sample allele frequencies at this locus, $\overline{x_l} = 1/k\sum_{i=1}^{K} x_{i,l}$. When mean-centering, we lose a degree of freedom, so, in calculating the likelihood, it is necessary to drop information from one population. Since the information from the dropped population is incorporated in the mean, the choice of the dropped population is arbitrary. In matrix form, the mean-centered allele frequencies with one dropped population can be expressed as

$$\overrightarrow{x}_i^{'} = \mathbf{T}\overrightarrow{x}_i \tag{A.16}$$

where $\mathbf{T}$ is an $K-1$ by K matrix with $K-1/K$ on the main diagonal and $-1/K$ elsewhere. Prior to mean-centering, we randomize the reference allele at each SNP to account for biases induced by defining the allele of interest.

Now, we model the mean-centered allele frequencies as multivariate normal around mean zero, with covariance proportional to a mean-centered parameterized covariance matrix ($\mathbf{F}^{(S)'}$) as

$$\overrightarrow{x_l}' \sim \mathcal{N}\left(\overrightarrow{0}, \overline{x_l}(1-\overline{x_l})\mathbf{F}^{(S)'}\right), \tag{A.17}$$

where we use the average present day allele frequency across populations at the locus, $\overline{x_l}$, as an estimate of ϵ_l in the site-specific term in the covariance. We note that $\overline{x_l}(1-\overline{x_l})$ is a slightly downwardly biased estimate of $\epsilon(1-\epsilon)$, but, for our purposes, it seems sufficient to include this term as a locus-specific adjustment to the expected covariance.

To obtain the corresponding mean-centered covariance matrix, dropping the same population, we can apply the following matrix operations,

$$\mathbf{F}^{(S)'} = \mathbf{T}\mathbf{F}^{(S)}\mathbf{T}^{\top}. \tag{A.18}$$

this new matrix is $K-1$ by $K-1$ and full rank.

Before mean-centering, $\mathbf{F}^{(S)}$, we apply a sample size correction to correct for the finite sampling bias. We add $1/n_i$ to the diagonal where n_i is the sample size in population i. We take twice the number of diploid individuals sampled in population i as n_i for data applications. In simulations, we use the number of chromosomes sampled in population i as n_i. Note that both this mean-centering and sample size correction is also performed on the neutral matrix, $\mathbf{F}$ before likelihood calculations under a neutral model with no selection.

To decrease some of the computational time involved in our likelihood calculations, we precompute the mean-centered covariance matrices with selection, $\mathbf{F}^{(S)'}$, for given bins of distance away from a putative selected site. We first divide our distances in our window into 1000 bins, and take the midpoint of the distances in these bins to calculate $\mathbf{F}^{(S)'}$, as this matrix is a function of distance. To avoid the costly step of recomputing the corresponding inverses, and determinants needed for likelihood calculations, we do this step first, and use these values for all SNPs in a given bin, and store them and reuse them over all locations of the selected site.

Thus, we calculate the likelihood of mean-centered allele frequencies, $\overrightarrow{x_l}'$, given our model M and its parameters Θ_M, a given locus l as

$$P\left(\overrightarrow{x_l}' \middle| \mathbf{F}^{(S)'}(r_l, M, \Theta_M)\right) = \frac{\exp\left(-\frac{1}{2}\overrightarrow{x_l}'^{\top}\left(\mathbf{F}^{(S)'}\right)^{-1}(\overline{x_l}(1-\overline{x_l}))^{-1}\overrightarrow{x_l}'\right)}{\sqrt{2\pi^k(\overline{x_l}(1-\overline{x_l}))^k \det \mathbf{F}^{(S)'}}} \tag{A.19}$$

where $k = K - 1$, the rank of matrix $\mathbf{F}^{(S)'}$.

A.3 Parametric Bootstrapping Approach Details

To carry out the parametric-bootstrapping approach, we again perform coalescent simulations using mssel for simulations with selection and ms for neutral simulations. We specify the number of populations and the sample size for each populations (twice the number of individuals sampled). Now, instead of specifying θ, we specify the number of segregating sites as the number of SNPs in our window of interest. We also simulate with the same population-scaled recombination rate and number of sites between which recombination can occur as the number of base pairs in our analysis window. To match the population-scaled recombination rate, we take the genetic map of our region r and scale it to be $4N_e r$, assuming that recombination is uniformly distributed over our region. We down-scaled the effective population size for computational efficiency in the generation of the simulations, which impacts both ρ, and the times in the trajectories of the beneficial allele, by a linear rescaling. Additionally, we specify the location of the selected site (ℓ) to be at the MCLE of the model used for simulation.

While, in the rest of the paper, we make use of stochastic trajectories, for the parametric-bootstrap simulations, we generated deterministic trajectories of the selected allele to be used as input for mssel. This is because we need to set our simulations up to accommodate both the MCLE selection coefficient and the coalescent times within and between populations, which is somewhat fiddly to automate with fully stochastic trajectories across all the models. Now, we fix the time of the sweep to be

$$\frac{1}{s}\log\left(\frac{p_{t_s}q_0}{q_{t_s}p_0}\right) \tag{A.20}$$

where p_0, the frequency of the beneficial allele at time 0, is $1/2N$ for a new mutation or g for the standing variant model, while p_{t_s}, the frequency of the beneficial allele at fixation, is set to 0.999. For the migration model, we start this trajectory (from $1/2N$) after the delay time (Equation 10) for recipient population(s). We simulate with migration after δ for a few generations. For the standing variant model with a source population, we start the selected allele trajectory (from frequency g) in the recipient population(s) after t generations. We simulate with a brief burst of migration at time t until the frequency of the beneficial allele goes to 0 in the recipient population(s), at a very low rate. This forces an instantaneous coalescent event back into our source population. The parameters (s, t, g, m, and the source population) are all set to the MCLE of the corresponding model.

We simulate each convergent and neutral model 100 times, and interpret the output and calculate the likelihood of our simulated data (as detailed in Appendix A.2) under the model used for simulations and the model with the largest composite likelihood for the original data. The mssel command lines can be found in Supplement S4 in File S1.

A.3.1 Approximating demography given a neutral F matrix

For the parametric bootstrap, we need to simulate under a model of population structure that approximately matches that in our data. To do so, we assume that our sampled populations are related through a bifurcating population phylogeny (with no neutral

migration). While this is a crude approximation, it allows us a good match to the observed F matrix of the data, and considerably simplifies the task of setting up the simulations. In practice, since our method works with these covariances, and, inferring the details of population structure is not our primary concern here, we view this as an acceptable compromise.

For simulating under the approximate population structure in our data, we need to estimate join times for population pairs. We use

$$f_{ij} \approx 1 - e^{-t_{ij}^{\text{coal}}}, \tag{A.21}$$

where t_{ij}^{coal} is in coalescent time units to approximate the shared branch length between populations i and j, assuming no migration. Migration will impact the coancestry coefficients, and, thus, our interpretations of the coalescent times. For example, migration between two populations will increase their relatedness, and can make their shared branch length appear longer. We also use this approximation to compare the split time between populations to the standing time for our adaptive alleles t, to judge whether they could have been standing for a given time between two populations, or if migration must be invoked.

To generate join times, we first solve for all t_{ij}^{coal} using A.21 from an estimated neutral **F** matrix. We find populations i and j with the largest t_{ij}^{coal}. We approximate the join time as the average of the differences between the total time associated with each population (*i.e.*, t_{ii}^{coal} and t_{jj}^{coal}) and the time between them (t_{ij}^{coal}). This follows from assuming that drift is acting additively, such that $f_{ii} \approx f_{ij} + f_i$, where f_i is the coancestry coefficient associate with population i in isolation (see Supplement S2 in File S1 for more). We then effectively join these two populations, updating all t_{ik}^{coal} and t_{jk}^{coal}, where k is any unjoined population to be the average of t_{ik}^{coal} where k, and t_{jk}^{coal} where k. We repeat this procedure, joining the two remaining populations with the largest t_{ij}^{coal} until all populations are joined. From this, we are able to specify join times for simulations that capture the general population structure of a given **F** matrix.

The population structure used for simulation is now represented in a bifurcating tree, which may fail to capture of the complexity represented in a given **F** matrix. Thus, when performing the composite-likelihood calculations, we use a modified **F** matrix estimated using the procedure detailed in A.1 with neutral data simulated with these join times, to parameterize our models.

Additionally, these estimates for the between-population coalescent times, assuming no migration and a bifurcating tree, can give us insight that it is possible for the beneficial allele to have been standing for a given $\hat{t}$ since the ancestral population, or whether it is necessary to invoke the model where migration has a role in spreading the beneficial allele prior to it standing. For example, in our *Mimulus* analysis, we estimate our join time to be 0.050 in coalescent units. Our MCLE for t under the classic standing model is 434 generations, or 0.00029 coalescent units, which is much shorter than the time in which our selected populations coalesce. We caution against assigning too much value to these inferences, given the assumptions, but do find these approximations to be broadly useful.

A.4 Standing Variant Model with a Source Population

When there are multiple selected populations, and they do not follow a bifurcating tree structure, it is necessary to incorporate a model that has a source population for the standing variant to have self-consistent mean-centered covariance matrices.

Let population l be a selected population and the source of the beneficial allele. In all other populations, the beneficial allele is standing for time t generations at frequency g before the lineage returns to the source population, where it still standing at frequency g (see Figure 11). We can define pairwise coancestry coefficients for all pairs of populations under this model. Let populations i and j represent populations that experience selection and population k be any unselected population.

Since population l is the source, its variance follows the same form as Equation 7.

$$f_{ll}^{(S)} = y^2\left(\frac{1}{1+4N_e rg} + \frac{4N_e rg}{1+4N_e rg} f_{ll}\right) + (1-y^2) f_{ll} \tag{A.22}$$

All other selected populations have a modified variance since lineages that fail to recombine off the beneficial background during the sweep, and fail to coalesce or recombine during the standing phase return to the source population. Thus,

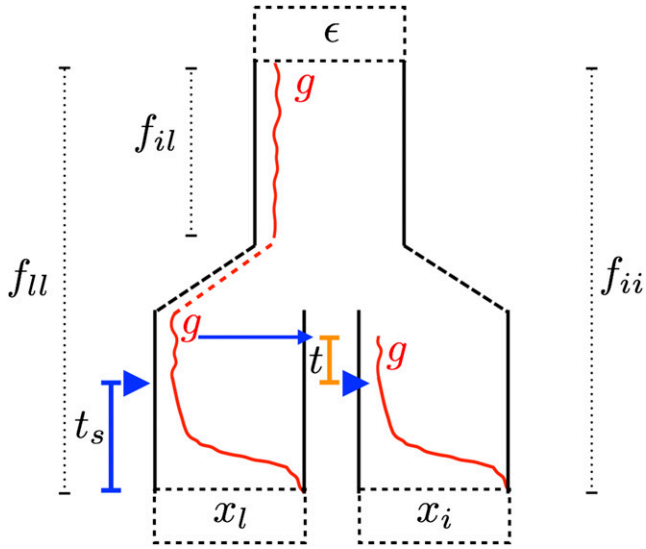

Figure 11 Trajectories of the beneficial allele (red) for the standing variant model with a source population. Populations *l* and *i* are under selection with present-day allele frequencies x_l and x_i at a neutral locus, derived from an ancestral population with allele frequency ε. The populations share some amount of drift proportional to f_{il} before reaching the ancestral population. The beneficial allele is standing at frequency *g* in the source population, *l*. It migrates into population *i* from *l*, where it is standing at frequency *g* for *t* generations prior to the onset of selection, indicated by the blue triangles.

$$f_{ii}^{(S)} = (1-y)^2 f_{ii} + 2y(1-y)\Big((1-r_t)f_{il} + (1-(1-r_t))f_{ii}\Big) + y^2\left(e^{-t\left(2r+\frac{1}{2N_e g}\right)}\left(\frac{1}{1+4N_e rg} + \frac{4N_e rg}{1+4N_e rg}f_{ll}\right)\right.$$
$$+\left(1-e^{-t\left(2r+\frac{1}{2N_e g}\right)}\right)\frac{1}{1+4N_e rg} + \left(\left(1-e^{-t\left(2r+\frac{1}{2N_e g}\right)}\right)\frac{4N_e rg}{1+4N_e rg} - \left(1-e^{-t\left(r+\frac{1}{2N_e g}\right)}\right)\frac{4N_e rg}{1+2N_e rg}(1-r_t)\right)f_{ii}$$
$$\left.+\left(1-e^{-t\left(r+\frac{1}{2N_e g}\right)}\right)\frac{4N_e rg}{1+2N_e rg}(1-r_t)f_{il}\right) \tag{A.23}$$

There is additional coancestry between pairs of selected populations. This takes a different form than Equation 9, as there, since if either lineage fails to recombines off the beneficial background during the sweep or standing phase, the lineage will be in population *l*. For selected populations *i* and *j*, now

$$f_{ij}^{(S)} = (1-y)^2 f_{ij} + y^2\left(r_t^2\left(\frac{1}{1+4N_e rg} + \frac{4N_e rg}{1+4N_e rg}f_{ll}\right) + (1-(1-r_t))^2 f_{ij} + (1-r_t)(1-(1-r_t))\Big(f_{il}+f_{jl}\Big)\right)$$
$$+y(1-y)\Big(2(1-(1-r_t))f_{ij} + (1-r_t)\Big(f_{il}+f_{jl}\Big)\Big) \tag{A.24}$$

If either population is the source, *l*, this reduces to

$$f_{il}^{(S)} = y(1-r_t)\left(y(1-r_t)\left(\frac{1}{1+4N_e rg} + \frac{4N_e rg}{1+4N_e rg}f_{ll}\right) + (1-y(1-r_t))f_{ll}\right) + (1-y(1-r_t))f_{il} \tag{A.25}$$

since, if the lineage fails to recombines off the beneficial background in population *i*, it is back in population *l*. If the lineage in *l* is still on the beneficial background after the sweep, and the initial *t* generations of standing, they can coalesce during the standing phase in population *l*. Else, the lineages will coalesce neutrally in population *l*. However, if the lineage sampled in

population i does not return to the source population (*i.e.*, it recombines during the sweep or standing phase of t generations), the lineages can coalesce with neutral probability f_{il}.

Lastly, we must incorporate the impact that linked selection has on the coancestry between lineages sampled from any pair of nonsource selected population i and nonselected population k.

$$f_{ik}^{(S)} = y\Big((1-r_t)f_{kl} + (1-(1-r_t)\Big)f_{ik}) + (1-y)f_{ik} \tag{A.26}$$

Since lineages that do not recombine off the beneficial background in population i go back into the source population l, nonselected populations may now have more or less coancestry with population i depending on whether l is neutrally has more or less coancestry with population l, respectively.

It may be possible to extend these models to allow the source population to be an unsampled population, u. In this case, we need information about how our unsampled source is related to our sampled populations. Specifically, we have f_{iu} and f_{uu} terms in the coancestry coefficients of any selected population i, as well as f_{iu}, f_{ju}, and f_{uu} for coancestry between any selected population pairs i and j and f_{kl} for unselected populations k. More work is needed to address this problem. It is possible to use all sampled populations, including nonselected populations, as proxies for the unsampled source to give us information about which sampled population our unsampled source is more closely related to. Additionally, if we assume the unsampled population is distantly related to our sampled populations, such that they span the root, the coancestry between u and any other sampled population will be 0.

A.5 Migration Model: More than Two Nonsource Selected Populations

In the main text, we consider two selected populations i and j, where population i is the source of the beneficial allele. We need to extend this model when we have more than two nonsource selected populations. Specifically, we need to define coancestry coefficients between selected nonsource pairs. Now, let population l be a selected population and the source of the beneficial allele.

The coancestry between nonsource selected populations is affected by migration, as there is some probability or either or both lineage failing to recombine off the beneficial background of the sweep and to migrate back into population l. Thus, for selected populations i and j,

$$f_{ij}^{(S)} = y^2e^{-2r\delta} + y^2\big(1-e^{-2r\delta}\big)f_{ll} + y(1-y)\Big(f_{il}+f_{jl}\Big) + y\big(1-ye^{-r\delta}\big)f_{ii} + (1-y)^2f_{ij} \tag{A.27}$$

If l is either population i or j, this reduces to Equation 13, up to a factor of 2δ, as now only one population experiences the delay, δ, as the other is the source. Thus, Equation 13 is more accurate for defining the coancestry coefficient between the source and selected populations. Equation 12 holds for the coancestry within all nonsource selected population, and Equation 14 for all nonselected and nonsource selected population pairs. Lastly, again, we assume the source coancestry within the source population l follows that of an independent sweep from new mutation (Equation 4).

Similar to the standing variant model with a source population above, we can think about extending this migration model to allow the source population to be unsampled. More work is needed to address the same issues related to estimating coancestry coefficients for unsampled populations.