The sweet spot for local polygenic adaptation: a case study motivated by stickleback

Jared Galloway, William A. Cresko, and Peter Ralph

August 17, 2018

Other title ideas? (Bad ideas allowed.)
How do stickleback share adaptations between lakes? A simulation study.
A simulation study of gene flow and local polygenic adaptation

A simulation study of gene flow and local polygenic adaptation in stickleback fish

JARGON WE SHOULD THINK ON ??
effect loci (locus) — QTL — causal: = ?
Original Lakes — initial — P_1, P_3, P_3 = ?
Migration Rate — Historical Introgression = ?
pre-existing freshwater adapted alleles — standing genetic variants = ?
Pending things to do? that maybe we do, maybe not
longer genomes (10^8 not 10^5 but rescaling everything accordingly)
verify that adaptation of introduced pops is due to initial diversity not subsequent migration

1 Abstract

Threespine stickleback fish provide one of the most striking examples of local adaptation. This hemisphere-wide metapopulation includes both marine populations and a large number of smaller freshwater populations that have apparently repeatedly adapted to freshwater conditions often using the same genetic basis. In this paper, we use simulations motivated by stickleback populations to examine what amounts of gene flow favor stable metapopulation polymorphism with allele sharing, and to further dissect the underlying dynamics. We find that rapid, repeated adaptation using alleles maintained at low frequency by migration-selection balance (the "transporter hypothesis") occurs over a realistic range of intermediate rates of gene flow, between slow, independent adaptation with low gene flow and large migration load at high gene flow. This is mediated mainly by the total downstream influx of alleles *(maybe??)*. We do not see evidence for strong effects encouraging genomic clustering of causal alleles or towards particular dominance coefficients. F_{ST} scans for adaptive alleles are more likely to succeed with higher rates of gene flow, The results support existing theory of local adaptation, and provide a more concrete look at a particular, empirically motivated example.

2 Introduction

intro outline: Start off with stickleback: repeated, rapid, parallel adaptation with obvious phenotypes. But, most species don't seem to have such obviously structured polymorphism: what's special about stickleback? The transporter hypothesis is a mechanism; what's required to make it work? Other species adapt rapidly in similar ways: e.g., Darwin's finches and island mice; do these differ through not having such a big, connected reservoir of genetic variation? or by not having such a large total population size of the "other" habitat? Or, is this situation more common than we know, with less obvious phenotypes? Also, can we actually tell what's happening from genetic data?

Recently, multiple instances of similar (parallel) underlying genetic basis of rapid local adaptation has brought about questions surrounding the origins and maintenance of genetic polymorphism. A empirical model is the Alaskan populations of freshwater and marine Ninespine Stickleback fish. 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, **?** observed significant phenotypic changes in less than 50 years that appear to be parallel to freshwater stickleback that have been separated for over 13,000 years. In 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. These results leave us with questions surrounding the nature of rapid adaptation. Does convergent evolution breed it's own solution (haplotype) for every new selective pressure, or can these solutions can be efficiently shared across multiple sub - populations facing similar selective pressures.

The leading hypothesis for stickleback is that rather than acting on new mutations, adaptation to freshwater environments is sped up through selection on standing genetic variation (SGV) found in marine populations. One clear example of this is the gene *eda* which has been shown to regulate the number of lateral plates. While this gene arose millions of years ago, it is found in freshwater ponds which have formed much more recently. Novel evidence from natural populations has provided evidence that most regions of the genome that distinguish marine-freshwater genetic differences share this pattern [**?**]. Schluter and Conte [2009] suggested the "transporter"-hypothesis. This outlined the flow of freshwater alleles into marine populations through offspring of hybridization events. It suggests freshwater haplotypes are distributed though marine individuals and are continually selected upon in introduced freshwater populations. The continued selection on freshwater favored alleles and introgression between the two sub-species, allows the marine to maintain the freshwater haplotype dispersed in low frequency among its' individuals. Once a new freshwater environment is introduced and inhabited by marine individuals who carry freshwater adapted alleles, selection reconstructs the freshwater haplotype [Schluter and Conte, 2009]. An alternative to this hypothesis is that individuals from other freshwater environments migrate directly to the new environments, and their haplotype is passed down directly. This hypothesis has been shown to be unlikely due to finding a high frequency of freshwater alleles in the ocean, and almost no freshwater individuals.

Bill: maybe you could make this section a little more specific

If selection on standing genetic variants is key in rapid adaptation, many questions are exposed concerning the surrounding population genetics parameters and underlying genomic architecture. What scale of historical introgression between populations with selective pressures X and Y, allows for rapid and parallel local adaptation of a population derived from X to an introduced environment with selective pressure Y? What is the origin of pre-existing adaptive variants? How are the variants structured, historically and across

Bring up the possibility of more than one freshwater haplotype here.

Re: should I? Doesn't Thom's work and parallel

the genome underlying the trait? Furthermore, how can we infer causal loci for regions of the genome that must be driving the rapid adaptation, from real data. Many biologists today make use of genome wide association studies (GWAS) and F_{st} across the genome to estimate regions responsible for certain traits. Unfortunately, the data can be heavily influenced by the level of introgression and population structure of the samples. This being the case, what are the biases we can expect to see in real data for certain levels of introgression.

There are many different parameters that carry a large impact on the questions above. **?** explored the dichotomy of selection on new alleles and those brought in by migrants from a similar selective pressure. Answering the fundamental questions surrounding spatial resolution of convergent evolution.

In this paper, we explore time to local adaptation as a function of migration rate through an environment (marine) with opposing selective pressure. Next, we compare to real data and explore how F_{st} data can be impacted by introgression.

It seems like this could be a good parallel to what we're doing here, or maybe for discussion?

3 Methods

We explored these questions using forwards-time simulations with explicit genomic representation of a quantitative trait in SLiM [**?**]. The details (described below) are motivated by previous population genomic studies of threespine stickleback, but remain simplistic in some aspects 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 population; we will consider the likely effects of this in the Discussion.

TODO

Habitat and geography Our model has two habitat types, which are considered separate populations with a fixed total size of 2,000 diploid individuals each. The marine population is treated as a single large deme, but the freshwater population is subdivided into local subpopulations or demes. The scheme, 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 10 units of length, and the "freshwater" habitat is divided into ten subpopulations (which we call "lakes"), each connected to the marine habitat at positions $i - 1/2$ for $1 \leq i \leq 10$.

Fitness in each habitat is determined by a single continuous trait with an inherent tradeoff such that alternative values of the trait have high or low fitness in tht two different habitats. This situation roughly models the cumulative effect of the various phenotypes such as armor morphology, body size, craniofacial variation and opercle shape (among others) on which divergent selection acts in the two environments. The optimal trait value in the marine habitat is $+10$, and in the freshwater habitat is -10. Fitness of a fish with trait value x in a habitat with optimal value x_{opt} is Gaussian with standard deviation, i.e.,

$$f(x; x_0) = \exp\left\{\frac{1}{2}\left(\frac{x - x_{\text{opt}}}{15}\right)^2\right\}.$$

Potential genomic architecture Each individual carries two copies of a linear chromosome of 10^5 loci (which could be thought of roughly as a kilobase each). New mutations occur with probability 10^{-7} per locus per generation. Ten "effect regions" of 100 loci each are spread evenly along the chromosome (separated

Initial genomic architecture? The header is a little confusing to me

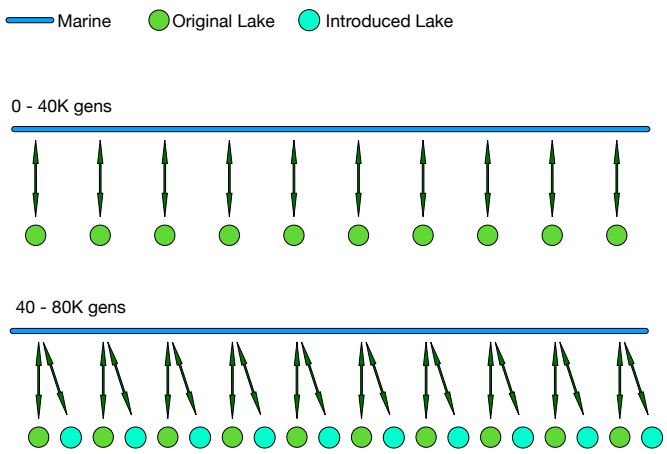

Figure 1: A representation of the geographic and evolutionary history of all populations throughout the simulation. The marine is a one-dimensional, continuous population with spatial positions ranging from [0.0 - 10.0]. Each freshwater lake$_i$ for both the introduced and freshwater populations is a discrete population connected by migration only through the marine, at position $i - 0.5$. The introduced lakes have the same spatial location (but separate?) and selective pressure as the original lakes, but arise at 40K generations. The marine has 2000 diploid individuals while each of the lakes fluctuate around 200. the introduced population is initialized as a copy of all marine individuals to model marine individuals inhabiting a newly created freshwater environment such as the ponds around Middleton Island. We then observe the selection process of marine individuals and following generations with the new selective pressure across a range of parameter values.

by 4950 loci) in which mutations affect the trait; mutations in other parts of the genome are neutral. Each mutation in these regions is either additive, completely recessive, or completely dominant (with equal probability), and with an effect sizes chosen randomly from an Exponential distribution with mean 1/2, either positive or negative with equal probability. Individual trait value are determined additively from the diploid genotypes. Concretely, an individual that is heterozygous and homozygous for mutations at loci H and D respectively has trait value $x = \sum_{\ell \in H} h_\ell s_\ell + \sum_{\ell \in D} s_\ell$, where s_ℓ is the effect size of the mutation at locus ℓ. Subsequent mutations at the same locus replace the previous allele.

Life cycle 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). The model is Wright-Fisher, so that each generation, in each habitat, 2000 new individuals are produced, and so to keep population sizes roughly constant within each lake, at the start of each generation we rescale fitness values in the freshwater habitat so that the sum of the fitness values within each lake are equal.

Migration 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. The between-habitat migration rate is denoted m, and will be called simply the "migration rate". 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 reweighted by a Gaussian function of the distance separating the two, with stanadard 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 $f(x_i)\exp(-2d_{ij}^2)$, 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 offpsring produced by marine individuals are assigned to the lake nearest to the position of the first marine parent.

New lakes To study adaptation in newly appearing freshwater habitats, we introduce a new set of lakes midway through the simulation. The initial set of individuals have parents chosen as above from the marine habitat, and act like an independent copy of the original set of lakes – in particular, the two sets of lakes each have 2,000 individuals at all times. Since this doubles the number of lake-to-marine immigrants, after this happens 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 *pre-existing freshwater adapted 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 lakes do not count if they are not also common in the original lakes. They are defined this

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

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 adaptaiton 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 $1 - p_f p_m/(\bar{p}(1-\bar{p}))$.

4 Results

We varied the migration rate, m, from 5×10^{-5} to 5×10^{-2} per individual, i.e., between 0.01 and 10 migrants per year to and from each of the ten lakes. This had a strong effect on many aspects of adaptation, including how fast adaptation occurred in each lake, how much alleles were shared between lakes, and the population genetic signals left behind. At the lowest migration rate, lakes adpated nearly independently, while at the highest migration rate, the habitats are beginning to act like a single metapopulation with substantial migration load. We will now dissect what happens between these extremes.

merge with the above Interestingly, the ability for separate populations to locally adapt to their own selective pressure was relatively unaffected until the highest rate of migration between the marine and freshwater environments. All rates of historical introgression aside from the lowest, helped both the initial and introduced freshwater populations share pre-existing freshwater adapted alleles. The sharing of alleles resulted in rapid adaptation ($\approx$ 100 generations) of the introduced population (split from the marine population) to adapt to the freshwater environment. We also found that larger amounts of migration allow for more statistical power and less false positives in the resulting population genetics data (F_{st} per SNP across the genome).

Local Adaptation: differentiation with gene flow

Local adaptation occurred in all simulations: as shown in Figure 2, freshwater and marine populations diverged until the trait means were close to the optimal values in each habitat. (Remaining simulations are shown in Figures **??** and **??**.) Greater migration decreased adaptedness only weakly except for at the highest migration rate (10 migrants per generation per lake), where the mean trait value equilibrated at roughly half its optimal value (Figures 7 and **??**). The establishment of new alleles in the lakes is visible in Figure 2 as jumps in the mean trait value; in the two simulations these occur on a time scale of 100 (lower migration) to 1000 (higher migration) generations, and move the trait by an amount of order 1. Trait variation within each population was small compared to the difference between populations, with interquartile ranges of around XXX.Across all parameter values, differences at around 16 commonly polymorphic sites (eight the shift the trait in each direction) were responsible for most of the adaptive differences between freshwater and marine habitats.

TODO

ALSO TODO: CHECK THIS

also, interesting point to make on second plot, you

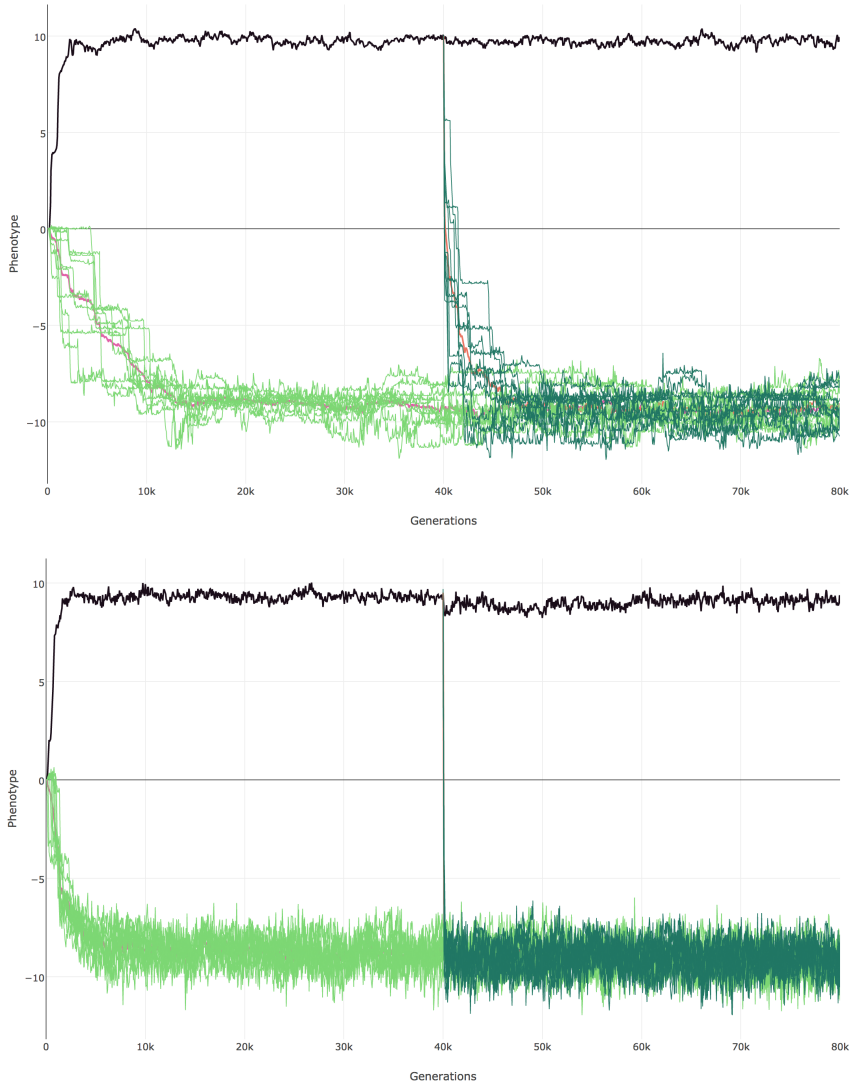

Figure 2: Mean individual trait values in the marine habitat (black line), the original lakes (light green lines; average in pink), and the introduced lakes (dark green lines; average in orange) across the course of two simulations, with migration rates of **(top)** $m = 5 \times 10^{-4}$ and **(bottom)** $m = 5 \times 10^{-3}$ per generation individual (i.e., 0.1 and 1.0 migrants per generation per lake, respectively). Optimal trait values in the two habitats are at ± 10.

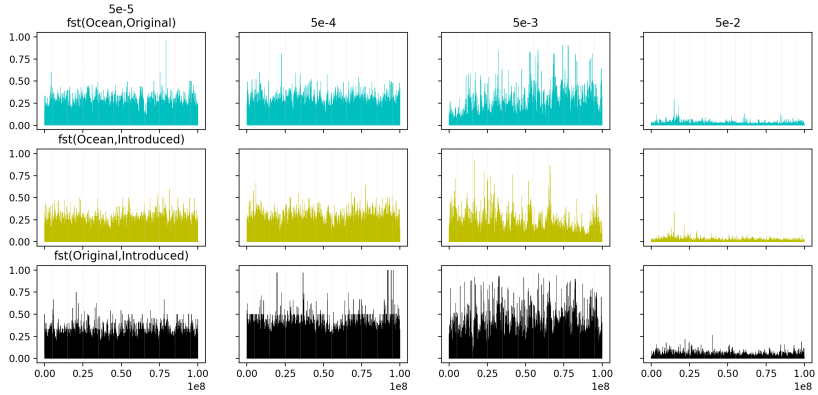

Figure 3: Distributions of F_{st} values between the three subpopulations throughout the simulations. On the left, F_{st} was calculated only on neutral mutation frequencies. On the right, F_{st} was calculated only on effect (Impact on phenotype) mutation frequencies. Effect mutations are acted upon by selection as they have an affect on fitness where neutral mutations

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. Local adaptation was still able to occur despite overall homogenization: if computed using only sites with alleles affecting the trait ("effect mutations"), F_{ST} between habitats was relatively constant across migration values.

Speed of adaptation Increasing migration rate strongly decreased the time until freshwater habitats could adapt, both in the initial and introduced sets of lakes. As shown in Figure 5, at the lowest migration rate it took over 20 thousand generations for the mean trait value across introduced lakes to get to within 0.5 of the original lakes value. Although many lakes had adapted before this time, the different rates of introduction of effect alleles is clearly seen in the traces of trait values against time (e.g., Figure 2). However, higher migration rates allowed the freshwater habitat to adapt nearly as quickly as the marine habitat.

Sharing of freshwater adapted alleles

The tenfold difference in speed of adaptation occurs because at low migration rates, adaptation occurs independently in each lake, and the marine habitat has ten times the influx of new alleles than any one lake. In other words, greater mixing at higher migration rates allows 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 further investigate sharing of locally adaptive alleles and the transporter hypothesis, we investigate the "pre-existing freshwater adapted alleles", defined for a particular generation to be effect mutations above 50% in at least one original lake and below 50% in the marine habitat in that generation. Figure 6A shows the distribution of the number of these alleles, across generations, as well as the average number of lakes each is found in. As migration rates increase, the number of these alleles decreases steadily, and each is concurrently found in a greater number of lakes.

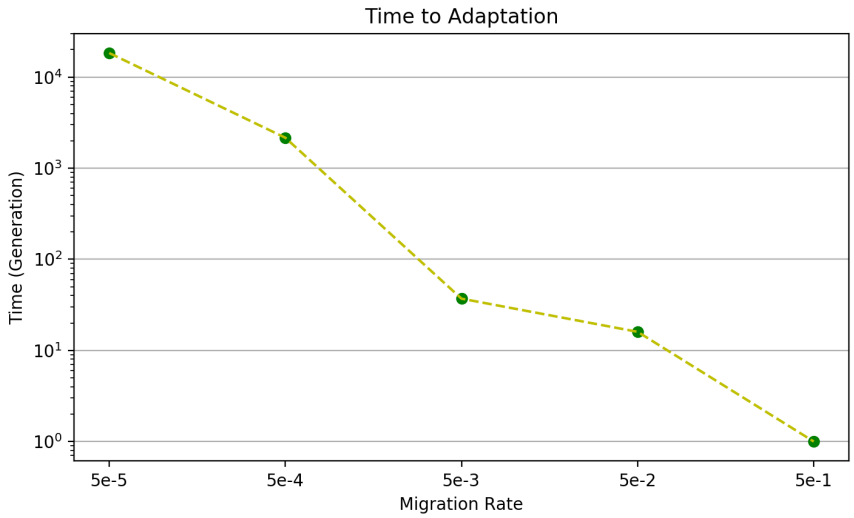

Figure 4: Time to adaptation as a function of migration rate (M) parameter value. This is where we measure how many generation it takes for the introduced population's mean phenotype to come within 0.5 of the original lakes average phenotype. Each point represents a simulation run at some value of M. The yellow dashed line is the average of all points at each respective parameter value

As the number of common, locally adaptive alleles decreases, the genetic basis of adaptation is more commonly shared. Figure 6B shows the distribution of the mean percentage of currently-defined freshwater adapted alleles that each genome in each of the populations carries, averaged across time and individuals. If all individuals across lakes carried the same set of alleles determining their trait value, this would be 100%. This value is nearly reached at the highest migration rate; but it is lower due to migration load. At the lowest migration rate, each genome in the original lakes have almost exactly $1/10^{th}$ of the total freshwater adapted alleles – since there are 10 lakes, this implies that each lake has adapted with a unique set of alleles. Since these are *pre-existing* alleles, the value is zero for introduced lakes.

Migration load and genetic variation

At first, increased migration allows sharing of adaptive alleles between lakes, but at the highest migration rate, the constant influx of alleles between the habitats creates substantial migration load. This rate, $m = .05$, 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 7. Significant gene flow constricts local adaptation as a consequence of a large number of offspring through hybridization events between subpopulations.

Standing genetic variation Despite a dramatic difference in the amount of allelic sharing between lakes, standing genetic variance (Figure **??**) was around 0.05 in the freshwater habitat across all but the highest migration rates. Concurrently, genetic variation in the marine population steady increased with migration to a similar value. On the one hand, it is not surprising that lakes, as a group, carry more genetic variation

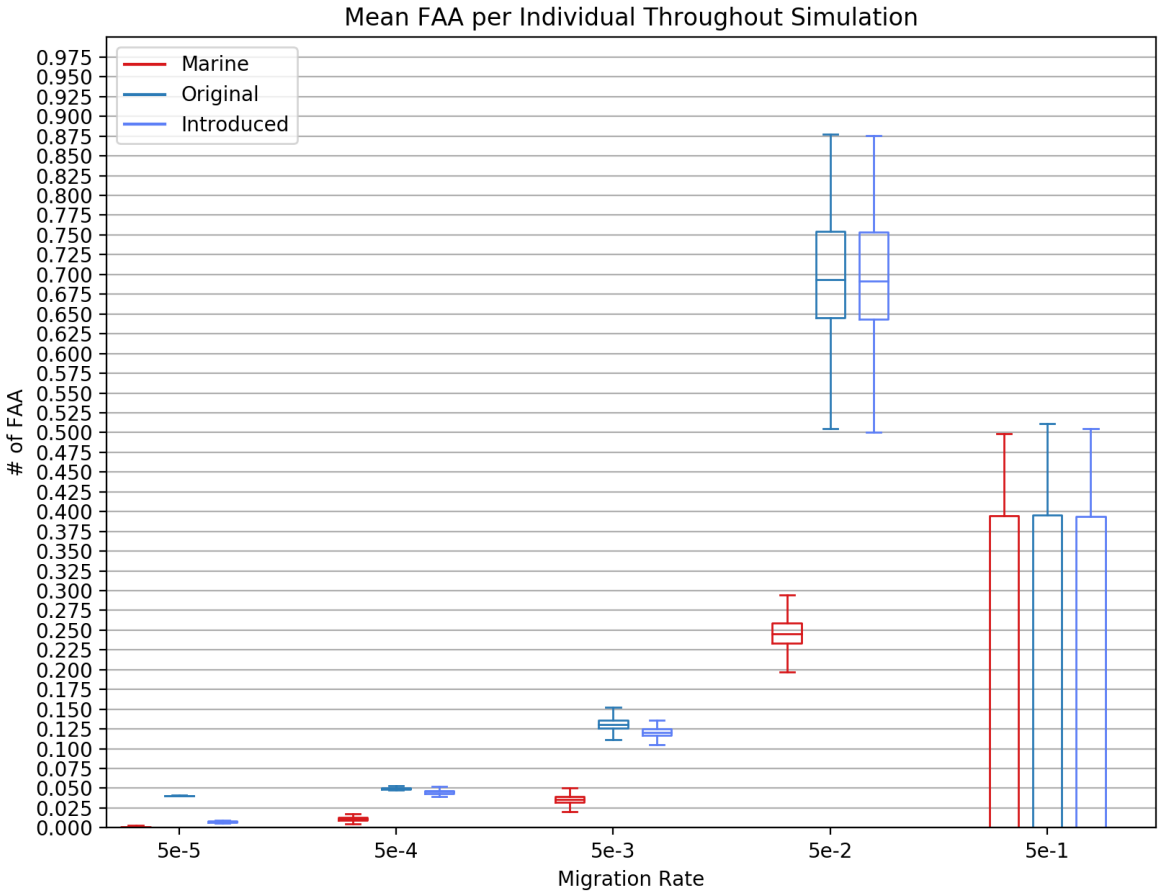

Figure 5: Distributions of mean percentage of freshwater adapted alleles (FAA) per individual throughout the simulation run, for each subpopulation. We count the total number of freshwater alleles for each individual before averaging them in each population and dividing by the total number of defined freshwater adapted alleles. Looking at total number of FAA per individual gives us an idea behind how many alleles underly a freshwater haplotype, while the percentage tells us the variance of the haplotype

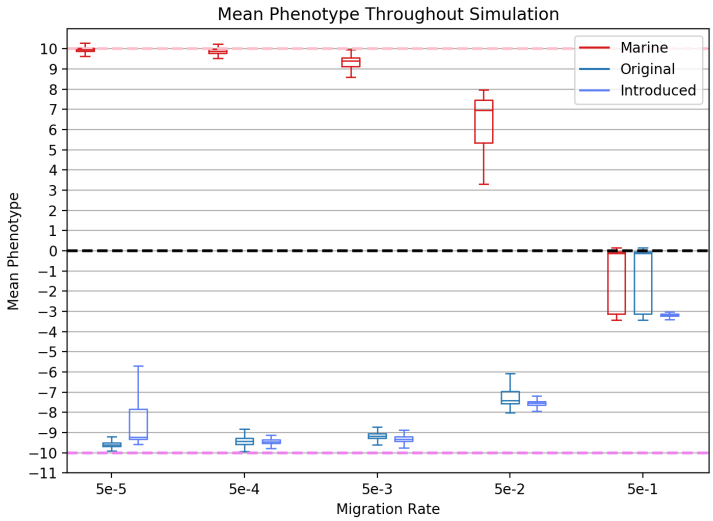

Figure 6: Distribution of mean phenotype throughout simulation runs at separate migration rate (M) parameter values, for each population. The dashed pink line (Pheno = +10) is the optimum phenotype any individual in the marine environment. In contrast the purple line at (Pheno = -10) represents the optimum for any individual in the freshwater environment. All individuals at generation 0 (beginning of the simulation)

than the marine habitat, since population subdivision allows different alleles to become common in each. However, it is striking that at $m = .005$, the marine habitat carries as much genetic variation, despite a lack of any substantial migration load.

Realized genomic architecture

Now we take a closer look at the genomic architecture of local adaptation between the two habitats. Do the alleles underlying trait differences cluster along the genome? Do measures of local differentiation identify the causal loci? Figure **??** shows plots along the genome of per-locus F_{ST} values between the marine and original freshwater habitats at the two intermediate migration rates. (Note that we are pooling freshwater habitats; a single lake would provide substantially less power.)

need to know where the FAA are in these figures

Higher migration rates showed more distinct F_{ST} peaks over polymorphic loci underlying trait differences between the habitats. As migration rate decreased, the "background" levels of F_{ST} increased, swamping out this signal until the regions under selection were indistinguishable. This is likely due to two reasons: first, stronger genetic drift with less migration leading to higher background F_{ST}, and second, greater sharing of ada~~ptive alleles providing a shared signal across populations.~~

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 8 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.

Labels are wrong on that figure: false pos rate should go down with cutoff. Is it

Figure 7: 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: x-axis label (F_{ST}).

With 10 effect regions across the genome, we can see in **??** that only about 1 or 2 of the region contain peaks. When looking closely at the peaks we see that each is usually a composition of 7-8 SNPs close in proximity. Because some of those mutations are selectively neutral (we can see this because the counts of FAA are not as high as the number in the clusters), this is suggestive of hitchhiking alleles along with the SNPs being brought to high frequency by a selective sweeps.

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 8 , 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.

Theoretical expectations

does this go first or second?

Here's some rough calculations to get a sense for what should be going on. Everything is done in more detail in OTHER PLACES WE SHOULD CITE. Fisher, Chevin, etc.

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 $2s$ [**??**]. 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 $s(u) = \exp(-\beta((z -$

$u)^2)/\exp(-\beta z^2)) \approx 2\beta z u$, 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 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 $\int_0^\infty 4\beta z u \exp(-2u) du = \beta z$, 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 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.

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 (which is close to optimal). 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 per generation divided by the selective disadvantage, which if $M = 2000m$ is the number of immigrants per generation, is $2M/\beta z^2$. With $z = 1$, the factor multiplying M is $2/\beta z^2 \approx 1/200$: since lakes have 400 genomes, as long as $M \geq 1$, 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 we'd need $M \geq 10$ to ensure enough pre-existing genetic variation that adaptation would happen entirely using the initial set of colonizers. This corresponds to our highest two migration rates, as in e.g., Figure 2B.

Migration If sufficient genetic variation is not present in a new lake initially, it must appear by new mutation or 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) adaptation is possible using variants present at colonization or arriving shortly thereafter, or (b) adaptation takes many multiples of $1/m$ generations.

These calculations depend on there being no bottleneck in colonization of the lake; if there is a bottleneck, then an additional factor must be added.

At what point do we expect new mutation to be more important than migration for adaptation? By the calculations above, if $M \geq 10$, we expect initial diversity in a lake to be sufficient for adaptation, corresponding to our third-highest migration rate. If this does not happen, then we expect adaptation to

take a multiple of $1/m$ generations; with our values, $1/m$ ranges from 20,000 to 20 generations. Above we estimated that adaptive alleles due to new mutation fix locally about every 2,000 generations, suggesting that at our second-lowest migration rate (where $1/m = 2,000$), the two contributions of migration and new mutation are roughly equal. This is in fact what we see: in Figure 6, we see that at the second migration rate, alleles start to be shared between lakes, while by the third migration rate, they are almost entirely shared.

5 Discussion

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 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.

We also have also explored the genomic architecture as a consequence of the nature of selection in our scenario. The alleles underlying individual trait value were of large effect and a low number. With a total of 10^5 loci, to be realistic, each loci should represent 1000 Kb in real data.

We 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.

This should be expanded upon, not sure what this means in terms of the genomic architecture

thresholds

We have found that too little migration leads to selection upon new mutations in all subpopulations and lakes alike. In contrast, at high migration rates we have seen that migration load limits the ability for species to locally adapt to the selective pressure of their environment. This leads us to consider a window of introgression which allows for the transportation of FAA's without migration load.

connect results back to real data?

The adaptive filter? RAMBLING THOUGHTS HERE Since larger effect alleles are more likely to establish, be it by mutation or migration, repeated colonization of new freshwater habitats will select for larger alleles, be it single alleles or haplotypes bound together by an inversion. However, these are more strongly selected against in the interstitial time. Being recessive would help with this, but would also make it more difficult to establish.

References

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`.

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`.

Peter L. Ralph and Graham Coop. Convergent evolution during local adaptation to patchy landscapes. *PLOS Genetics*, 11(11):1–31, 11 2015. doi: 10.1371/journal.pgen.1005630. URL `https://doi.org/10.1371/journal.pgen.1005630`.

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`.

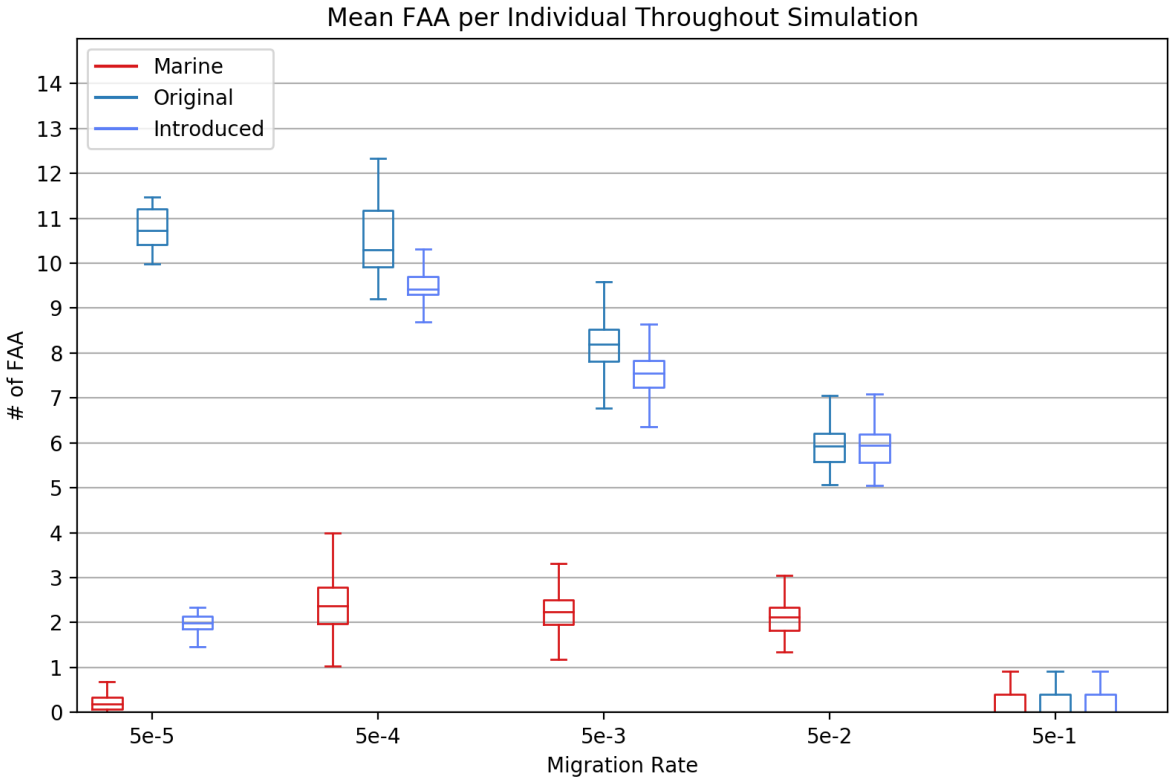

Figure 8: Distributions of mean number of freshwater adapted alleles (FAA) per individual throughout the simulation run, for each subpopulation. We count the total number of freshwater alleles for each individual before averaging them in each population and dividing by the total number of defined freshwater adapted alleles. Looking at total number of FAA per individual gives us an idea behind how many alleles underly a freshwater haplotype,

A Supp.