A few stickleback suffice to transport adaptive alleles to new lakes

Jared Galloway, William A. Cresko, and Peter Ralph

December 12, 2018

Abstract

Threespine stickleback fish provide a striking example of local adaptation despite recurrent gene flow. The species is distributed around the Northern hemisphere in both marine and freshwater habitats. It is thought that these numerous, smaller freshwater populations have been established "de novo" from marine fish, and that a shared freshwater phenotype is often established using standing genetic variation. Here we use genealogical simulations to determine the levels of gene flow that best matches observed patterns of allele sharing among habitats in stickleback, and more generally to better understand how gene flow and local adaptation in large metapopulations determine speed of adaptation and reuse of standing genetic variation. We find that rapid, repeated adaptation using a shared set of alleles maintained at low frequency by migration-selection balance occurs over a realistic range of intermediate rates of gene flow. Low gene flow leads to slow, independent adaptation of distinct habitats, whereas high gene flow leads to large migration load. We quantify how F_{ST} scans for adaptive alleles are more likely to succeed with higher rates of gene flow. In addition, we find the origin of many freshwater adapted alleles in the introduced lakes to be propagated from the original generation of marine individuals that had inhabited the lake. The results support existing theory of local adaptation, and provide a more concrete look at a particular, empirically motivated example.

Introduction

The canonical model for the genetics of adaptation, first formulated by Fisher in the early part of the 20th century, involves the sequential fixation of new mutations. While it has proven valid in numerous studies in the field as well as in the lab, this model is now rightfully understood as incomplete for many species in nature that have more complicated population structures. What role does genomic variation play in metapopulations with connection to an array of selective pressures? How could migration-selection balance and ecological speciation actually benefit species as a whole? A growing number of studies have identified instances of convergent or shared adaptive evolution in small demes that utilize standing genetic variation (SGV) found in the static population from which they split. However, it is still widely unknown which variations of evolutionary processes such as gene flow, recombination, selection and mutation, promote the maintenance and re-use of such genetic polymorphisms. In addition, we would like to know how complex geographic systems, with a variety of selective pressures, interact with these evolutionary forces to help or hinder the use of SGV in newly colonized subpopulations.

these q's make sense to me, but could be utter nonsense to an experienced pop geneticist.

Recently, a large flood of data resulting from evolutionary systems of organisms have come from population genomic studies powered by advances in sequencing technologies. One such organism is the threespine stickleback. Many studies of species have exhibited a long standing evolutionary history of migration-selection balance between marine and freshwater populations across the globe. The rigidity of the stickleback's ability to frequently prosper in newly created freshwater environments has made this fish a good model for understanding the genetic basis of adaptive evolution. The ancestral marine form of stickleback has given rise to millions of independently derived populations in recently de-glaciated regions of the Northern Hemisphere. While geographic isolation prevents direct migration between a large majority of these patches of freshwater populations, They appear to exhibit very similar phenotypes, as well as many shared, adaptive alleles often found to be identical by decent (IBD). Impressively, independent local adaptation of marine individuals to freshwater environments has been observed to take place in tens of generations, and the adaptive alleles are found in freshwater populations that have been geographically isolated since the end of the last ice age $\approx 13,000$ years ago (Cite Kristin). For example, in 1964 the Great Alaskan Earthquake caused an uplift of Middleton island and in turn, introduced a group of freshwater ponds around the perimeter of the island. Quickly inhabited by the surrounding marine population of stickleback, Lescak et al. [2015] observed significant phenotypic changes in less than 50 years that appear to be parallel to freshwater stickleback that have been separated for over thousands of years. In these freshwater stickleback, the number of lateral plates are reduced and the opercle shapes shows the same expansion of the dorsal region and reduction of the ventral region as observed in the large majority of freshwater demes.

But how can evolution occur at such a rapid pace? Waiting for new mutations to arise in each lake would take much longer, and the genotypes being identical by state is even more improbable. This would seem to suggest that freshwater alleles are maintained in the marine individuals allowing the accelerated selection on SGV found in marine individuals which colonize the lake. The first clear example of the global reuse of SGV was the gene *eda* which has been shown to be an important regulator for the number of lateral plates. While the low lateral plate version of this gene arose millions of years ago, it is found in freshwater ponds which have formed much more recently. Contemporary population genomic studies employing genome-wide haplotype analyses has provided evidence that *most* regions of the genome that distinguish marine-freshwater genetic differences share this pattern [Nelson and Cresko, 2017].

These empirical data generally support the "transporter"-hypothesis proposed by Conte and Schluter in 2009 Schluter and Conte [2009], which is a conceptual model for the flow of freshwater alleles from multiple smaller freshwater populations into much larger and less structured marine populations through hybridization events. Alleles in this asymmetrically structured meta-population can then be recycled for subsequent freshwater adaptation. Several questions remain, however, about the manner and degree to which these alleles are scattered among the marine fish allowing the haplotype to be re-assembled. Is it more akin to the atomization and rebuilding that occurred in the Star Trek series that motivated this hypothesis? Alternatively does it occur in more of a patchwork by individuals, or their early generation hybrids in geographically adjacent freshwater habitats moving quickly through the marine environment to seed new freshwater populations. More generally, how do different levels of gene flow and local adaptation interact with the asymmetric nature of stickleback population structure affect the dynamics of the origin and rate of adaptive alleles, as well as their eventual organization into marine and freshwater typical genomes?

Here we develop a forward simulation approach in SLiM to model the stickleback evolutionary history in

marine and freshwater habitats. We then record the effect of variation in gene flow on the genetic and genomic architecture of local adaptation. We ask, how rapidly can selection act on standing genetic variants at a given value of migration, M? Furthermore, how might the inference of causal loci through the use of genome wide association studies (GWAS) and F_{st} scans across the genome be skewed by (lack of) introgression between sub-populations. Finally, we trace back through the genealogical history of adaptive variants in new lakes to find the precise origin is of those variants.

We find that only a few stickleback per generation, per lake, suffice to maintain freshwater alleles from a (global?) haplotype at an acceptable frequency in marine environments for efficient "transportation" of the freshwater haplotype. The selection on SGV at this level results in rapid local adaptation in ≈ 60 generations, this follows closely to what we observe in nature (such as middleton island, cite this). Surprisingly, we find that the the continued gene flow (subsequent migration) of marine individuals is less important than the initial amount of SGV in the initial generation marine individuals at time of colonization. This suggests downstream gene flow from freshwater populations is key to this system of "transportation". We show that low levels of introgression also results in noise that often presents itself as F_{st} peaks when scanning across the genome suggesting introgression plays a key role in inferring causal loci . With higher $M > 10$ individuals per lake per generation we see the impact of migration load, preventing the demes from fully adapting to their selective pressure.

Methods

We explored these questions using forwards-time simulations with explicit genomic representation of individuals selecting upon a single continuous quantitative trait using SLiM [Haller and Messer, 2017, 2018]. The details of the model were motivated by current understanding of threespine stickleback evolutionary history and demography. We set up a realistic model of the relationship between marine and freshwater selective pressures with continuous space and migration rates between the populations to emulate gene flow. Certain aspects of the the model remain simplistic due to computational constraints. Possibly the most important caveat is that simulated population sizes are much smaller than the census size of the threespine stickleback populations observed in nature (see the Discussion for more on this).

Habitat and geography Throughout the simulation there are two habitat types – Marine and Freshwater– defined by their selective pressures. We start the simulations with two populations of size 5,000 diploid individuals in each of the habitats, for a total of 10,000 diploid individuals. The arrangement of these habitats, depicted in Figure 1, roughly models a set of freshwater habitats along a stretch of coastline. The marine habitat is a continuous, one-dimensional range of length 25 units, while the freshwater habitat is divided into 25 discrete subpopulations (which we call "lakes"), each connected to the marine habitat at regularly spaced intervals (positions $i - 1/2$ for $1 \leq i \leq 25$).

Divergent selection is mediated by a single quantitative trait with different optima in marine and freshwater habitats. This situation roughly models the cumulative effect of the various phenotypes such as armor morphology, body size, craniofacial variation and opercle shape on which divergent selection is thought to act on in the two environments. Concretely, the optimal trait values in the marine and freshwater habitats are $+10$ and -10 respectively, and fitness of a fish with trait value x_{ind} in a habitat with optimal value x_{opt}

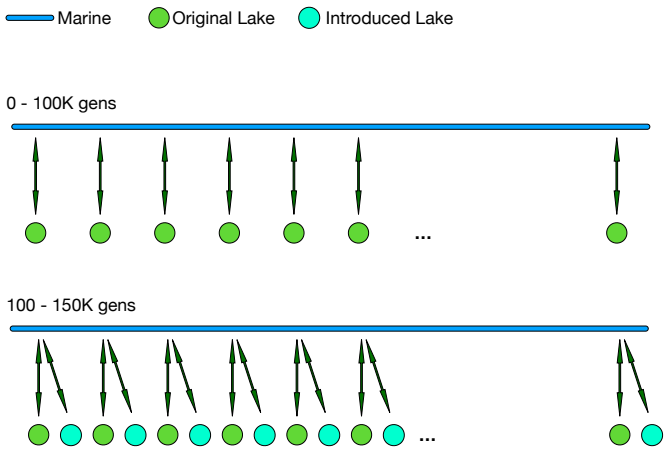

Figure 1: **Diagram of simulated populations:** a single, continuous, one-dimensional marine habitat (blue) is coupled to randomly mating "lakes" at discrete locations with dark green arrows representing migration patterns. After an initial period of 100K generations with 25 lakes, an additional 25 lakes are added (at the same set of locations) and populated with a copy of the *marine* individuals to simulate the appearance of newly accessible freshwater habitats colonized by marine stickleback. The marine habitat, and each set of 25 lakes, each contain 5,000 individuals at all times.

is determined by a Gaussian kernel with standard deviation 15, i.e.,

$$f(x_{\text{ind}}; x_{\text{opt}}) = \exp\left\{\frac{1}{2}\left(\frac{x_{\text{ind}} - x_{\text{opt}}}{15}\right)^2\right\}.$$

In short, the difference between an individual's trait value and the optimum largely determines that individual's fitness. Note that the scale we measure trait on is arbitrary as only the scale of trait values relative to mutation effect size is relevant. We chose the difference between optima and strength of stabilizing selection in each habitat so that (a) around 10 (diploid, homozygous) mutations were sufficient to move from one optimum to the other, and (b) well-adapted fish from one habitat would have low, but nonzero, fitness in the other habitat.

Genetic architecture of the trait Each individual carries two linear chromosomes, each of size 10^8 loci. Mutations that can affect the trait under selection can occur at rate 10^{-7} per locus per generation in ten "effect regions" of 10^5 loci each, spread evenly along the chromosome. Each mutation in these regions is either additive, completely recessive, or completely dominant (with equal probability). Effect sizes for these mutations are chosen randomly from an exponential distribution with mean 1/2, either positive or negative with equal probability. Individual trait values (x_{ind}) are determined additively from the diploid genotypes. Concretely, an individual that is heterozygous and homozygous for mutations at sets of loci H and D respectively has trait value $x_{\text{ind}} = \sum_{i \in H} h_i s_i + \sum_{j \in D} s_j$, where h_i and s_i are the dominance coefficient

and the effect size of the mutation at locus i. Subsequent mutations at the same locus replace the previous allele.

Population dynamics We use SLiM to simulate a Wright–Fisher population with non-overlapping generations and a fixed population size of 5,000 diploid individuals in each habitat. Each generation, the two parents of each new offspring are chosen proportional to their fitness (all individuals are hermaphroditic), and the contributing genomes are produced by Poisson recombination with an average of one crossover per chromosome per generation (10^{-7} per locus per generation). Since the total population across *all* 25 lakes is fixed at 5000, we normalize the fitnesses of each individual such that approximately 200 offspring are generated in each lake, each generation. to do this, we divide fitness values of each freshwater individual by the mean fitness in their lake, so that the mean fitnesses of all lakes are equal before selection happens.

The paragraph below has a confusing structure, skipping for now ... Dispersal occurs both locally along the coastline in the marine habitat and between the marine habitat and the lakes, and can be thought of as occurring at the juvenile stage. There is no dispersal directly between lakes. The lake–ocean migration rate is denoted m, which we refer to as the rate of *gene flow* between habitats. Each new individual in the marine habitat has freshwater parents with probability m; to obtain the pair, a first parent is chosen proportional to fitness, and a mate is chosen from the same lake as the first, also proportional to fitness. The resulting offspring is given a spatial location in the marine habitat at the location of the parent's lake. Parents for a new marine individual who is not a migrant are chosen similarly (with probability $1 - m$): first, a single parent is chosen proportionally to fitness in the marine habitat, and then a mate is chosen, also proportionally to fitness but re-weighted by a Gaussian function of the distance separating the two, with standard deviation 1/2. Concretely, if the first parent is marine individual i, then marine individual j is chosen as the mate with probability proportional to $f(x_j)\exp(-2d_{ij}^2)$, where d_{ij} is the distance between the two locations. Finally, each new marine offspring is given a position displaced from the first parent's position by a random Gaussian distance with mean 0 and standard deviation 0.02, and reflected to stay within the population. New offspring in the freshwater habitat are chosen in the same way, except the probability that the parents are marine individuals is m; any new freshwater offspring produced by marine individuals are assigned to the lake nearest to the position of the first marine parent.

New freshwater populations To study how newly appearing freshwater habitats adapt and select on the standing genetic variation in the marine, we introduce a new set of 25 lakes midway through the simulation at 100K generations. These new lakes are populated with a copy of the marine individuals to emulate a freshwater lake being colonized by its neighboring marine population. In particular, after this point of introduction in the simulation there are: two *sets* of lakes, and one marine population, each with 5,000 individuals for a total of 15,000 individuals being tracked in the simulation. Since this introduction of lakes doubles the number of lake-to-marine immigrants, the probability that a new marine individual has freshwater parents is $2m$ instead of m.

Descriptive statistics To assess whether new lakes adapt using existing genetic diversity, we define a *freshwater allele* to be an effect mutation that has frequency higher than 0.5 in at least one of the original lakes, while remaining lower than 0.5 in the marine. This categorization is made for each generation using the allele frequencies from that generation, and so changes with time. Alleles common in the newly introduced

Population genetic analyses? Instead of the more general 'Descriptive statistics'?

lakes do not count if they are not also common in the original lakes. They are defined this way because the transportation hypothesis does not specify where or when an advantageous mutation arises, but simply suggests that any sufficiently common freshwater adapted allele could participate in adaptation in new habitat [Schluter and Conte, 2009].

Time to adaptation of the introduced population, denoted T_{adapt}, is defined to be the generation at which the difference between the average trait value in the original and the introduced freshwater populations is less than 0.5.

We describe overall genetic differentiation between the habitats using F_{ST}, calculated on a per-locus basis. Concretely, if p_f and p_m are the frequencies of a given mutant allele in the freshwater and marine habitats, respectively, and $\bar{p} = (p_f + p_m)/2$, then we compute F_{ST} for that mutation as $1 - p_f p_m/(\bar{p}(1-\bar{p}))$.

Recording genealogical history We used SLiM's ability to record *tree sequences* [Haller et al., 2018] to output the genealogical history of all individuals at the time of introduction of new lakes, at the time of adaptation, and at the end of the simulation. This allowed us to directly query the true origins of adaptive alleles. In addition it allowed for much larger simulations by avoiding the computationally expensive task of simulating neutral mutations which were retroactively added to the gene trees at a rate of 10^{-7} per locus per generation, as described in Kelleher et al. [2018]).

The output tree sequence from each simulation allows us to explore the origin of the genetic basis of adaptation in the new lakes. To do this, we constructed the genealogical tree relating all extant chromosomes at each locus along the genome. Using these trees we classified each adaptive allele, in each genome in the new lakes at the time of adaptation, into four categories:

1. *a "De novo" allele:* deriving from a new mutation that occurred in a new lake.
2. *a "Migrant" allele:* deriving from a migrant that was not in the initial generation that colonized the lake
3. *a "Captured" allele:* present in the individual that initially colonized the new lake, and both common (above 50%) in the original lakes, and uncommon (below 50%) in the ocean.
4. *a "Marine" allele:* present in the individual that initially colonized the new lake, and not a captured allele.

We then looked at the ratio of alleles from each origin to the total number of alleles being used as the basis for local adaptation in the new lakes. This essentially gives a measure on whether selection in the new environments made use of (1) new mutation, (2) post-colonization migration, (3) standing variation at migration–selection balance, and (4) standing variation at mutation–selection balance. In other words, we were able to quantify the contributions of these possible origins of adaptive alleles to understand where the genomic basis of adaptation had derived from.

We were also able to use the tree sequence to get information about the individual "effect regions" in all initial genomes of the new lakes at time of introduction. From this we determined the ability for the new lakes to select upon the standing variation in the marine. Given the probability that effect region fixes at a given trait value, $2 * x/45$, where x is the effect size of a haplotype, we calculated the expected total effect

sizes of the haplotypes that fix. Concretely, it is the sum across the 10 effect regions of

$$\sum_{i=1}^{N} \prod_{j<i} (1 - p_j) p_i x_i$$

where N is the number of genomes per lake, p_i is the probability of fixation of the i^{th} haplotype, x_i is the effect size of the i^{th} haplotype, and the haplotypes are sorted in decreasing order by x_i. These numbers were computed assuming additivity of mutations, meaning the numbers produced are a slight overestimate.

Results

To observe the impact of gene flow on selection in the new lakes, we varied the ocean–lake migration rate, m, across separate simulations from 5×10^{-5} to 5×10^{-1}. Since each lake contains 200 individuals, we refer to these migration rates as between 0.01 and 100 migrants per lake, per generation. Many aspects of adaptation changed substantially across this range, including the speed of adaptation, degree of sharing of adaptive alleles between lakes, and the population genetic signals left behind. At very low rates of gene flow, each new lake's population adapted completely separately from *de novo* mutation, which took a very long time ($\approx 20,000$ generations). At very high rates of gene flow, local adaptation was constrained by the large influx of maladaptive alleles. Between these two extremes, genetic variation that allowed adaptation to freshwater habitats could move relatively easily between lakes. Surprisingly, only a few migrants per generation from the lakes to the marine were needed to see the effects of rapid local adaptation accelerated by selection on standing genetic variants in the new lakes – despite being deleterious in the intervening marine habitat.

Local Adaptation: differentiation with gene flow

As shown in Figure 2, local adaptation occurred at migration rates of $m \leq 10$. Figures 3 and S3 show freshwater and marine populations diverging until the trait means were close to the optimal values in each habitat. The establishment of new alleles in the lakes is visible in the first few thousand generations of Figure 3 as jumps in the mean trait value. At the lowest rate of gene flow, $m = 0.01$, differences at ≈ 20 commonly polymorphic sites (≈ 10 that shift the trait in each direction) were responsible for most of the adaptive differences between freshwater and marine habitats. In Figure 6A, we can see that more gene flow simplified the genetic basis. As expected, increasing migration rate decreased differentiation between habitats. As seen in Figure 4, F_{ST} between marine and freshwater habitats at neutral sites steadily declines as migration increases.

Speed of adaptation Adaptation occurred much more quickly at higher migration rates, both in the old and new sets of lakes. We measured this "time to adaptation" as the number of generations until average trait values in old and new lakes were within 0.5 of each other, shown in Figure 5 for different rates of gene flow. Adaptation of new lakes took over 18,000 generations at the lowest rate of $m = 0.01$, while at $m = 1$ migrant per, lake per generation, new lakes managed to adapt in just under 60 generations.

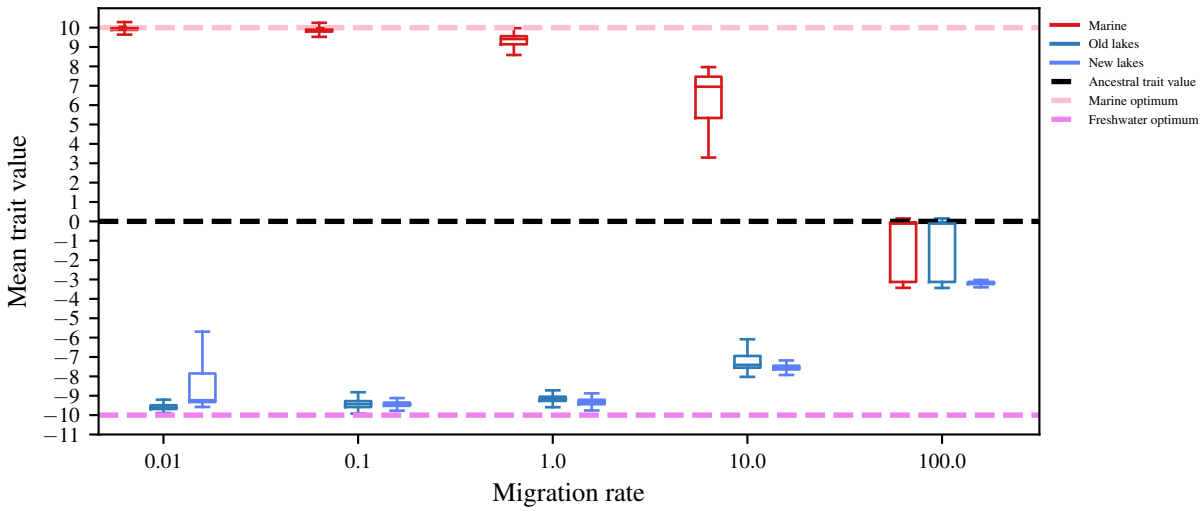

Figure 2: Distribution of mean trait values across generations of the simulation, for different migration rates. The dashed pink and purple lines at ±10 give the optimum phenotypes in the marine and freshwater environments, respectively.

Sharing of freshwater adapted alleles

At low migration rates, the *initial* period of adaptation takes roughly 25 times longer for lakes than it does for the ocean. This difference occurs because at low migration rates, all subpopulation must wait for novel mutations to arise in order to adapt. Because the marine is one continuous population with 25 times more individuals than any one lake, there is a much larger influx of new mutations to be selected upon. At higher migration rates, greater mixing allows the initial lakes to share alleles instead of developing their own genetic basis of adaptation. As a first indication of this, Figure 4 shows that F_{ST} between the "original" and "introduced" sets of lakes at effect mutations decreased with migration rate.

To investigate in more depth how locally adaptive alleles found in the original lakes are shared between lakes, as well as how they spread to the new lakes, we defined and tracked the distribution of "pre-existing freshwater adapted alleles" at the beginning of each generation. To be considered in this category, a mutation must participate in the genetic basis of any freshwater adaptation. Concretely, these are any non-neutral mutations whose frequency is above 50% in at least one original lakes and below 50% in the marine habitat. Figure 6A shows the distribution of the number of these alleles across generations. At $m = 0.01$, each lake has a private set of about 10 mutations nearly fixed in that lake but not elsewhere: the new lakes acquire new adaptive alleles and therefor have little to none of these adaptive alleles. At $m = 0.1$, we again observe the original lakes adapting nearly independently from each other. However, in Figure 6A we can see the average marine individual is shown to hold ≈ 2 pre-existing freshwater adapted alleles. This being the case, we can see new lakes adapt using a subset of this large repertoire of standing variation deriving from a multitude of independently derived haplotypes. As migration rate increases past this, we can see in Figure 6C that the total number of pre-existing freshwater adapted alleles declines. This suggests that the alleles move between

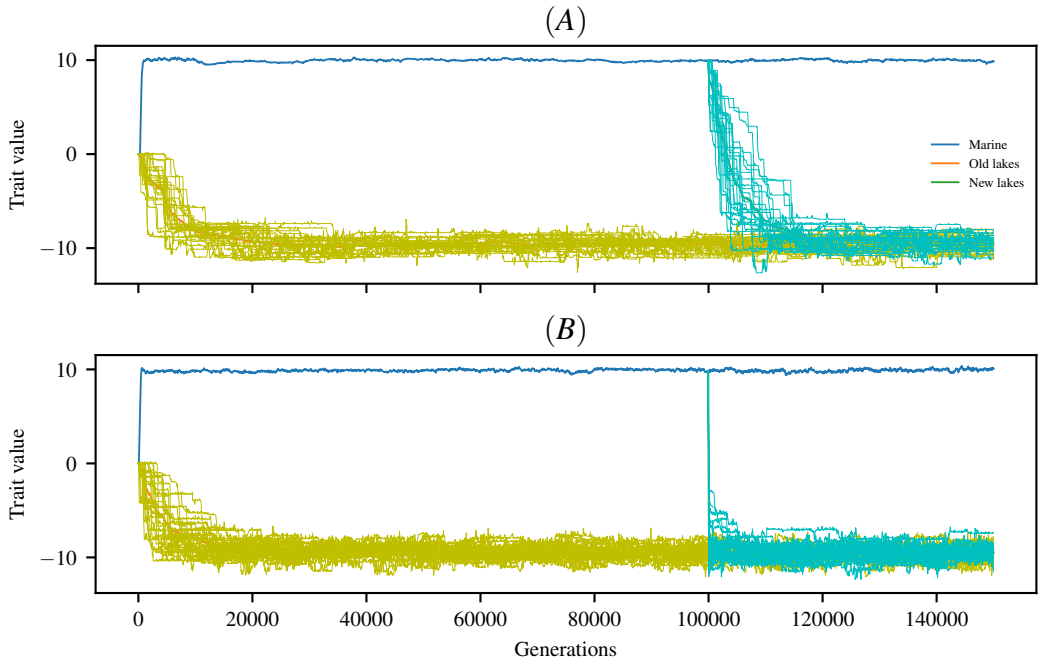

Figure 3: Mean individual trait values in the marine habitat (blue line), the original lakes (yellow lines; average in orange), and the new lakes (light green lines; average in dark green), across the course of two simulations, with migration rates of **(top)** $m = 5 \times 10^{-5}$ and **(bottom)** $m = 5 \times 10^{-4}$ per generation per individual (i.e., 0.01 and 0.1 migrants per lake per generation, respectively). Optimal trait values in the two habitats are at ± 10. Analogous plots for other migration rates are shown in Figure S3.

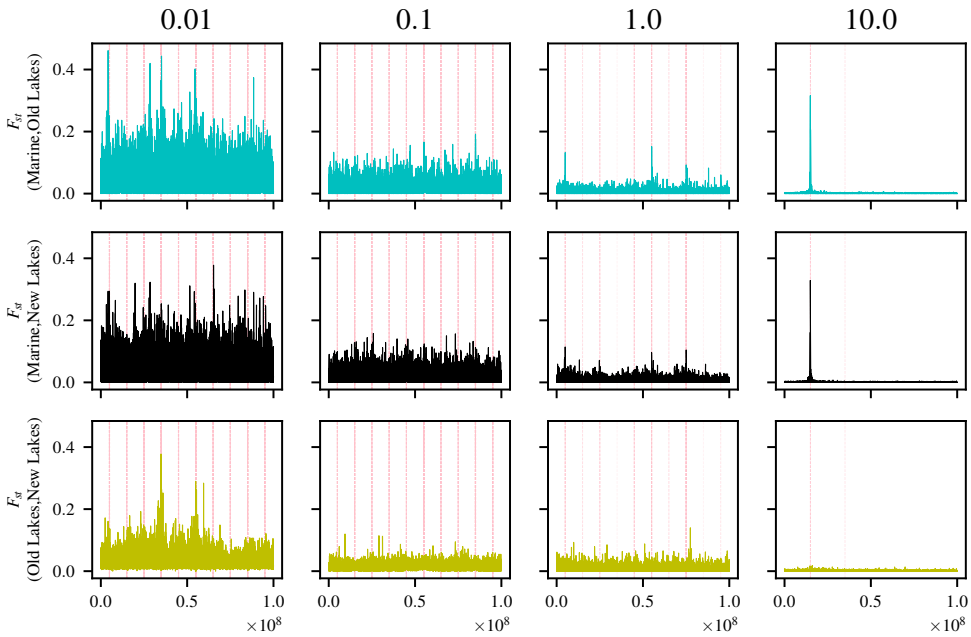

Figure 4: Average F_{ST} in windows of 500Mb between: **(top)** marine habitat and old lakes; **(middle)** marine habitat and new lakes; and **(bottom)** old and new lakes. Each plot shows F_{ST} values for a separate simulation, with columns corresponding to increasing gene flow from left to right. All locations of pre-existing freshwater adapted alleles have been highlighted by pink dashed vertical lines with an alpha value of 0.3, so darker shades of blue imply more freshwater adapted alleles at that location.

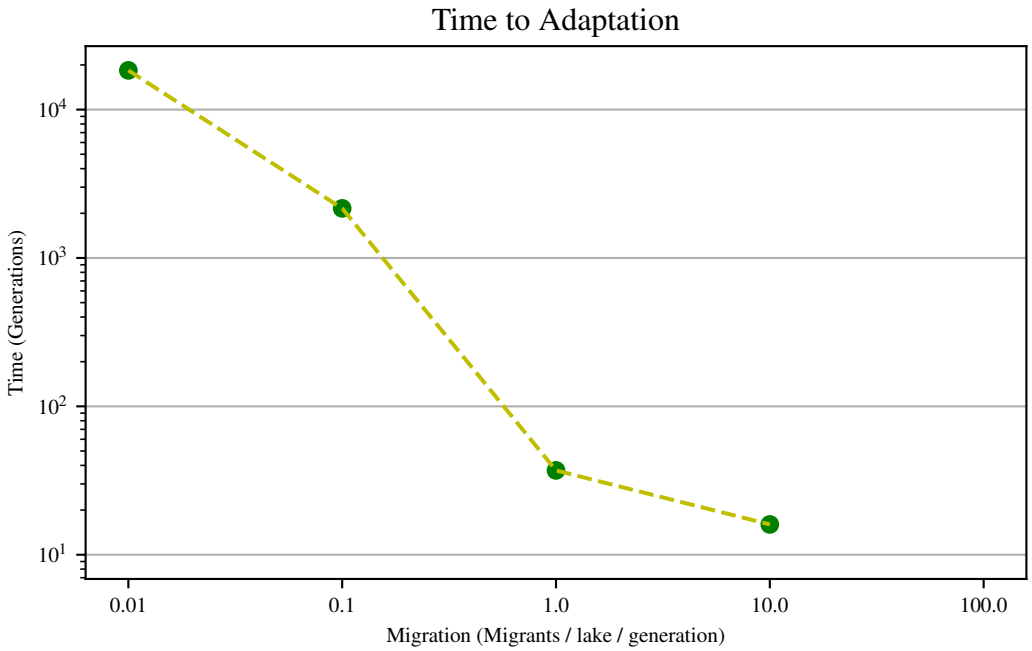

Figure 5: Time to adaptation as a function of migration rate. The time to adaptation is measured as the number of generations until the introduced population's mean phenotype comes within 0.5 of the original lakes average phenotype. Each point represents a single simulation run. (Adaptation did not occur at the highest rate of gene flow.)

populations by migration before selection acts on new mutation for neighboring lakes. Interestingly, the frequency of these alleles in the ocean stays relatively constant across the reasonable rates of gene flow.

Figure 6B shows the distribution, through time, of the mean percentage of currently-defined freshwater adapted alleles that each genome in each of the populations carries. If all individuals across lakes carried the same set of alleles determining their trait value, this would be 100%. At the lowest migration rate $m = 0.01$, each genome in the original lakes have almost exactly $1/25^{th}$ of the total number of pre-existing freshwater adapted alleles – this is because each one of the 25 lakes has adapted with a unique set of alleles. Since these are *pre-existing* alleles, the value is zero for introduced lakes. Figure 6A shows us that at 0.1 migrants per lake per generation and above, the average individual across the new lakes has nearly the same amount of pre-existing freshwater adapted alleles as individuals across the old lakes. As expected, the genetic basis of the freshwater phenotype seems to simplify as migration increases – higher rates of migration allow adaptive alleles of higher effect to travel more efficiently through the population, even though they are deleterious in the ocean.

The numbers of Figure 6 suggest that the dramatic increase in speed of local adaptation we observed above occurs because higher gene flow between populations allows sharing of freshwater alleles between populations. We confirmed this by using recorded tree sequences to identify the origin of each trait-affecting allele, in each individual, common in any of the new lakes at time of adaptation as defined in the Methods. Figure 7 shows that at the lowest rate of gene flow the majority of adaptive alleles are derived from *"de novo"* mutation. As gene flow increases, a larger fraction of adaptive alleles derive from pre-existing variation in the marine population at the time of introduction. In other words, greater mixing at higher migration rates allows lakes to share alleles instead of developing their own genetic basis of adaptation.

While increased migration allows sharing of adaptive alleles between lakes, at $m = 10.0$ the constant influx of alleles between the habitats creates substantial migration load. The rate at which migration load becomes substantial, 10 migrants per lake per generation, only replaces 5% of each population each generation with migrants from the other habitat, but this is sufficient to shift the mean trait values to nearly half their optimal values, as seen in Figure 2.

Realized genetic architecture

Here, we take a closer look at the genomic architecture of local adaptation between the two habitats. Do measures of local differentiation identify the causal loci? Figure 4 shows plots along the genome of per-locus F_{ST} values between habitats. (Note that we are pooling freshwater habitats; a single lake would provide substantially less power.) Higher migration rates showed more distinct F_{ST} peaks over polymorphic loci underlying trait differences between the habitats. "Background" levels of F_{ST} increase as gene flow decreases, swamping out this signal until the regions under selection are indistinguishable. This is likely due to two reasons: first, stronger genetic drift with less migration leads to higher background F_{ST}, and second, greater sharing of adaptive alleles providing a shared signal across populations.

need to know where the FAA are in these figures

check the new plot

This suggests that genome scans for local adaptation based purely on measures of differentiation will only be successful given enough migration between habitats. To quantify this, Figure S1 shows the power and false positive rates that would be obtained by an F_{ST} cutoff that declared everything above a certain value to be a causal locus.

Need to fix caption on figure S1 before editing this bit.

There is still something

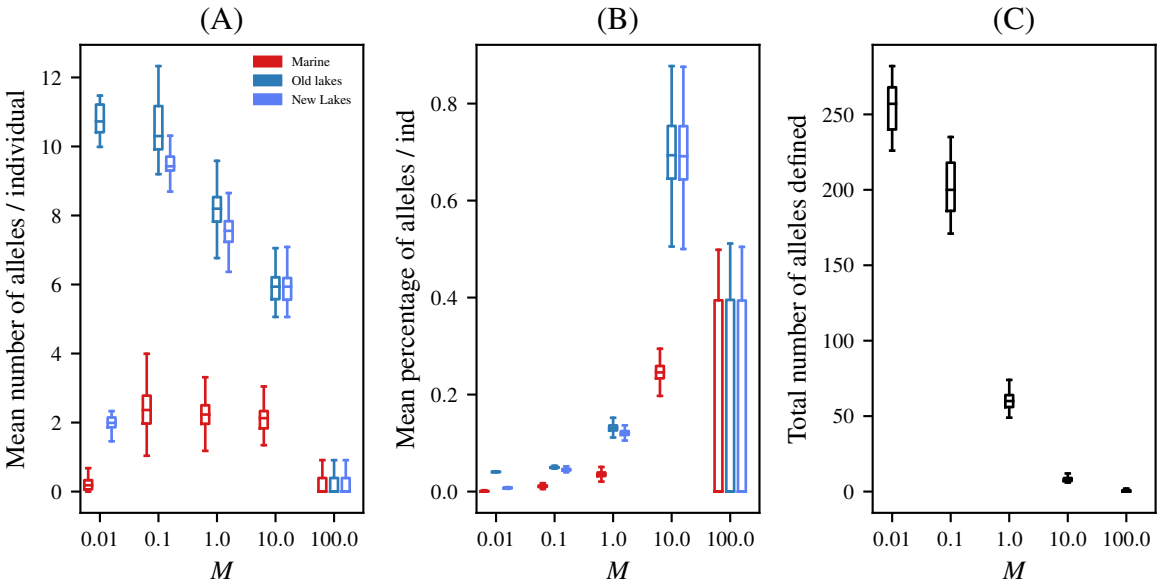

Figure 6: Amount of standing freshwater variation by habitat, across migration rates. Each plot counts "pre-existing freshwater adapted alleles", that are common in the original lakes but rare in the ocean (see text for definition). **(A)** Mean number of these alleles per individual. **(B)** Mean percentage of these alleles per individual. **(C)** Total number of these alleles (so, $B = A/C$). The number of alleles meeting these conditions changes over the course of the simulation, and each plot shows distributions of these values across generations. The horizontal axis shows M, the mean number of migrants per lake per generation.

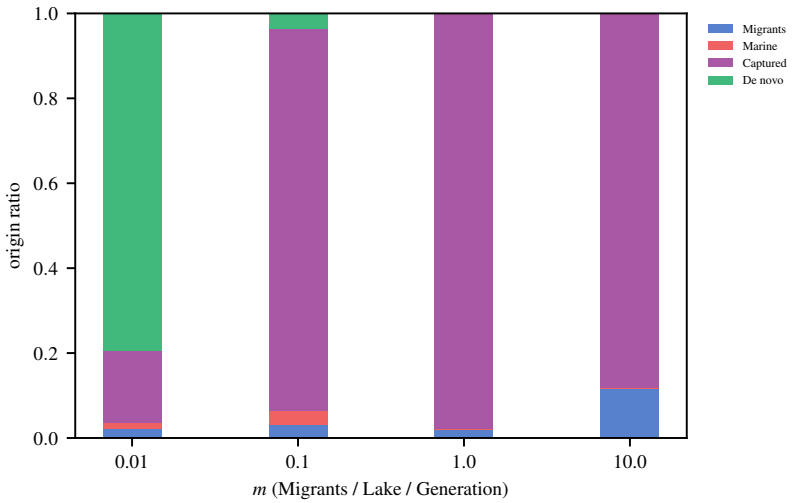

Figure 7: **(Origin of adaptive alleles:)** Each bar plot shows the origins of all trait-affecting alleles above frequency 50% in at least one new lake, classified as **(red)** new mutations, **(blue)** post-colonization migrants, **(green)** "captured" from pre-existing lakes, or **(orange)** standing marine variation. See Methods for precise definitions of these categories.

In regions of the genome underlying individual trait value, we observed Given that migration increases the gene flow between subpopulations, how valid are F_{st} peaks at different M. Knowing exactly which mutations effect phenotype in our simulations, we can look at the statistical power and false positives given F_{st} per SNP across the genome. In Figure S1 , looking at an F_{st} threshold greater than 1, we see the two lowest migration rates 10^{-5} and 10^{-4} having very little statical power. This along with low false positive rate across all F_{st} threshold values is fairly predictable when you consider the high F_{st} values across the genome.

Origin of introduced freshwater adaptive alleles

We have thus far found that the speed of adaptation depends strongly on the degree to which alleles can be shared between populations. However, there remains a mystery. Figure 6A shows that with $m = 0.1$ migrants per lake per generation, there is substantial sharing of alleles between populations, and yet Figure 5 shows that new lakes take around 2,000 generations to adapt. This is surprising because we see a similar quantity of allele sharing at $m = 1.0$, but much more rapid adaptation (≈ 30 generations). Perhaps this $60\times$ discrepancy occurs because new lakes must wait for an additional influx of freshwater alleles through migration beyond what is present in the initial generation? This hypothesis is disproven by Figure 7, which shows that at both $m = 0.1$ and $m = 1.0$, the vast majority of alleles underlying the local adaptation derived from individuals in the original generation ("captured").

	m=0.01	m=0.1	m=1	m=10
best	2.77	-12.02	-21.02	-7.02

Table 1: Haplotypic variation present in the new lakes at time of colonization, across rates of gene flow, quantified as the most negative trait value achievable with intact haplotypes, averaged across populations (see text for details).

An alternative explanation is that at lower rates of gene flow, freshwater alleles present as standing variation in the marine habitat are more tightly linked to marine alleles, and so adaptation in the new lakes must wait for recombination. This might be expected since freshwater alleles present in the ocean, to remain present at migration-selection balance at low migration rates, must be "masked" by nearby compensatory alleles. Reversing this "masking" then slows the process of adaptation that uses this genetic variation. Said another way, higher gene flow from freshwater to the ocean maintains relatively intact freshwater haplotypes that can be more easily rebuilt in the marine environments. Recall that trait-affecting mutations only occur in relatively small regions of 10^5 loci in which recombination occurs at rate 10^{-2}. Even if there is a sufficient amount of variation in the initial population of a lake to shift the trait from the marine optimum ($+10$) to the freshwater optimum (-10), these may be tightly linked to alleles that counteract their effect. For example, suppose there are 10 variants segregating at low frequency with effect size -1 each, but each is paired with a compensatory allele with effect size $+1$. Each local haplotype is therefore neutral (and so likely to be at higher frequency in the marine population).

Indeed, this appears to be the case. To quantify the genetic variation available *without* recombining within the ten genomic regions, we first found, within each population, the haplotype with the largest net negative effect at each of the ten genomic regions. Summing these ten numbers, we get the maximum amount that selection could move the population in the freshwater direction without recombining within regions.

The mean of this value across the 25 populations is shown in the first row ("best") of Table 1 – typical populations at $m = 0.1$ migrants per lake per generation could shift to a phenotype of -12.02 (and so have sufficient variation to adapt without recombination), but the "best" haplotypes populations at $m = 1.0$ have effect sizes nearly twice as big (a mean total of -21.02).

Theoretical expectations

We now revisit our observations above in the light of population genetics theory. The discussion will be rough, with a focus on intuition rather than precise calculations. Since within each population we model stabilizing selection on an additive trait controlled by a moderately large number of loci, more precise expectations might be obtained through quantitative genetics [Svardal et al., 2014] or even Fisher's geometric model [Barton, 2001, Chevin et al., 2014].

Suppose a new allele enters a lake, either by migration or mutation. If, when it is rare but present in n copies, it has fitness advantage s – i.e., the expected number of copies in the next generation is $(1 + s)n$ – then the probability that it escapes demographic stochasticity to become common in the population is approximately $2s$ [Lambert, 2006, Haldane, 1927] (assuming Poisson reproduction, as we roughly have here). If the current population all differed from the optimum trait by z, and the allele has effect size $-u$ in heterozygotes, then the fitness advantage of the allele would be $s(u) = \exp(-\beta((z-u)^2)/\exp(-\beta z^2)) \approx 2\beta zu$, where in our parameterization, $\beta = 1/450$. This tells us two things: (1) the rate of adaptation decreases as the population approaches the optimum, and (2) larger mutations (in the right direction) are more likely to fix.

New mutations The total rate of appearance of new mutations per lake is $\mu_L = 0.04$, which are divided evenly in seven categories: neutral, and then additive, dominant, and recessive in either direction. This implies that a new additive or dominant effect mutation appears once every 87.5 generations, on average. Effect sizes are randomly drawn from an Exponential distribution with mean 1/2, and so the probability that a dominant mutation manages to establish in a population differing from the optimum by z is roughly $\int_0^\infty 4\beta zu \exp(-2u)du = \beta z$, and so the rate of establishment of dominant mutations is $\beta z/87.5$, i.e., about one such mutation every $2461/z$ generations. The distribution of these successfully established mutations has density proportional to $u\exp(-2u)$, i.e., is Gamma with mean 1 and shape parameter 2. Since additive alleles have half the effect in heterozygotes, they have half the probability of establishment. During the initial phase of adaptation, the populations begin at around distance $z = 10$ from the optimum. Combining these facts, we expect adaptive alleles to appear through mutation within lakes at first on a time scale of 250 generations, with the time between local fixation of new alleles increasing as adaptation progresses, and each to move the trait by a distance of order 1. This agrees roughly with what we see in Figures **??** and XXX.

refer to additional trace plots in the supp

Standing variation An allele that moves the trait z units in the freshwater direction in heterozygotes has fitness roughly $\exp(-\beta z^2) \approx 1 - \beta z^2$ in the marine environment. The product of population size and fitness differential in the marine environment for a mutation with $z = 1$ is therefore $2Ns = 8.9$, implying that these alleles are strongly selected against but might occasionally drift to moderate frequency. The average frequency of such an allele in the marine environment at migration-selection equilibrium is equal to the total

influx of alleles into the ocean per generation divided by the selective disadvantage, which if there are M immigrants per generation, is $2M/\beta z^2$. Each new lake is likely to contain a few copies of alleles at frequency above 1/400 (since each new lake is begun with 400 genomes). With $z = 1$, the factor multiplying M is $2/\beta z^2 \approx 1/200$, so if $M \geq 1$, the chances are good that any particular lake-adapted allele that is present in all pre-existing lakes will appear at least once in the fish that colonize a new lake. However, an allele with $z = 1$ only has probability of around 1/20 of establishing locally, suggesting that the migration rate should be somewhat higher to ensure enough pre-existing genetic variation that adaptation would happen entirely using the initial set of colonizers. However, this calculation treats each allele independently; in practice we found that standing freshwater variation in the ocean were masked by linkage to compensatory marine variants.

This M is the total number in the ocean! Make sure this differs from notation above for the number per lake.

Migration The key quantity regulating the amount of standing variation in the ocean was the *downstream* migration rate, from lakes to the ocean. How important is the upstream migration rate? If sufficient genetic variation is not present in a new lake initially, it must appear either by new mutation or by migration. Since a proportion m of each lake is composed of migrants each generation, it takes $1/m$ generations until the genetic variation introduced by migrants equals the amount initially present at colonization. This implies a dichotomy: either (a) migration is high, and adaptation is possible using variants present at colonization or arriving shortly thereafter, or (b) migration is low, so adaptation takes many multiples of $1/m$ generations. Since in our model lower migration also reduces the amount of variation available in the ocean, we expect very little contribution of subsequent migration across any value of m, as seen in Figure 7.

Discussion

In this paper, we have described a realistic simulation model of a coastal metapopulation to observe the effects of selection of genetic variation (SGV) across a wide range of gene flow rates. We have shown that historical introgression, at our given parameter sets, is able to reproduce rapid and parallel adaptation similar to what we've seen in real populations such as Middleton island. Selection is able to rebuild the freshwater haplotype at a rapid pace (in tens of generations) from marine populations as a medium between all freshwater populations. Almost all rates of migration were helpful in the efficiency of the population to locally adapt except for the highest at which migration load limited the ability of the populations to reach the local optimum.

While our findings prove useful to the understanding of stickleback, they are not autonomous to this species. Now, we will discuss the broader implications of our findings on the more general framework of similar systems, we also explore what other questions remain. Small patches of similar selective pressures can be found in many places such as high altitudes, underwater volcanic vents, or post forest fire environments. Recently, (Cite Nelson Ferretti et al), discovered seven new species of Hapalotremus (Theraphosinae), living at the extremely high altitudes distributed along the Andes and Yungas in western South America. With the ancestral species at much lower altitudes, beneficial high altitude alleles could potentially be carried to help in the colonization of neighboring mountain environments. Systems of metapopulations like this could potentially arise in many places where the ancestral species carries around with it beneficial alleles forselective pressures it has encountered in the past and remains connected to through hybridization events.

Atmosphere's ability to carry whoever scotty is beaming up ... Perhaps the most notable finding in this study is that large, ancestral populations have the capacity to maintain alleles which are potentially deleterious for an extended amount of time. This principle, deriving solely from the properties of sex and natural selection, could be applicable to almost any species experiencing the influence of divergent selection with introgression. And to explore further how this process is carried out exactly, many of the questions lie in understanding which properties of genomic architecture and demographics allow this transportation to be possible. his is largely analogous to exploring the specific properties of the atmosphere / space which allow for the heroes to be disassembled, then put back together. In our simulations, the continued gene flow from freshwater populations meant that the alleles were maintained at a constant rate with fixed demographics. Two possible questions that might arise from this are: (1) How do population sizes, demographics, and evolutionary histories impact an ancestral population's capacity for carrying maladaptive alleles in its opposing environment, and (2) What role does genomic architecture play in the ability for a haplotype to be rebuilt.

Origins of genetic basis of local adaptation in new lakes In Figure 7 we found that the genetic basis for local adaptation, when selecting upon SGV, is propogated largely from the initial generation of inhabitants. This contrasts the belief that subsequent migration is a necessity in newly established populations ability to rebuild a haplotype. This finding suggests that with high probability, a randomly chosen subset from ancestral species in the metapopulations like marine stickleback, carry the capacity to rapidly adapt without *any* continued connection to the rest of of the species.

Realized genomic architecture Additionally have also shown introgression is beneficial for inferring causative loci from divergence (F_{st}) along the genome. This is generally because noise of selectively neutral alleles divergence can appear causative when genetic drift causes more differences between populations that have little gene flow between them. It's important to know that in all scenarios, hitchhiking of selectively neutral alleles could also be mistaken for being causative as they often display the same amount of divergence.

References

N. H. Barton. The role of hybridization in evolution. *Molecular Ecology*, 10(3):551–568, 2001. doi: 10.1046/j.1365-294x.2001.01216.x. URL `http://www.ncbi.nlm.nih.gov/pubmed/11298968`.

Luis-Miguel Chevin, Guillaume Decorzent, and Thomas Lenormand. Niche dimensionality and the genetics of ecological speciation. *Evolution*, 68(5):1244–1256, 2014. ISSN 1558-5646. doi: 10.1111/evo.12346. URL `http://dx.doi.org/10.1111/evo.12346`.

J. B. S. Haldane. A mathematical theory of natural and artificial selection, part V: Selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(07):838–844, 7 1927. ISSN 1469-8064. doi: 10.1017/S0305004100015644. URL `http://journals.cambridge.org/article_S0305004100015644`.

Benjamin C. Haller and Philipp W. Messer. Slim 2: Flexible, interactive forward genetic simulations. *Molecular Biology and Evolution*, 34(1):230–240, 2017. doi: 10.1093/molbev/msw211. URL `http://dx.doi.org/10.1093/molbev/msw211`.

Benjamin C Haller and Philipp W Messer. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, page msy228, 2018. doi: 10.1093/molbev/msy228. URL http://dx.doi.org/10.1093/molbev/msy228.

Benjamin C. Haller, Jared Galloway, Jerome Kelleher, Philipp W. Messer, and Peter L. Ralph. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *bioRxiv*, 2018. doi: 10.1101/407783. URL https://www.biorxiv.org/content/early/2018/09/04/407783.

Jerome Kelleher, Kevin R. Thornton, Jaime Ashander, and Peter L. Ralph. Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, 14(11):1–21, 11 2018. doi: 10.1371/journal.pcbi.1006581. URL https://doi.org/10.1371/journal.pcbi.1006581.

A Lambert. Probability of fixation under weak selection: a branching process unifying approach. *Theor Popul Biol*, 69(4):419–441, June 2006. doi: 10.1016/j.tpb.2006.01.002. URL http://www.ncbi.nlm.nih.gov/pubmed/16504230.

Emily A. Lescak, Susan L. Bassham, Julian Catchen, Ofer Gelmond, Mary L. Sherbick, Frank A. von Hippel, and William A. Cresko. Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences*, 112(52):E7204–E7212, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1512020112. URL http://www.pnas.org/content/112/52/E7204.

Thomas C. Nelson and William A. Cresko. Ancient genomic variation underlies recent and repeated ecological adaptation. *bioRxiv*, 2017. doi: 10.1101/167981. URL https://www.biorxiv.org/content/early/2017/07/25/167981.

Dolph Schluter and Gina L. Conte. Genetics and ecological speciation. *Proceedings of the National Academy of Sciences*, 106(Supplement 1):9955–9962, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0901264106. URL http://www.pnas.org/content/106/Supplement_1/9955.

Hannes Svardal, Claus Rueffler, and Joachim Hermisson. A general condition for adaptive genetic polymorphism in temporally and spatially heterogeneous environments, 2014. URL http://arxiv.org/abs/1411.3709. cite arxiv:1411.3709.

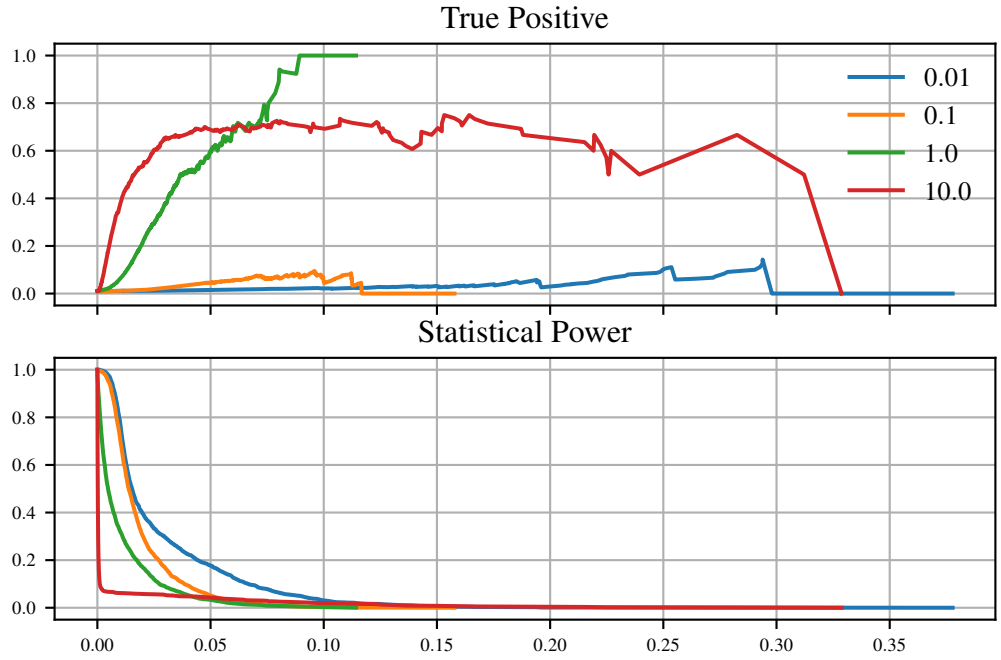

Figure S1: Statistical Power and False Positives as a function of F_{st} threshold. Statistical Power is the likelihood that a SNP will be predicted to have an effect on phenotype when there is an effect to be detected (?). False Positives give us the ratio of SNPs that effect phenotype to total SNPs greater than the F_{st} threshold. **TODO: fix this caption; x-axis label (F_{ST}).**

Supplementary material

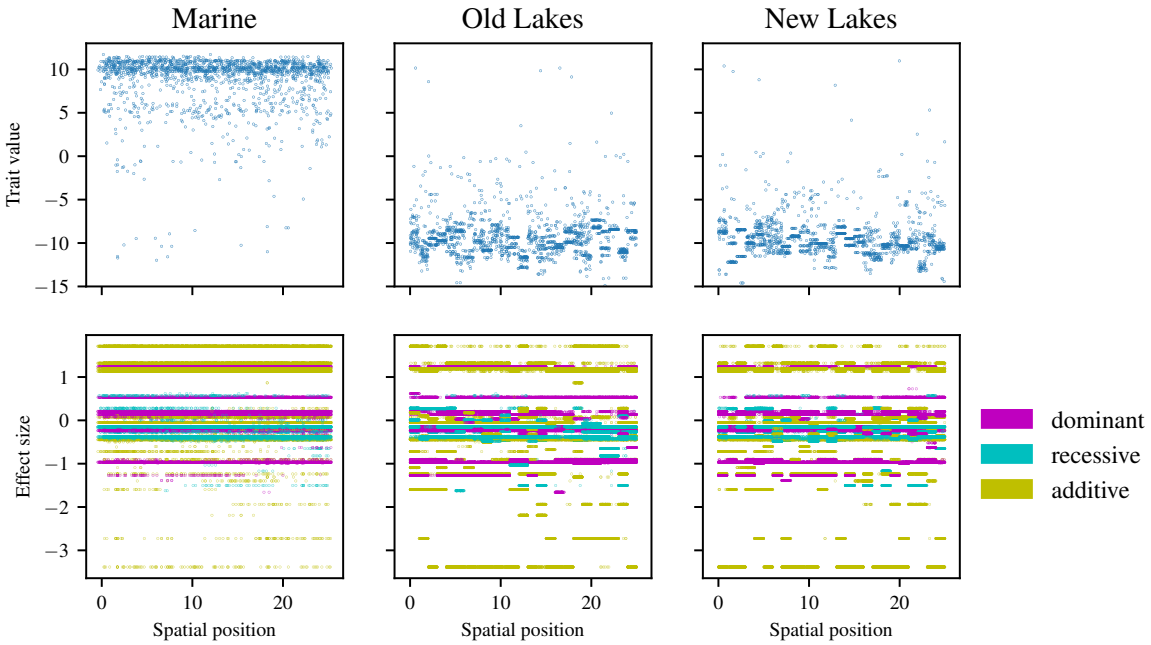

Figure S2: Phenotypes (**Top**) (effect size and mutation type) and haplotypes **(Botton)** as a function of spatial position of the individual possessing the genome. The phenotype of each haplotype is, by definition, the sum of the effect sizes of each mutation in the genome.

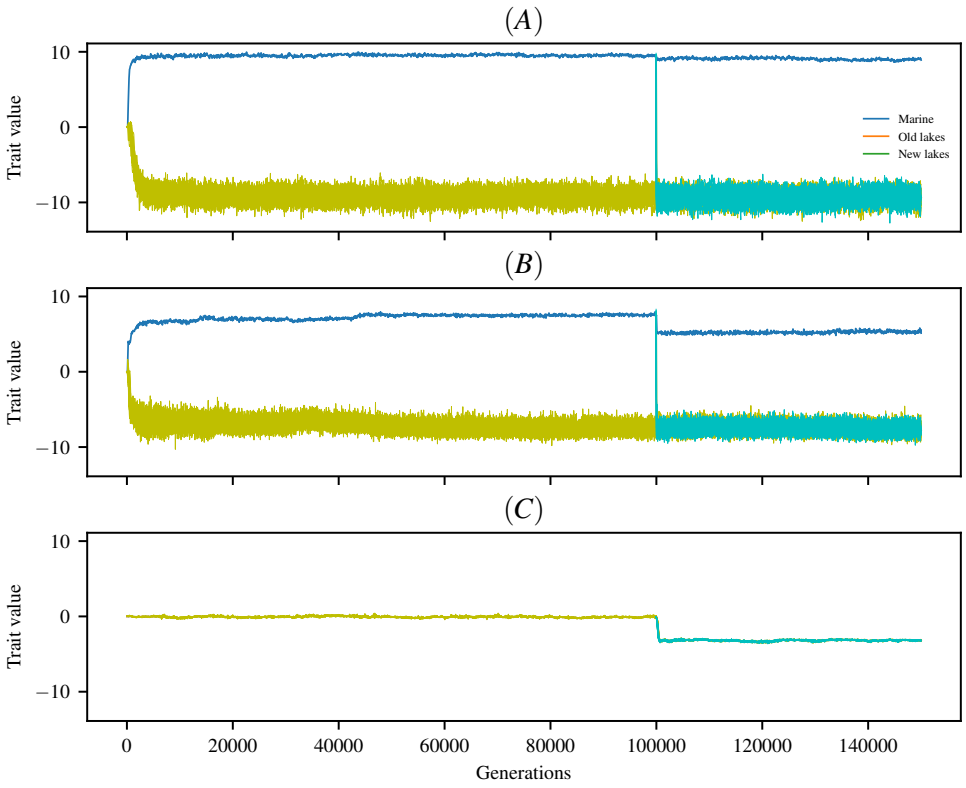

Figure S3: Mean individual trait values in the marine habitat (blue line), the original lakes (yellow lines; average in orange), and the new lakes (light green lines; average in dark green), across the course of two simulations, with migration rates of **(A)** $m = 5 \times 10^{-3}$, **(B)** $m = 5 \times 10^{-2}$ and **(B)** $m = 5 \times 10^{-1}$. (i.e., 1, 10, and 100 migrants per lake per generation, respectively). Optimal trait values in the two habitats are at ± 10.