# 1 Abstract

Note: Finish after the whole paper has been written

# 2 Introduction

In the mid 19th century, Darwin theorized that the process of natural selection resulted in speciation, meaning all living species derived from a common ancestor. He also believed that this process took place on ecological timescales that would not allow for this process to be observed in a single lifetime. Today, we have found evidence of evolution in tens of generations and are starting to grasp the complexity behind such a simple idea. Famously, Drs. Peter and Rosemary Grant went on to study Darwin's finches and found that phenotypes could adapt to changes in the food source within mere generations. We have have found many examples of rapid adaptation in many commonly studied models but many questions still remain surrounding the genomics which allow for this.

Recently, rapid and parallel adaptation has brought about question of selection of standing genetic variation in natural populations of Stickleback. In 1964, The Great Alaskan Earthquake caused Middleton island to raise up and in turn, introduced a group of new freshwater ponds around the perimeter of the island. Quickly inhabited by the surrounding marine population of Ninespine Stickleback, we've 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. [2] 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 (Cite Kristin, Cresko et. al)

The leading hypothesis is that rather than acting on new mutations, ecological speciation is sped up through selection on standing genetic variation 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. [3]

In 2009, Dolph Schluter and Gina L. Conte suggested the "transportation"-

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 rebuilds the freshwater haplotype. [4] The 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.

Assuming they were correct about transportation, we might ask; how much migration and gene flow is to be expected for this relationship to be maintained? (Other questions) how does selection, at the genomic level, alter given a range of migration and recombination rates? Here, we use forward moving simulations to model marine and freshwater populations of stickleback. We then observe the effect (focusing on migration and recombination, for now) of different parameters on local adaptation of introduced freshwater environments. We compare to real data and make predictions about realistic parameters.

Simulations are a powerful tool when there is some map between the parameters we want to learn about and the genetic data we observe in nature. We don't know what that map is, so we invert the relationship. This means going from genomic data that we see to the parameter values that must be driving the system, assuming the model is correct.

# 3 Methods

Our models are forward-moving evolutionary models which emulate the geography, selective pressure, and genomic architecture of coastal marine and freshwater lake populations. Here we give a brief overview of the software used and the models created to observe the effects of selection.

Note: Maybe talk about some things are stochastic: (1) Correlation of mutation frequencies at low migration rates (2) Distribution of Effect Size? (run

Each set of simulations was run multiple times to ensure the data we observed was not stochas-

another set to make sure)

## 3.1 SLiM

To model the stickleback populations, we used a flexible evolutionary framework, SLiM. (cite SLiM 2) The model-type we used was based off of an extended wright-fisher. The simulation life cycle of each generation is as follows. First, generate offspring by: (1) choosing parents based on cached fitness values and migration rates, (2) performing recombination of parental genomes, (3) allow for mutation, (4) modify each child by a defined callback. Next, offspring become individuals before fitness is evaluated for the next generation. Finally, the generation counter is incremented. The population structure in SLiM can be arranged in any number of subpopulations, continuous or discrete, connected by any rate of migration.

## 3.2 Model

### 3.2.1 Geography

What we've modeled geographically is a coastal marine population connected by migration to multiple, smaller freshwater populations that are located along the coast. The population structure for the marine is modeled as a one-dimensional, continuous population ranging from 0.0 to 10.0. There are two sets of lakes which represent freshwater populations; the first which evolves in parallel to the marine, and a set introduced later in the simulation as a subset of the marine. In SLiM, these freshwater sets are defined as just two continuous subpopulations. Using SLiM?s $modifyChild()$ callback, we?ve modeled both to act as 10 separate discrete subpopulations. Each of the 10 lakes, $i$, is located along the marine at $i - 0.5$, and connected by migration only through the marine environment. The marine has 2000 individuals while each of the lakes fluctuate around 200.

Should I put my geography picture here?

### 3.2.2 Selective Pressure

To emulate the freshwater and marine selective pressure, we set up a quantitative genetics model in which fitness is phenotypically based. Freshwater and Marine environments are distinguished by a single numeric value. This value is the

optimum phenotype for each environment and acts as a representation of lateral plate number and opercle shape in the stickleback populations. The fitness of an individual is then determined by a probability density for a normal distribution at the quantile (?) of the difference between the optimum and the individual's phenotype

$$\text{Individual.Fitness} = X \sim \mathcal{N}(\text{Optimum - Individual.Phenotype, } 0.0\,15.0).$$

Where an individual?s phenotype is also defined by a single numeric value; determined as a summation of the selection coefficients of all mutations an individual possesses.

$$\text{Individual.Phenotype} = \sum_{m \in mutations} \text{m.selectionCoeff}$$

We allow six mutation types which affect phenotype along with one neutral mutation type. Among the effect mutations; there are additive, dominant, and recessive mutations which effect phenotype in either the positive or negative direction. The selection coefficients are pulled from a beta distribution with a shape parameter of 1 and a mean of either 0.5, or $-0.5$.

This isn't quite right because of how we calculate additive, recessive, and dominant

$$\text{Mutation.selectionCoeff} = X \sim \Gamma(1, \pm 0.5)$$

### 3.2.3 Genomic Architecture

Although still uncertain about how much of the genome is directly associated with the distinguished phenotypes, GWAS has indicated clusters of loci (linkage groups? Operon?) along the genome to be causative. To mimic this and compare dynamics of neutral vs. effect under selection, we create regions of effect. These genomic elements are the only loci which allow for mutation that impact phenotype. This allows us to observe differences in selection at the genomic level. In SLiM, the genome is conceptually a linear array of loci for which we can define different amount

finish this

### 3.3 Sampling

Here, we're interested in observing how different parameter values (mainly migrations rates and recombination rates) impact the sharing of alleles between populations. we define freshwater adapted alleles (FAA), at any given generation, to be a mutations with a frequency higher than 0.5 in *any* of the original lakes, while remaining lower than 0.5 in the marine. This is because the transportation hypothesis does not specify where or when an advantageous mutation arises; but simply suggests that when one comes to high enough frequency, it too, could participate in the transportation process. [4] To observe transportation of freshwater alleles in our simulations, we use a variety of metrics when sampling. Throughout the simulation, we measure; mutation frequencies, $F_{st}$ values, average phenotype, mean number of original lakes each FAA appears in, and total number of unique FAA. to understand where, as well as how much, of these freshwater haplotypes are being transported, we track the mean percentage of FAA per individual (MPFAI) across all populations throughout the simulation: computed as the sum of the FAA mutation frequencies among each population. we're particularly interested in using MPFAI to measure the marine population' capacity to carry FAA, as well as observing the introduced populations' ability to select upon them.

Because we are interested in the local adaptation of the introduced freshwater populations, we define time to adaptation ($T_{adapt}$) of the introduced population to be the generation at which the difference between the average phenotype of the original and the introduced freshwater populations is less than 0.5. At this generation, we look at: the correlation (B. Pearson) between the mutation frequencies of all effect loci and FAA, and the number of shared high frequency alleles. These metrics tells us how much of the FAA were used in the lakes once they have locally adapted.

## 4 Results

### 4.1 Local Adaptation: differentiation with gene flow

Local adaptation occurred across all simulated parameter values. Starting from the same baseline freshwater and marine populations diverged phenotypically, until they reached an equilibrium, where the population means were close to or at the optimal phenotype. phenotypic variation within each population was

small compared to the difference between populations. Polymorphic loci that effected the phenotype, showed greater differentiation compared to neutral loci between marine and freshwater populations, with the exception of migration rate at a value of $5e-5$. $F_{st}$ plots along the genome, show peaks at loci underlying local adaptation.

add more general observations and caveats about not all parameter values.

## 4.2 The Effect of Migration Rate

### 4.2.1 Migration is Mixing

Across increasing parameters of migration, we observed more gene flow across the populations. As can be seen in Figure (neutral fst figure) $F_{st}$ values for neutral Alleles steadily decline between all sets of populations as migration increases, this illustrates less difference between sections of the genome which are not acted up upon by selection, in all populations. Additionally, Standing Genetic Variance (Seen in SGV Figure) values in the marine population steady increased with migration which suggests that more alleles from the lakes surrounding were exporting alleles into the marine individuals.

computed as?

Something about SGV rates for the the other population

### 4.2.2 migration affects the speed of adaptation

We observed a dramatic shift in time until adaptation for the introduced populations as migration rates increased. As can be seen in Figure (Time Until Adaptation) at the lowest migration rate of $5 * 10^{-5}$, It took the introduced population over 30 thousand generations for the average phenotype of all the lakes to get to within 0.5 of the original lakes original phenotypes. This suggests that selection was acting primarily on new mutations, meaning the introduced lakes needed to wait for a beneficial mutation to arise before it was selected upon. There was a positive correlation between the amount of standing genetic variation in the marine and the rapid adaptation of the introduced populations.

### 4.2.3 sharing of freshwater adapted alleles

To further investigate this correlation, we take a look at the FAA driving local adaptation of the introduced population. One common metric we investigate is the percentage of FAA per individual in each of the populations. This gives us a general look at the where these alleles are being distributed. Looking at

the lowest migration rate for the original lakes population In Figure (MPAA / IND) We see that the average individual has almost exactly $1/10^{th}$ of the total defined FWAA throughout the simulation. Because FAA are defined among *any* population, this suggests that each one of the 10 lakes has created their own solution to adaptation of the the freshwater selective pressure.

As we can see in Figure (Total/Avg shared) as migration rate parameter increases for the simulations, we see the distribution of the total number of FAA's decrease, and the average number of lakes each allele appears in at high frequency ($p > 0.5$), increase. With the distribution of effect size remaining the same across all simulations, these are both suggestive of the original lakes sharing solutions to the same selective pressure.

#### 4.2.4 We see Migration Load after a certain threshold of migration

In the distributions we have shown across migration rate parameter values, We have experienced the most dramatic shifts of the population dynamics at $M = 5x10^{-3}$. After a threshold between this and $M = 5x10^{-2}$, we start to experience migration load. Significant gene flow constricts local adaptation as a consequence of a large number of offspring through hybridization events between subpopulations. In Figure (Phenotype distribution) at $M = 5x10^{-2}$, we can see the distribution of average phenotype throughout the simulation pull towards the opposing selective pressure value and away from the local optimum in all subpopulations.

#### 4.2.5 There is a Qualitative Threshold of migration rates

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 (Goldilocks Zone) of introgression which allows for the transportation of FAA's without migration load.

#### 4.2.6 Low Recombination Causes Clustering

## 5 Discussion

These simulation we're

# References


[1] Daniel I. Bolnick and Patrik Nosil. Natural selection in populations subject to a migration load. *Evolution*, 61(9):2229–2243, 2007.

[2] 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.

[3] Thomas C. Nelson and William A. Cresko. Ancient genomic variation underlies recent and repeated ecological adaptation. *bioRxiv*, 2017.

[4] Dolph Schluter and Gina L. Conte. Genetics and ecological speciation. *Proceedings of the National Academy of Sciences*, 106(Supplement 1):9955–9962, 2009.