A few stickleback suffice for the transport of alleles to new lakes

Jared Galloway, William A. Cresko, and Peter Ralph

June 16, 2019

Abstract

Threespine stickleback populations provide a striking example of local adaptation to divergent habitats while still being connected by recurrent gene flow. These small fish occur in marine and freshwater habitats throughout the Northern Hemisphere, and in numerous cases the smaller freshwater populations have been established "de novo" from marine colonists. Previous results have found similar phenotypes between independently evolved freshwater populations are derived largely from the same standing genetic variants found to be Identical-by-decent. Geographic isolation preventing direct migration between the freshwater populations strongly suggests that these shared adaptive alleles are transported through the widely distributed marine population. However it is still largely unknown how gene flow, recombination, and selection jointly impact standing variation to fuel adaptation. Here we use genealogical simulations to determine the levels of gene flow that best match observed patterns of allele sharing among habitats in stickleback. We aim to better understand how gene flow and local adaptation in large metapopulations determine the speed of adaptation and re-use of standing genetic variation. In our simulations we find that repeated adaptation uses a shared set of alleles that are maintained at low frequency by migration-selection balance in oceanic populations. This process occurs over a realistic range of intermediate levels of gene flow that match previous empirical population genomic studies in stickleback. Examining these simulations more deeply reveals that lower levels of gene flow leads to slow, independent adaptation to different habitats, whereas higher levels gene flow leads to significant mutation load – but an increased probability of successful population genomic scans for locally adapted alleles. Surprisingly, we find that the genealogical origins of most freshwater adapted alleles can be traced back to the original generation of marine individuals that colonized the lakes, as opposed to subsequent migrants. These simulations provide deeper context for existing studies of stickleback evolutionary genomics, and guidance for future empirical studies in this model. More broadly, our results support existing theory of local adaptation but extend it by more completely documenting the genealogical history of adaptive alleles in a metapopulation.

Introduction

The canonical model for the genetics of adaptation has long been the sequential fixation of new mutations [Smith and Haigh, 1974, Endler, 1977, Orr, 2005]. While it has proved to be a useful baseline for understanding the genetic variation we see in species today, this model is now rightfully understood as incomplete for many species in nature that have more complicated population structures [Lai et al., 2019, Schrider and Kern, 2017]. In particular, empirical studies have increasingly identified the need to more deeply incorporate standing genetic variation (SGV) into adaptation dynamics for metapopulations inhabiting an array of habitats [Hermisson and Pennings, 2005, Barrett and Schluter, 2008]. Populations experiencing diverse selective pressures while still exhibiting significant gene flow often result in more complex genomic signals that are still not fully understood [Charlesworth, 1997, Charlesworth et al., 2003, Nosil et al., 2009, Flaxman, 2013, Samuk et al., 2017]. Concurrently, a growing number of empirical studies have identified instances of convergent evolution using SGV [Schrider and Kern, 2017, Barrett and Schluter, 2008, Nelson and Cresko, 2018, Nelson et al., 2019, Bassham et al., 2018]. However, it is still not clear how variation in evolutionary processes - such as gene flow, recombination, selection and mutation - can promote the maintenance and re-use of SGV, particularly during colonization and adaptation to new environments [Nelson and Cresko, 2017, Pritchard, 2010, Yeaman and Whitlock, 2011, Schrider and Kern, 2017]. It is similarly unclear whether variation in these evolutionary processes can determine the genetic architecture of evolving traits via SGV. While empirical studies in the field and lab are essential, modeling specific systems can help us understand the nature of standing genetic variants to fuel local adaptation as well as the genomic signals we can expect to see.

The ancestral marine form of threespine stickleback fish has given rise to millions of independently derived freshwater populations in recently de-glaciated regions around the Northern Hemisphere [Hunt et al., 2008, Thompson, 1997, Cresko et al., 2007]. This model organism has provided some of the earliest data in a recent flood showing the heterogeneous nature of divergence across genomes and the much more extensive use of SGV than once thought [Schluter and Conte, 2009, Roesti et al., 2014, Nelson and Cresko, 2018, Bassham et al., 2018, Nelson et al., 2019, Hohenlohe et al., 2010, Terekhanova et al., 2014, Marques et al., 2016]. While geographic isolation often prevents direct migration between freshwater populations, stickleback in them frequently evolve similar phenotypes [Cresko et al., 2004, Colosimo et al., 2004]. The most recent evolutionary genomic studies on stickleback document that while the overall dynamic of local adaptation to marine and freshwater habitats has been occurring for millions of years [Nelson and Cresko, 2018], independent local adaptation of marine individuals to freshwater environments has been observed to take place in just tens of generations [Terekhanova et al., 2014, Lescak et al., 2015, Bassham et al., 2018]. For example, in 1964 the Great Alaskan Earthquake caused an uplift of many islands and coastal regions throughout the Gulf of Alaska. Studies of stickleback populations on uplifted Middleton Island showed that newly created freshwater ponds were invaded by the surrounding marine population of stickleback which evolved the freshwater syndrome of phenotypes in less than 50 years [Lescak et al., 2015, Bassham et al., 2018]. Amazingly, despite very little change in a majority of the genome, nearly a quarter of it diverged significantly in this same time in a pattern that mirrored previous genomic studies of freshwater stickleback populations that have been geographically separated for thousands of years [Hohenlohe et al., 2010, Bassham et al., 2018].

But how can evolution occur at such a rapid pace? Waiting for new mutations to arise in each lake or pond would take much longer than the decades since the Alaskan earthquake. Even more improbable is having divergence cluster in such similar genomic regions across independent populations. An alternative hypothesis, supported by the most recent population genomic analyses, is that the majority of alleles important for freshwater adaptation are maintained in the marine individuals due to recurrent gene flow from freshwater back in to marine populations. In 2009 Schluter and Conte [2009] proposed a conceptual model they termed the "transporter"- hypothesis to describe the process of alleles beneficial in freshwater environment being maintained at migration-selection balance in the larger oceanic population and thus available for subsequent adaptation to new freshwater habitats. The first clear example of the global reuse of SGV in stickleback was the gene *eda* which has been shown to be an important regulator for the number of bony lateral plates [Colosimo et al., 2004, Cresko et al., 2004, Colosimo et al., 2005]. While the low lateral plate version of this gene arose millions of years ago, it is found in much younger freshwater ponds around the Northern Hemisphere [O'Brien et al., 2015]. More recently, genome-wide haplotype analyses have provided evidence that *most* regions of the genome that distinguish marine-freshwater genetic differences share this pattern [Nelson and Cresko, 2017].

While this growing body of population genomic data on stickleback evolution support the transporter hypothesis, a number of questions remain. What are the actual population sizes, migration rates, and fitness differentials consistent with this hypothesis? How many differentially selected alleles exist, how many are used at any one time, and how are they arranged within the genome? A curious natural history observation underlying many of these questions is the fact that some newly formed freshwater habitats, such as the ponds on Middleton Island, are quite small and presumably the number of initial marine migrants is few.

Here, we use forward simulations implemented in SLiM that incorporate linkage and realistic length genomes to model the stickleback metapopulation and address these questions [Haller and Messer, 2017, 2018]. We ask how variation in amount of gene flow affects the genetic architecture of local adaptation to newly created freshwater ponds. Because we are documenting the entire genealogy of all alleles, we can determine the distribution and abundance of haplotypes across all in marine and freshwater populations to know the timing and proportion of potentially adaptive alleles that are actually utilized in each population. In addition, we can document how these adaptive alleles are distributed across the genomes, and as a consequence determine how this may affect the efficacy of genome scans for between-habitat differentiation. We find that only a few stickleback migrants per lake per generation are sufficient to maintain SGV in the ocean, allowing new freshwater populations to adapt in only tens of generations. We also find that continued gene flow of marine individuals into recently established lakes is not necessary for adaptation as the majority of alleles are present in the initial marine colonizers. Finally, we find that this rapid, local adaptation can only occur over an intermediate range of levels of gene flow. Low gene flow prevents variation from being transported, while high gene flow prevents local adaptation. Together these simulation findings enhance and extend the existing rich empirical work on stickleback evolutionary genomics.

Methods

To explore these questions, we used SLiM [Haller and Messer, 2017, 2018] to implement complex forward-time simulation of populations of individuals with explicitly represented genomes in which selection acted upon a single continuous quantitative trait. The details of the model were motivated by current understanding of threespine stickleback evolutionary history and demography, and included divergent selection in the two habitats which had substantial spatial structure. Certain aspects of 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).

I don't know where we should talk about our small population sizes ... Is this even worth

Habitat and geography Our simulations include two habitat types – Marine and Freshwater – defined by the nature of their selective pressures, each with 5,000 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 thought to be under divergent selection between the habitats, such as armor morphology, body size, craniofacial variation and opercle shape (among many other traits). 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} is determined by a Gaussian kernel with standard deviation 15, i.e.,

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

Thus, the difference between an individual's trait value and the optimum determines that individual's fitness. 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^{-10} per locus per generation in ten 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 $x_{\mathrm{ind}} = \sum_{i \in H} h_i s_i + \sum_{j \in D} s_j$, 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^{-8} per locus per generation). Since the total population across *all* 25 lakes is fixed at 5000, and the Wright–Fisher model assumes unrealistic global population regulation, 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.

As depicted in Figure 1, dispersal occurs both locally along the coastline in the marine habitat as well

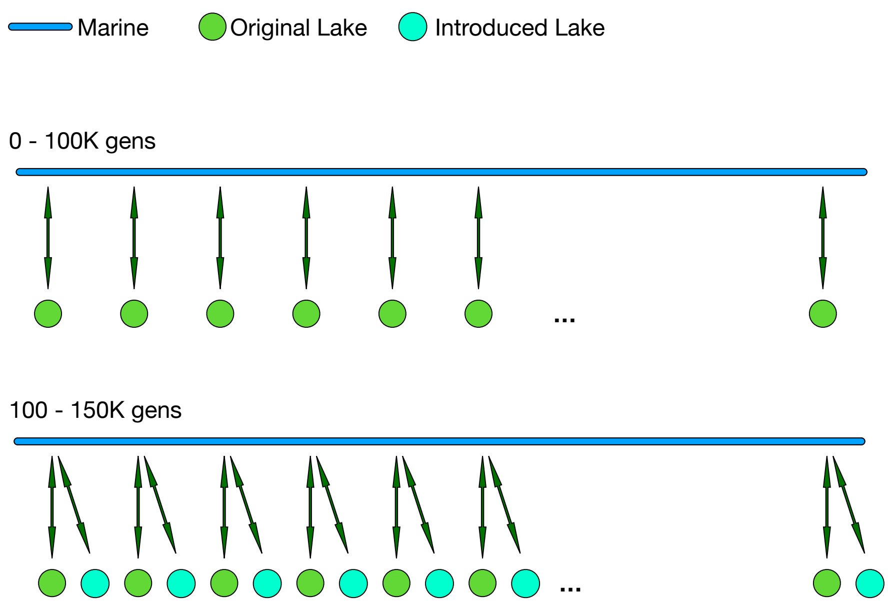

Figure 1: **Diagram of simulated populations:** a single, continuous, one-dimensional marine habitat (blue) is coupled to "lakes", in which stickleback randomly mate, at discrete locations with arrows representing migration patterns. After an initial period of 100K generations with 25 lakes, an additional 25 lakes are added (at the same set of locations) and populated with 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.

as between the marine habitat and the lakes, with a lake–ocean migration rate denoted m. No migration occurs among the freshwater populations. All individual dispersal events can be thought of as occurring at the juvenile stage in the life cycle of the simulation. Each new individual in each habitat has parents from the other habitat with probability m (in which case we call it a "migrant"), and parents from the same habitat with probability $1 - m$. The first parent of each non-migrant individual in the freshwater habitat is chosen from the freshwater habitat proportional to fitness, and a mate is chosen from the same lake as the first, also proportional to fitness. The offspring then lives in the same lake as the parents. Parents for each non-migrant marine individual are chosen similarly: 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 $f(x_j)$ is the fitness of individual j and 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 0.5, and reflected to stay within the habitat range. Parents for each freshwater migrant are chosen in the same way as for non-migrant marine individuals, and are assigned to the lake nearest to the position of the first marine parent. Similarly, parents for each marine migrant are both chosen from the same lake as before, and the offspring is given a spatial location in the marine habitat at the location of the parent's lake.

Colonization of newly formed lakes To study how marine-derived populations adapt after colonizing newly appearing freshwater habitats, we introduce a new set of 25 lakes after 100,000 generations. These new lakes are populated with marine individuals to emulate a freshwater lake being colonized by oceanic stickleback that had previously been exchanging alleles with older freshwater populations. This creates two *sets* of lakes, which along with the marine population have a total of 15,000 individuals. 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.

Recording genealogical history We used SLiM's ability to record *tree sequences* [Haller et al., 2018, Kelleher et al., 2016] 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^{-8} 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 not in the initial generation that colonized the lake

3. *a "Captured" allele:* present in initial colonists of 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 initial colonists of the new lake, but not a "captured" allele.

The proportion of trait-affecting alleles in new lakes that fall in these categories measures the degree to which 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.

We used neutral mutations to calculate F_{ST}, calculated on a per-locus basis and then averaged in windows as a measure of between-population relative differentiation. 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 $1 - p_f p_m / (\bar{p}(1 - \bar{p}))$.

Results & Discussion

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}. Below, we often refer to these migration rates in terms of the number of migrants per lake, per generation, which we denote M. Since each lake contains 200 individuals, $M = 200m$. 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 almost completely independently through *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 locally maladaptive alleles [Bolnick and Nosil, 2007]. Between these two extremes, genetic variation that allowed adaptation to freshwater habitats could move relatively easily between lakes. Perhaps surprisingly, only a few migrants per generation from the lakes to the ocean were needed to maintain sufficient genetic variation in the ocean to dramatically accelerate adaptation in new lakes.

In general I suggest having the finding or statement at the beginning of the sentence with a reference to the figure or statistic at the end of the sentence. Notice that in many cases below (e.g. first sentence of the first paragraph in the 'Local Adaptation' section) the sentence starts with the Figure (e.g. 'As shown in Figure 2'). You can just start with the declarative statement (e.g.

Local Adaptation: differentiation with gene flow

As shown in Figure 2, local adaptation occurred in all simulations with the exception of the highest gene flow at which half of each population was composed of migrants. Figures 3 and S3 show how mean trait values in freshwater and marine populations diverged over time, until the trait means were close to the optimal values in each habitat. Establishment of new alleles in the lakes is visible in the first few thousand generations as jumps in the mean trait value which move the trait by an amount of order 1 every few hundred generations (Fig. 3). At the lowest rate of gene flow, $M = 0.01$, differences at around 20 commonly polymorphic sites (about 10 that shift the trait in each direction) were responsible for most of the adaptive differences between freshwater and marine habitats. 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.

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

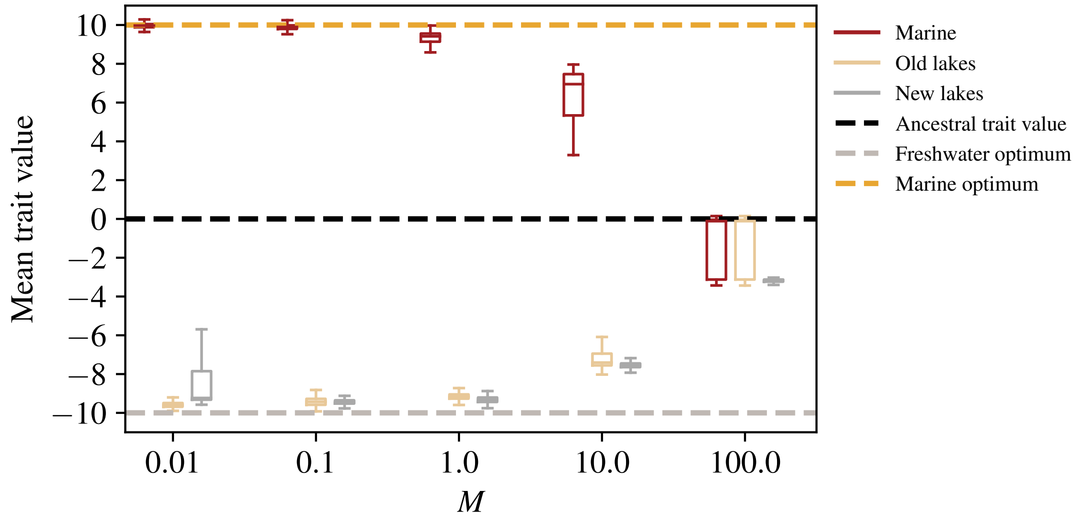

Figure 2: Distribution of mean individual trait values across generations of the simulation, for different migration rates. The dashed pink and purple lines at ± 10 give the optimum phenotypes in the marine and freshwater environments, respectively.

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.

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, each lake must wait for its own novel mutations to arise in order to adapt. Because the marine habitat is continuous, 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 for adaptation.

same comment as above, is there a declarative statement of finding? e.g. Freshwater adapted alleles are shared extensively and primarily via initial migrants

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, an allele must participate in the genetic basis of local adaptation for at least one of the original lakes. Concretely, these are any non-neutral mutations whose frequency is above 50% in at least one original lake 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: new lakes independently acquire new adaptive alleles rather than pre-existing ones. At $M = 0.1$, we again observe the original lakes adapting nearly independently from each other, but now the new lakes adapt using pre-existing alleles present in the original set of lakes. Concurrently, the average marine individual carries ≈ 2 pre-existing freshwater adapted alleles, standing variation which was nearly absent at $M = 0.01$. As

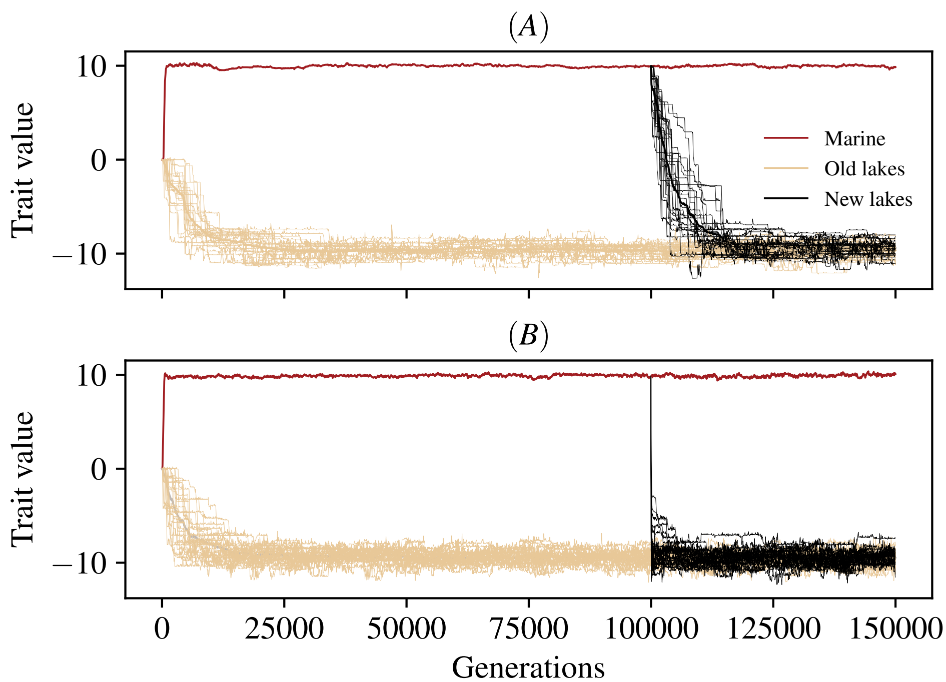

Figure 3: Mean individual trait values in the marine habitat (blue line), the original lakes (yellow lines; average in orange), and the new lakes (light green lines; average in dark green), across the course of two simulations, with migration rates of **(top)** $M = 0.01$ and **(bottom)** $M = 0.1$ migrants 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.

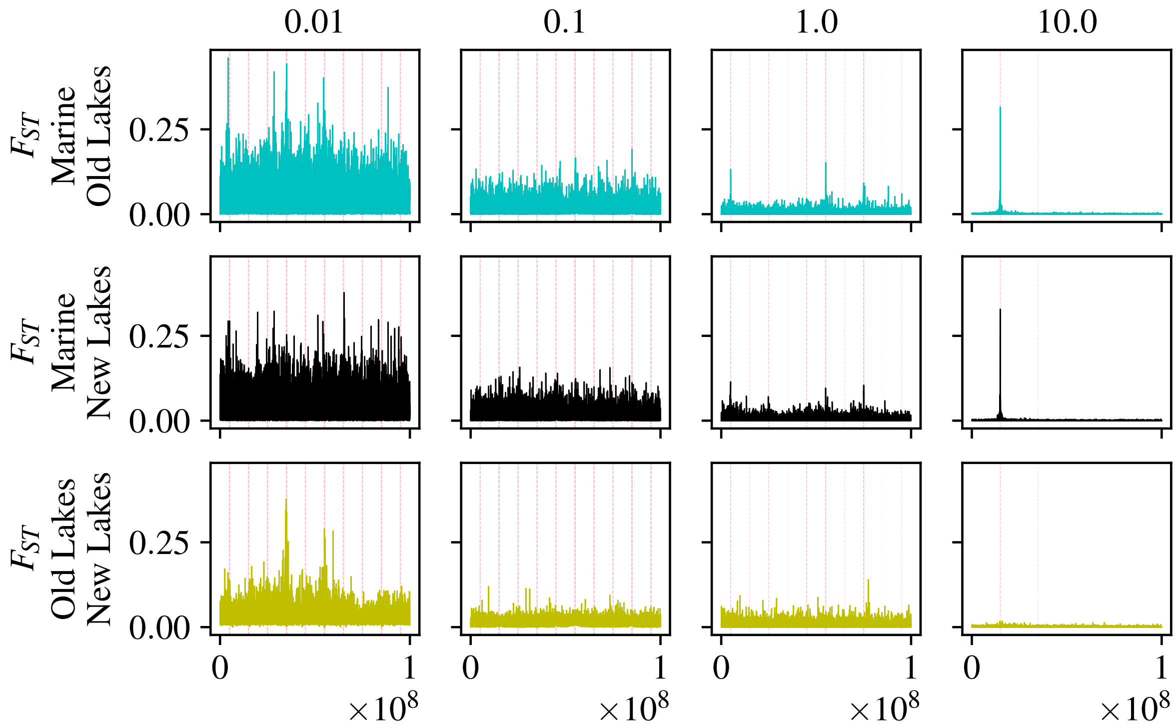

Figure 4: Average F_{ST} in windows of 500Mb between: **(top)** marine habitat and old lakes; **(middle)** marine habitat and new lakes; and **(bottom)** old and new lakes. Each plot shows F_{ST} values for a separate simulation, with columns corresponding to increasing gene flow from left to right. F_{ST} is calculated between all marine individuals, and all lakes pooled together. 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 pink imply more freshwater adapted alleles at that location.

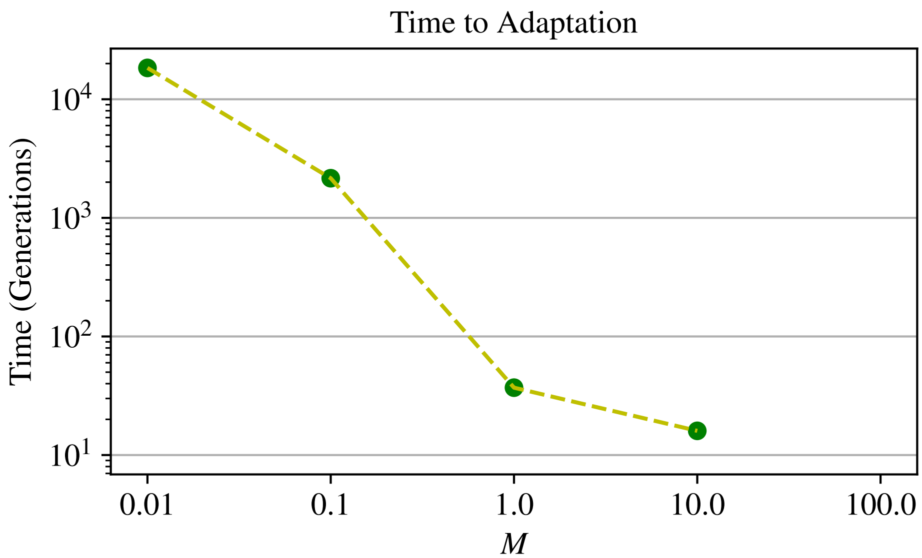

Figure 5: Time to adaptation as a function of migration rate. The time to adaptation is measured as the number of generations until the introduced population's mean phenotype comes within 0.5 of the original lakes average phenotype. Each point represents a single simulation run. (Adaptation did not occur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

migration rate increases past this, the total number of pre-existing freshwater adapted alleles declines (Figure 6C). 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 larger effect to travel more efficiently through the population, even though they are deleterious in the ocean.

The numbers in Figure 6 strongly suggest that the dramatic increase in speed of local adaptation we observed above occurs because higher gene flow allows sharing of freshwater alleles among populations. We confirmed this by using recorded tree sequences to identify the origin of each trait-affecting allele in each individual in the new lakes, 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

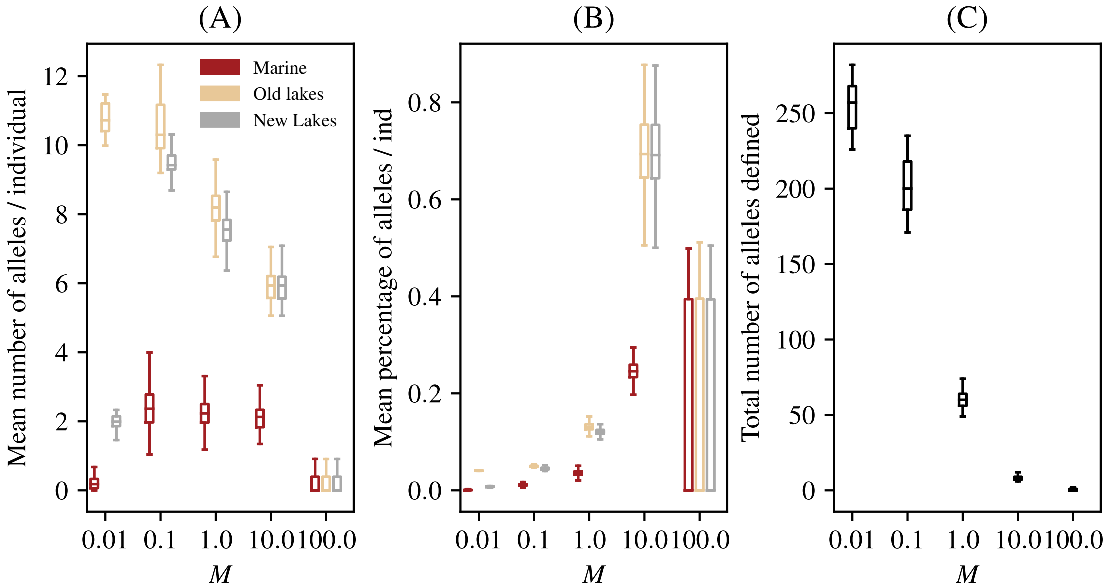

Figure 6: Amount of standing freshwater variation by habitat, across migration rates. Each plot counts "pre-existing freshwater adapted alleles", that are common in the original lakes but rare in the ocean (see text for definition). **(A)** Mean number of these alleles per individual. **(B)** Mean percentage 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 throughout each simulations. The horizontal axis shows M, the mean number of migrants per lake per generation.

their own genetic basis of adaptation.

While increased migration allows sharing of adaptive alleles among 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. These results present the sensitivity of a population to the amount of gene flow it experiences from neighboring species. In our simulations, the continued gene flow from freshwater populations meant that the alleles were maintained at a constant rate with fixed demographics. However, future modeling efforts must be made to understand what population sizes, demographics, and evolutionary histories impact an ancestral population's capacity for carrying maladaptive alleles in its opposing environment, as well as what role a haplotype's genomic architecture play it to be rebuilt in a novel environment.

The efficacy of genomic scans of selection depends on gene flow

just made a suggested change

Here, we take a closer look at the genomic architecture of local adaptation between the two habitats. Can measures of local differentiation such as F_{ST} be used to identify the causal loci? Figure 4 shows plots along the genome of average per-locus F_{ST} values in 500bp windows between the marine habitat and all freshwater habitats pooled together. Higher rates of dispersal showed more distinct F_{ST} peaks over polymorphic loci, while "background" levels of F_{ST} increase as gene flow decreases, swamping out this signal until the regions

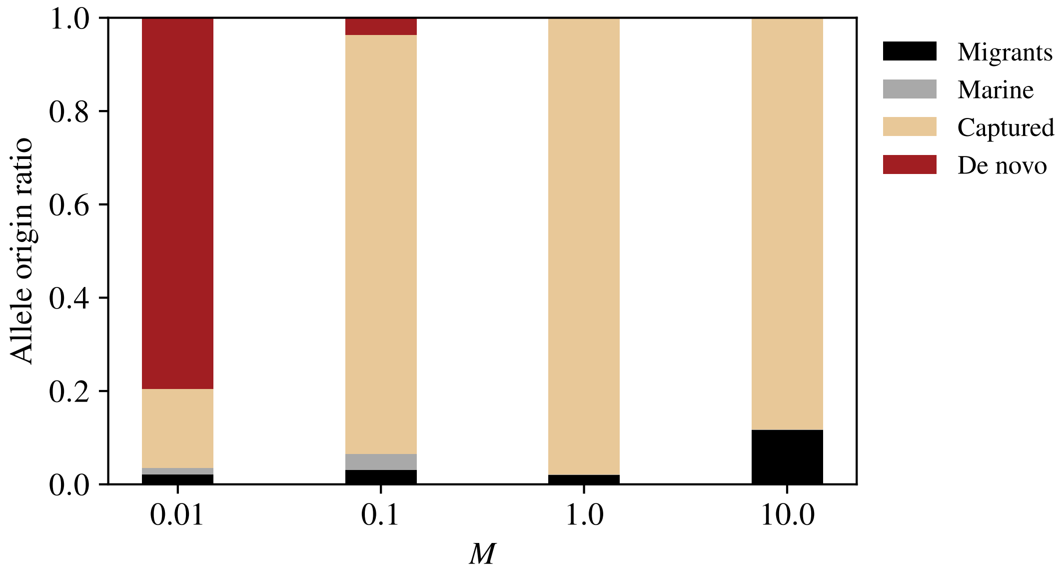

Figure 7: **(Origin of adaptive alleles:)** Each bar plot shows the origins of all trait-affecting alleles above frequency 50% in at least one new lake, classified as **(red)** new mutations, **(blue)** post-colonization migrants, **(green)** "captured" from pre-existing lakes, or **(orange)** standing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

under selection are indistinguishable. This is likely the contribution of two separate forces of natural selection: 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.

At first glance, this suggests that genome scans for local adaptation based purely on measures of differentiation will only be successful given enough migration between habitats. But how many of these peaks are actually underlying trait differences that form the basis for local adaptation? 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.

What we observe most plainly in this graph is that $M = 1.0$ and $M = 10.0$ migrants per lake, per generation are the only rates of dispersal which provide reliable true negatives when observing peaks. This means that a large percentage of peaks past an F_{ST} cutoff of 0.05 actually lie on top of regions which could impact trait differences. Unfortunately, the low statistical power for all rates of M would seem to suggest that many regions which impact trait differences do not appear as F_{st} peaks. This may imply that genome scans for local adaptation based purely on measures of differentiation may not be as reliable as previously thought. We therefore find that with high rates of dispersal between population, F_{ST} peaks may be a good indication of causal loci that are identified, but that not all causal loci will appear as F_{ST} peaks. It's important to note that our study in Figures 4 and S1 pooled together all lakes and all marine individuals for the calculations. For comparison, Figures S4 & S2 shows the same scans but calculated only between individuals in the marine with spacial location ≤ 5.0 and the individuals in the respective 5 lakes. These graphs show even less power and True Positives only at $M = 10.0$ migrants per lake, per generation

This study shows introgression is beneficial for inferring causative loci from divergence (F_{ST}) along the genome. This is generally because noise of neutral 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. In general, our study may suggest that diversity scans may not prove as useful in causal loci inference as previously thought. It is important to note that these summary statics could prove to be very useful if used beyond just the peaks seen when looked at as a function of genomic position.

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, the *origin* of the alleles underlying the phenotype is still unknown. What is the nature of selection on standing genetic variants? In our simulations migration from marine individuals into the new lakes continues throughout introduction and adaptation: but is consistent influx of SGV a necessity for rapid adaptation of the population? To answer this, we traced the genealogy of *all* adaptive alleles in *all* individuals from the introduced lakes, after local adaptation. Surprisingly, for all cases with the exception of $M = 0.01$, we found that the majority of adaptive allele origins traced back the original generation of inhabitants in the lake as seen in Figure 7. This suggests that given enough SGV, any reasonable subset of the ancestral population has the potential for rapid adaptation without the need post-colonization hybridization events.

same thing here - a statement of the major finding of this section?

If effectively capturing standing genetics variants is the key to rapid adaptation, as we have presented thus far, why is there a large difference in speed of adaptation between dispersal rates of $M = 0.1$ and $M = 1.0$ seen in Figure 5. This is surprising because we see a similar quantity of allele sharing at $M = 1.0$, but much more rapid adaptation (≈ 60 generations).

	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, **"Best"** quantified as the most negative trait value achievable with intact haplotypes, averaged across populations (see text for details).

A possible 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. This implies the adaptation in the new lakes must wait for recombination to separate freshwater and marine adapted alleles. This might be expected at low dispersal rates since freshwater alleles may need to be "masked" by nearby compensatory alleles to remain in the ocean for long periods of time. Reversing this "masking" then slows the process of adaptation that uses this genetic variation.

In other words, higher dispersal 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 only once in every one thousand meioses. This implies that 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), rebuilding the most beneficial haplotype may prove to be a non-trivial task for selection. 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. This might also explain why at $M = 0.1$, marine individuals still hold on average several alleles that shift the trait in the freshwater direction. (Figure 6A).

To quantify the genetic variation available *without* recombination 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 these regions. The mean of this value across the 25 lake populations is shown in 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 in the populations at $M = 1.0$ have effect sizes nearly twice as big at a mean total of -21.02. The amount of variation available at $M = 10.0$ is lower (only -7.02), presumably because of migration load, while at $M = 0.01$ almost no alleles with negative effect are present. Note that some of these haplotypes will likely be lost to drift – indeed, if they did all fix, then populations at $M = 0.1$ would adapt much more quickly. However, this calculation still provides contrast and a reasonable explanation for the difference in time to adaptation.

Alignment with theoretical expectations

We now revisit our observations above in the light of population genetics theory. The discussion will be roughly based on estimates, with a focus on intuition rather than precise calculations. Because we model stabilizing selection on an additive trait controlled by a moderately large number of loci within each population, 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 $(1 + s)n$ it is 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 populations all differ from the optimum trait by z, and the allele has effect size $-u$ in heterozygotes, then the fitness advantage of the allele would be $s(u) = \exp(-\beta((z-u)^2)/\exp(-\beta z^2)) \approx 2\beta z u$, 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 six categories: additive, dominant, and recessive, in either direction. This implies that a new additive or dominant effect mutation appears once every 75 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 $\int_0^\infty 4\beta z u \exp(-2u) du = \beta z$, and so the rate of establishment of dominant mutations is $\beta z/75$, i.e., about one such mutation every $33750/z$ generations. At the beginning of the simulation where $z = 10$, we would then expect the fixation of alleles to take around 3,000 generations. The distribution of the effect sizes of these successfully established mutations

This subsection is a little incongruous with the others sections with respect to how it is written. Seems that the bold headers for the paragraphs was residual from notes? Also, this section seems to be in a different voice from the rest of the Results/Discuss section. I think it's worth a bit of time trying to hone it 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 3,000 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 3 and S3.

Standing variation An allele that moves the trait u units in the freshwater direction, in heterozygotes, has fitness roughly $\exp(-\beta u^2) \approx 1 - \beta u^2$ in the marine environment, i.e, a fitness differential of $s = \beta u^2$. The product of population size and fitness differential in the marine environment for a mutation with $u = 1$ is therefore $2Ns \approx 22$, implying that these alleles are strongly selected against but might drift to moderate frequency if recessive. The average frequency of such an allele in the marine environment at migration-selection equilibrium is equal to the proportion of individuals in the ocean replaced by migrants per generation divided by the selective disadvantage, i.e., around $m/\beta u^2$. Each new lake is likely to contain a few copies of alleles at frequency above 1/400 (since each new lake is initialized with 400 randomly selected genomes). With $u = 1$, this occurs if $m/\beta > 1/400$, and so since $1/\beta = 450$, if $m \geq 5 \times 10^{-6}$ 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 $u = 1$ only has probability of around 1/20 of establishing locally, so an allele of this size must be present in about 20 copies to ensure adaptation. Putting these together, we expect migration-selection balance to maintain sufficient genetic variation for new lakes to adapt if $m \geq 10^{-4}$, which corresponds to $M \geq 0.4$, agreeing with our results. However, this calculation treats each allele independently; in practice we found that standing freshwater variation in the ocean was masked by linkage to compensatory marine variants.

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.

Conclusion

In this paper, we have described a realistic simulation model of a coastal meta-population to observe the effects of selection on genetic variation across a wide range of gene flow rates. We have shown that over a realistic range of parameter values historical introgression is able to reproduce rapid and parallel adaptation similar to that documented in natural stickleback populations, such as Middleton island (cites), and supporting the "Transporter"-hypothesis suggested by Schluter and Conte [2009]. 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 contributed to the efficiency of local adaptation except the highest due to migration load. A finding of practical consequence is that the efficacy of F_{ST} based genomic scans of selection varies significantly with migration rate. A surprising new finding of our simulations is that the majority of adaptive variant's genealogy can be traced back to the original generation of inhabitants as opposed to post-colonization migrant alleles, and thus a randomly chosen subset of marine stickleback carry the capacity to rapidly adapt without *any* continued connection to the rest of of the species. Thus, the large marine population was able to harbor and distribute deleterious alleles to the surrounding freshwater populations with only a few migrants per population, per generation. Our simulation findings significantly extend previous empirical and theoretical population genomic studies of stickleback to make testable predictions, and provide a framework for similar studies in other organisms to determine the generality of our findings.

References

Rowan D. H. Barrett and Dolph Schluter. Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, 23(1):38–44, Jan 2008. ISSN 0169-5347. doi: 10.1016/j.tree.2007.09.008. URL `https://doi.org/10.1016/j.tree.2007.09.008`.

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`.

Susan Bassham, Julian Catchen, Emily Lescak, Frank A. von Hippel, and William A. Cresko. Repeated selection of alternatively adapted haplotypes creates sweeping genomic remodeling in stickleback. *Genetics*, 209(3):921–939, 2018. ISSN 0016-6731. doi: 10.1534/genetics.117.300610. URL `https://www.genetics.org/content/209/3/921`.

Daniel I. Bolnick and Patrik Nosil. Natural selection in populations subject to a migration load. *Evolution*, 61(9):2229–2243, 2007. doi: 10.1111/j.1558-5646.2007.00179.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2007.00179.x`.

Brian Charlesworth. A big book and a small book on selection. *Evolution*, 51(4):1337–1338, 1997. doi: 10.1111/j.1558-5646.1997.tb03983.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1997.tb03983.x`.

Brian Charlesworth, Deborah Charlesworth, and Nicholas H. Barton. The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution, and Systematics*, 34(1):99–125, 2003. doi: 10.1146/annurev.ecolsys.34.011802.132359. URL `https://doi.org/10.1146/annurev.ecolsys.34.011802.132359`.

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`.

Pamela F Colosimo, Catherine L Peichel, Kirsten Nereng, Benjamin K Blackman, Michael D Shapiro, Dolph Schluter, and David M Kingsley. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLOS Biology*, 2(5), 03 2004. doi: 10.1371/journal.pbio.0020109. URL `https://doi.org/10.1371/journal.pbio.0020109`.

Pamela F. Colosimo, Kim E. Hosemann, Sarita Balabhadra, Guadalupe Villarreal, Mark Dickson, Jane Grimwood, Jeremy Schmutz, Richard M. Myers, Dolph Schluter, and David M. Kingsley. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, 307(5717):1928–1933, 2005. ISSN 0036-8075. doi: 10.1126/science.1107239. URL `https://science.sciencemag.org/content/307/5717/1928`.

William A. Cresko, Angel Amores, Catherine Wilson, Joy Murphy, Mark Currey, Patrick Phillips, Michael A. Bell, Charles B. Kimmel, and John H. Postlethwait. Parallel genetic basis for repeated evolution of armor loss in alaskan threespine stickleback populations. *Proceedings of the National Academy of Sciences*, 101 (16):6050–6055, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0308479101. URL `https://www.pnas.org/content/101/16/6050`.

William A. Cresko, Katrina L. McGuigan, Patrick C. Phillips, and John H. Postlethwait. Studies of threespine stickleback developmental evolution: progress and promise. *Genetica*, 129(1):105–126, Jan 2007. ISSN 1573-6857. doi: 10.1007/s10709-006-0036-z. URL `https://doi.org/10.1007/s10709-006-0036-z`.

John Endler. Geographic variation, speciation and clines. *Monographs in population biology*, 10:1–246, 02 1977. doi: 10.2307/2259234.

Samuel M. Flaxman. Surfing downhill: when should population range expansion be characterized by reductions in fitness? *Molecular Ecology*, 22(24):5963–5965, 2013. doi: 10.1111/mec.12564. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12564`.

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`.

Joachim Hermisson and Pleuni S. Pennings. Soft sweeps. *Genetics*, 169(4):2335–2352, 2005. ISSN 0016-6731. doi: 10.1534/genetics.104.036947. URL `https://www.genetics.org/content/169/4/2335`.

Paul A. Hohenlohe, Susan Bassham, Paul D. Etter, Nicholas Stiffler, Eric A. Johnson, and William A. Cresko. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLOS Genetics*, 6(2):1–23, 02 2010. doi: 10.1371/journal.pgen.1000862. URL `https://doi.org/10.1371/journal.pgen.1000862`.

Gene Hunt, Michael A. Bell, and Matthew P. Travis. Evolution toward a new adaptive optimum: Phenotypic evolution in a fossil stickleback lineage. *Evolution*, 62(3):700–710, 2008. doi: 10.1111/j.1558-5646.2007.00310.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2007.00310.x`.

Jerome. Kelleher, Alison M. Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, 12(5):e1004842, May 2016. doi: 10.1371/journal.pcbi.1004842. URL `https://doi.org/10.1371/journal.pcbi.1004842`.

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`.

Yu-Ting Lai, Carol K. L. Yeung, Kevin E. Omland, Er-Li Pang, Yu Hao, Ben-Yang Liao, Hui-Fen Cao, Bo-Wen Zhang, Chia-Fen Yeh, Chih-Ming Hung, Hsin-Yi Hung, Ming-Yu Yang, Wei Liang, Yu-Cheng Hsu, Cheng-Te Yao, Lu Dong, Kui Lin, and Shou-Hsien Li. Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences*, 116(6):2152–2157, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1813597116. URL `https://www.pnas.org/content/116/6/2152`.

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`.

David A. Marques, Kay Lucek, Joana I. Meier, Salome Mwaiko, Catherine E. Wagner, Laurent Excoffier, and Ole Seehausen. Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLOS Genetics*, 12(2):1–34, 02 2016. doi: 10.1371/journal.pgen.1005887. URL `https://doi.org/10.1371/journal.pgen.1005887`.

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`.

Thomas C. Nelson and William A. Cresko. Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. *Evolution Letters*, 2(1):9–21, 2018. doi: 10.1002/evl3.37. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/evl3.37`.

Thomas C. Nelson, Matthew R. Jones, Jonathan P. Velotta, Abhilesh S. Dhawanjewar, and Rena M. Schweizer. Unveiling connections between genotype, phenotype, and fitness in natural populations. *Molecular Ecology*, 28(8):1866–1876, 2019. doi: 10.1111/mec.15067. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15067`.

Patrik Nosil, DANIEL J. Funk, and DANIEL ORTIZ-BARRIENTOS. Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, 18(3):375–402, 2009. doi: 10.1111/j.1365-294X.2008.03946.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2008.03946.x`.

Eleanor O'Brien, John Hunt, and Jason B. Wolf. Quantitative genetic variation and heritability. *Oxford Bibliographies in Evolutionary Biology*, 2015. doi: 10.1093/OBO/9780199941728-0064.

H. Allen Orr. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, 6(2):119–127, 2005. ISSN 1471-0064. doi: 10.1038/nrg1523. URL `https://doi.org/10.1038/nrg1523`.

Jonathan K. Pritchard. How we are evolving. *Scientific American*, 303(4):41–47, 2010. URL `http://www.jstor.org/stable/26002210`.

Marius Roesti, Sergey Gavrilets, Andrew P. Hendry, Walter Salzburger, and Daniel Berner. The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology*, 23(16):3944–3956, 2014. doi: 10.1111/mec.12720. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12720`.

Kieran Samuk, Gregory L. Owens, Kira E. Delmore, Sara E. Miller, Diana J. Rennison, and Dolph Schluter. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular Ecology*, 26(17):4378–4390, 2017. doi: 10.1111/mec.14226. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14226`.

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`.

Daniel R. Schrider and Andrew D. Kern. Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, 34(8):1863–1877, Aug 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx154. URL `http://dx.doi.org/10.1093/molbev/msx154`.

John Maynard Smith and John Haigh. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23 (1):23?35, 1974. doi: 10.1017/S0016672300014634.

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.

Nadezhda V. Terekhanova, Maria D. Logacheva, Aleksey A. Penin, Tatiana V. Neretina, Anna E. Barmintseva, Georgii A. Bazykin, Alexey S. Kondrashov, and Nikolai S. Mugue. Fast evolution from precast bricks: Genomics of young freshwater populations of threespine stickleback gasterosteus aculeatus. *PLOS Genetics*, 10(10):1–13, 10 2014. doi: 10.1371/journal.pgen.1004696. URL `https://doi.org/10.1371/journal.pgen.1004696`.

John N. Thompson. Evaluating the dynamics of coevolution among geographically structured populations. *Ecology*, 78(6):1619–1623, 1997. doi: 10.1890/0012-9658(1997)078[1619:ETDOCA]2.0.CO;2. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/0012-9658%281997%29078%5B1619%3AETDOCA%5D2.0.CO%3B2.

S Yeaman and M C Whitlock. The genetic architecture of adaptation under migration-selection balance. *Evolution*, 65(7):1897–1911, July 2011. doi: 10.1111/j.1558-5646.2011.01269.x. URL http://www.ncbi.nlm.nih.gov/pubmed/21729046.

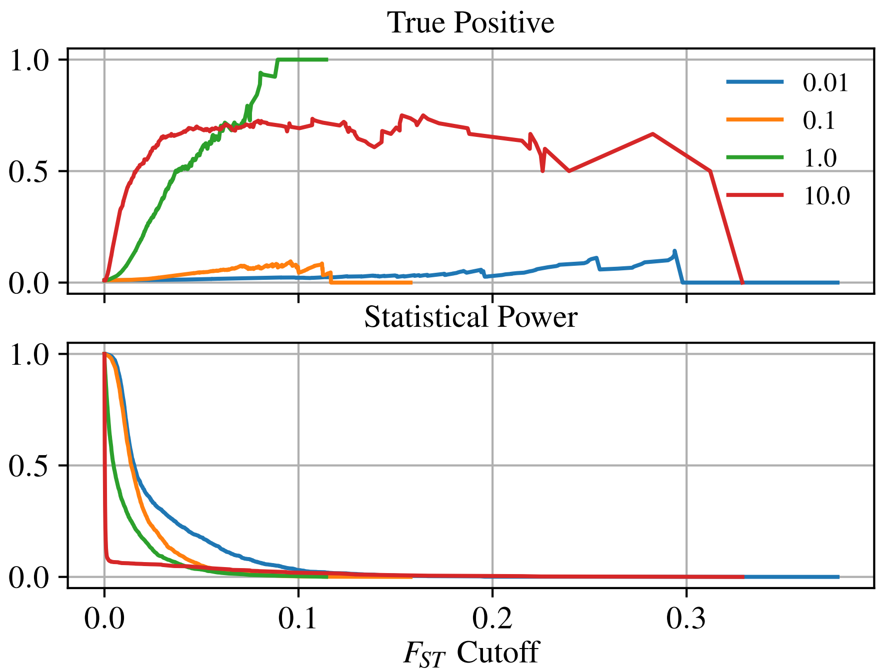

Figure S1: Statistical Power and False Positives as a function of F_{ST} threshold. Statistical Power is the likelihood that a SNP will be predicted to have an effect on phenotype when there is an effect to be detected. False Positives give us the ratio of SNPs that effect phenotype to total SNPs greater than the F_{ST} threshold.

Supplementary material

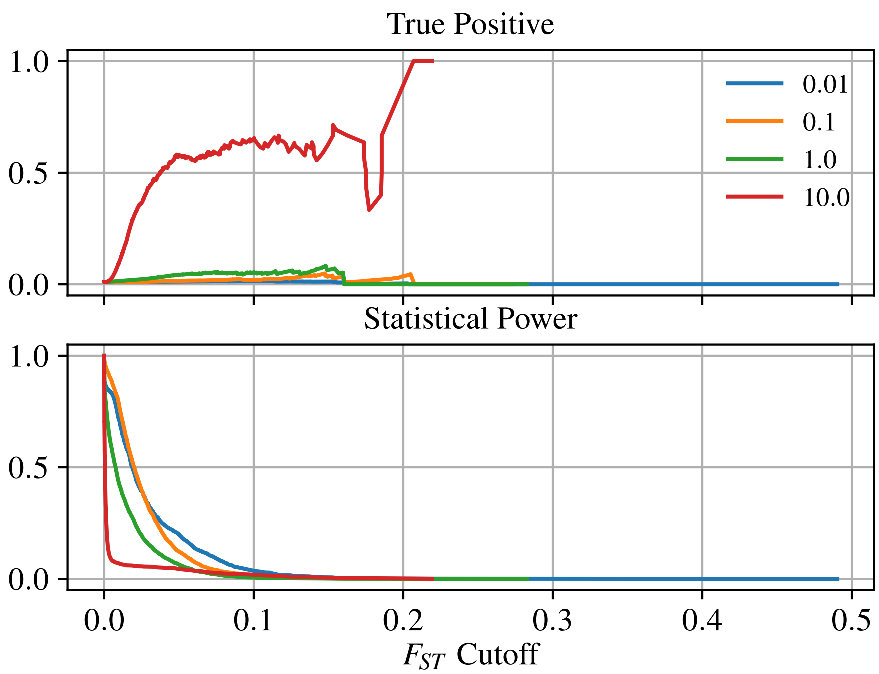

Figure S2: Statistical Power and False Positives as a function of F_{ST} threshold. Statistical Power is the likelihood that a SNP will be predicted to have an effect on phenotype when there is an effect to be detected. False Positives give us the ratio of SNPs that effect phenotype to total SNPs greater than the F_{ST} threshold.

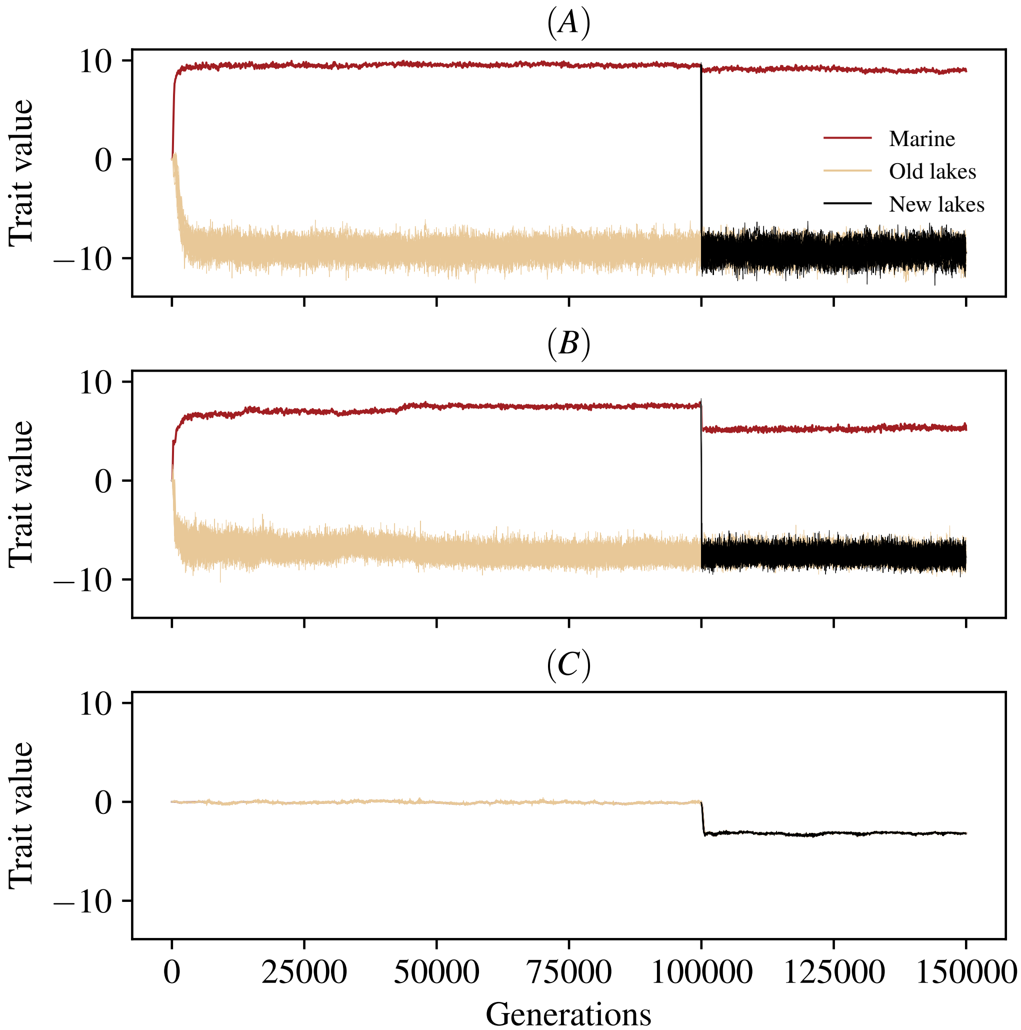

Figure S3: Mean individual trait values in the marine habitat (blue line), the original lakes (yellow lines; average in orange), and the new lakes (light green lines; average in dark green), across the course of two simulations, with migration rates of **(A)** $m = 5 \times 10^{-3}$, **(B)** $m = 5 \times 10^{-2}$ 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.

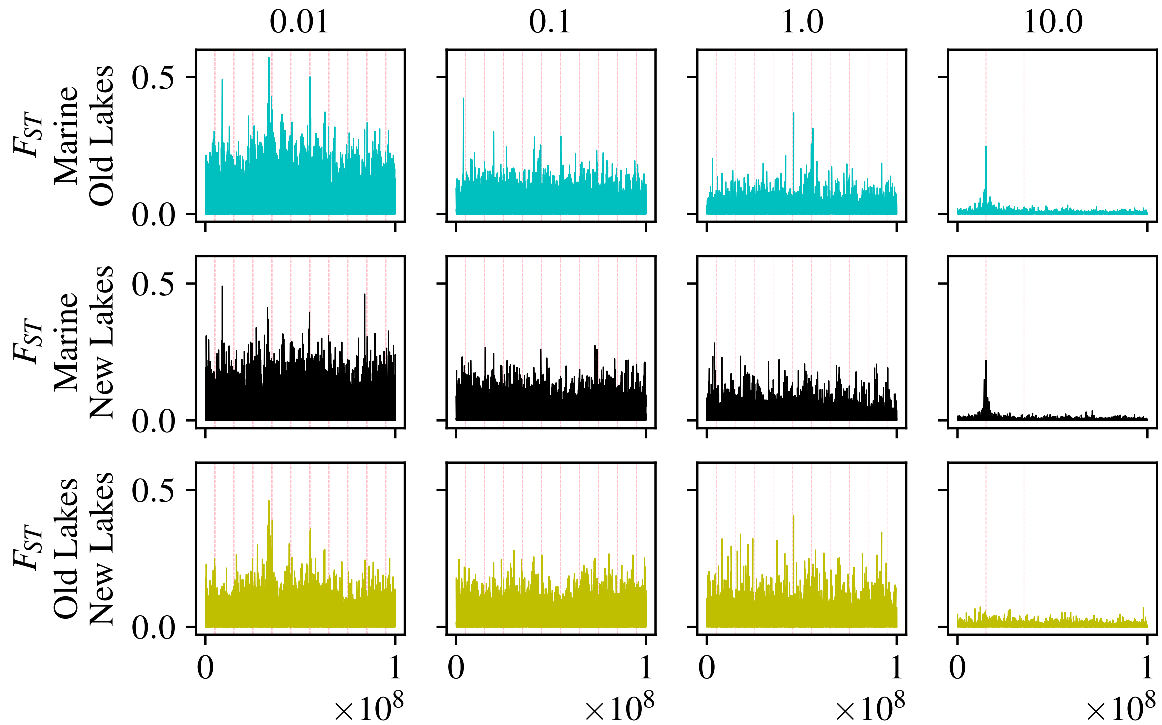

Figure S4: Average F_{ST} in windows of 500Mb between: **(top)** marine habitat and old lakes; **(middle)** marine habitat and new lakes; and **(bottom)** old and new lakes. Each plot shows F_{ST} values for a separate simulation, with columns corresponding to increasing gene flow from left to right. F_{st} is calculated between all marine individuals with spatial location ≤ 5.0, and the five corresponding lakes. 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.