The effect of local population structure on genetic variation at selected loci in the *A. majus* hybrid zone

by

Mara Jean Julseth

April 2023

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Master of Science

Committee in charge:
Prof. Nicholas H. Barton
Prof. Beatriz Vicoso

ISTA Master's Thesis, ISSN: 2791-4585

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Mara Jean Julseth

April, 2023

Signed page is on file

The effect of local population structure on genetic variation at selected loci in the *A. majus* hybrid zone

Introduction

The evolutionary processes that brought about today's plethora of living species and the many billions more ancient ones all underlie biology. Evolutionary pathways are neither directed nor deterministic, but rather an interplay between selection, migration, mutation, genetic drift and other environmental factors. How selection counteracts the random processes of mutation, genetic drift and migration is key to understanding how differences within and between organisms evolve and are maintained. In reality, the life history of an organism's dispersal and the distribution of individuals can vary greatly, playing a central role in the latter two processes. Organisms exist in a heterogeneous distribution: their dispersal is limited by environmental barriers, and mating almost never happens in panmixis, but rather with individuals close by, thereby generating isolation-by-distance. The interplay between these processes can lead to differences in local population structure and create local barriers to gene flow. Understanding the influence of local population structure is therefore a key question in population genetics. Inferring these influences from either experiments or the field needs tremendous time and many generations of organisms and even scientists. However, since evolution is ongoing we can find so-called hybrid zones all over the world, where genetically distinct species meet and produce viable hybrids. These usually narrow but often long stretches are natural crossing experiments and give us a great opportunity to infer evolutionary processes using population genetic tools. (Westram et al 2021, Barton & Gale 1993, Barton and Hewitt 1985). In this thesis, I will analyse clines across an *Antirrhinum majus* hybrid zone to look at the effect of different demographic factors on selected loci and the creation of genetic barriers.

Hybrid Zones

Within hybrid zones we find clines - geographic gradients - for different loci, karyotypes or quantitative traits (Barton and Hewitt 1985, Bruno et al. 1994, Bridle et al. 2001). Clines can be dispersal independent with solely selection maintaining the cline (Moore 1977), or can be maintained by a balance between dispersal and selection. In the second case, different forms of selection can maintain the cline, including differing selection pressures on differing environments, selection against hybrids and frequency dependent selection as well as pure epistasis (Bazykin 1968, Mallet 1986, Mallet and Barton 1989, Barton and Gale 1993).

Different forms of selection can act simultaneously and yet have little to no effect on the shape of the cline (Barton and Gale1993) For example *Heliconius* butterfly hybrid zones are maintained by frequency dependent selection, selection against heterozygotes and epistasis. The difference in effect of these different types of selection pressure on hybrid zone formation and on parameters such as width is insignificant under weak selection (Barton and Mallet 1989, Barton and Gale 1993). The dynamic behavior of hybrid zones elaborately described by Barton (1979) indicate that hybrid zones maintained by a balance between selection and dispersal can move due to variation in density, selection pressures or dispersal rate. However, it is expected that most of them remain stuck where they started due to either natural barriers or local variation in population structure (Barton 1979).

The Theory behind Clines

Haldane (1948) developed the first theoretical model, using a cline in deer mice to infer the strength of selection. In Haldane's model, selection depends on geographic location with each side of the hybrid zone favoring a different genotype, leading to a gradual change of genotype frequencies between the two habitats at a single selected locus (Haldane 1948). Bazykin (1968) extended this model to an environment-independent cline, maintained by dispersal and disruptive selection, accounting for the possibility of varying allele frequency in different geographic locations by incorporating the effect of migration (diffusion) in addition to selection against heterozygotes. This leads to a stable 1D solution, where allele frequency is only dependent on one of the two spatial coordinates:

$$q(x) = \frac{1}{1+e^{2\sqrt{s}(x-\mathrm{X0})/\sigma}} \quad \textbf{(1)}$$

where (x) = allele frequency at given x,

s = selection pressure against heterozygotes

σ = standard deviation of the distance between parent and offspring

X_0 = arbitrary constant depending on the zero point of the coordinates selected

(Bazykin 1968).

This basic cline is a sigmoid curve. A variation on this can be a stepped cline, which occurs due to barriers to gene flow and is described in more detail below.

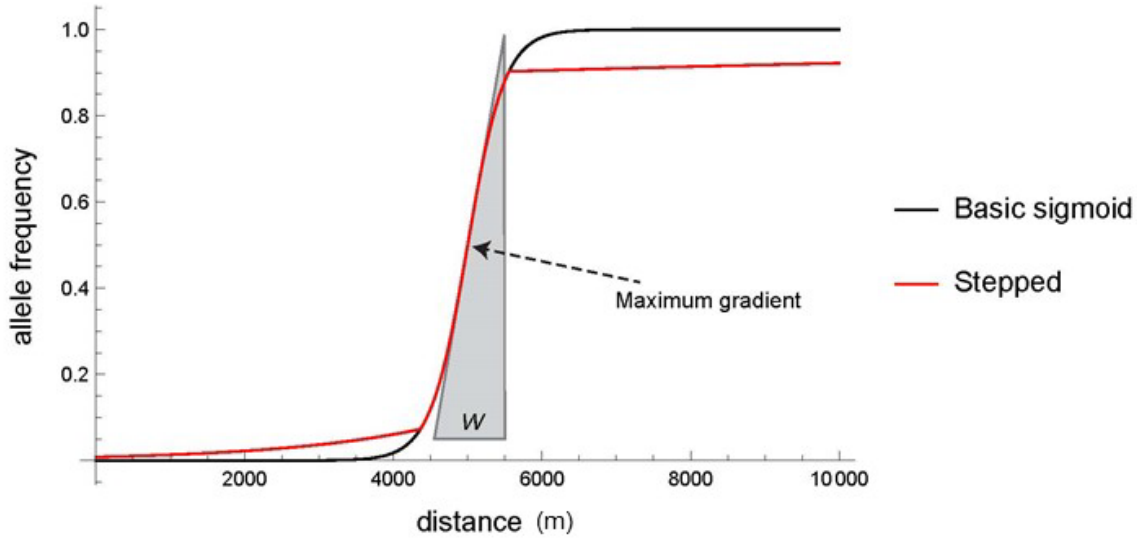

Figure1: Example clines, one classic sigmoid, one stepped cline with a width of 1000 m and center at 5000 m.

Cline width and shape

Cline width is proportional to the ratio between dispersal and the square root of selection ($\sigma/\sqrt{s}$), and is defined as the inverse of the maximum gradient (Barton and Gale 1993). Therefore, selection strength can be estimated when dispersal rates are known (Haldane 1948, Barton and Hewitt 1985, Mallet 1986).

The cline shape and position produced by different forms of selection on a single locus varies only slightly. The greatest deviation in shape is caused by asymmetric selection, favoring alleles differentially strongly on the side of a sharp ecotone, or by complete dominance in both selection against hybrids and frequency-dependent selection. While the first creates a longer tail where the recessive or less heavily selected alleles are common, the second leads to ineffectiveness of selection on the recessive allele, and therefore higher frequency than expected. (Haldane 1948, Mallet 1986, Barton and Gale 1993). In most hybrid zones, we find multilocus clines, with different loci contributing to diverging traits. When looking at multiple loci in hybrid zones one often finds strong associations between unlinked genes referred to as linkage disequilibria (LD), which are gametic correlations (Barton and Gale 1993).

Clines found in many hybrid zones are often wider than estimated by suggested dispersal rates and selection strength. This is most probably an effect of underestimated long-range dispersal, which in addition to widening the clines also leads to higher than expected rates of LD at both ends of the cline (Barton and Hewitt 1985).

Stepped clines and LD

A shape deviation from the theoretically predicted curve is found in a few different hybrid zones as for example in *Bombina*, with allele frequency changes in a sharp step flanked by shallow tails of introgression (Szymura and Barton 1986). Stepped clines can be caused by LD, by physical barriers to gene flow or by long-range migration, which can be hard to distinguish (Barton and Gale 1993). Stepped clines no matter whether produced by an environmental barrier to gene flow, or by LD with loci under selection, indicate a barrier to gene exchange at linked loci. Long-range dispersal in contrast increases gene flow and would counteract barrier effects.

Analysis and comparison of cline parameters and shape against theoretical predictions are commonly used to infer not only strength but also mode of selection and rate of gene flow (e.g Haldane 1948, Barton and Hewitt 1985, Mallet 1986, Szymura and Barton 1991). With some empirical studies relating displacement of clines to variation in density, Nagylaki (1975) first considered the effects of local barriers to gene flow. However, there has been relatively little subsequent work; further theoretical work incorporating local population structure on cline parameters in tension zones has largely been missing (Nagylaki 1975, Wang et al. 2011, Westram et al. 2018).

Varying Density in Clines

Areas of low density correspond to a natural barrier in the hybrid zone, as for example a river, which would make dispersal across this barrier hard to impossible. When encountering such patches of low density, theory predicts a step in allele frequency at the barrier location in the cline. This step is proportional to change in slope. Therefore, the barrier strength can be calculated from the width *x* of the low-density region, and the factor, *k*, by which density is reduced.

$$Bs = x(k^2 - 1) \quad \textbf{(2)}$$

with *k* being the factor of density reduction and *x* the distance of the density reduction zone (Nagylaki 1975, Barton 1986). The barrier being:

$$B = \frac{\Delta p}{p'} \quad \textbf{(3)}$$

with Δp being the difference in allele frequency across the step and p' the gradient of allele frequency dp/dx (Szymura and Barton 1986).

Population Structure

Models of population structure have a long history, dating back to Wright's Island model (Wright 1931). The downside of most existing models of population structure is that they are limited to uniform distributions at equilibrium with homogenous migration, or to discrete populations. A recent model for neutral genetic variation considering population structure, while simulating a continuous population was created by Surendranadh et al (2022) inspired by an *Antirrhinum majus* hybrid zone. This model, which considered neutral loci, showed better results in estimates of variance in inbreeding and pattern of isolation by distance than a uniformly distributed simulation (Surendranadh et al 2021). My goal is to look at the role of demography on variation of selected loci in a similar manner, using data on clinal loci from the same *Antirrhinum majus* hybrid zone.

The *Antirrhinum majus* Hybrid Zone

Antirrhinum majus forms hybrid zones in the Spanish Pyrenees, where two subspecies meet: *A. m. pseudomajus* with magenta colored flowers and *A. m. striatum* with yellow colored flowers, with intermediate colored hybrids. *Antirrhinum* is hermaphroditic, self-incompatible, diploid and carries 8 chromosomes, with a genome size of roughly 500Mb. Flower color is mainly regulated by the tightly linked loci ROSEA and ELUTA-SULFUREA and FLAVIA, with the first two being responsible for the intensity and pattern of magenta pigmentations and the latter two affecting the distribution of yellow pigmentation (Whibley 2006, Tavares 2018). Two further loci, RUBIA and UDP, also play a role in flower color, as inferred from field data and crossing experiments (Barton Group, unpublished). The 5 unlinked loci are hypothesized to be under positive frequency dependent selection due to pollinator preference, assuming that rare colors have less probability of being perceived by pollinators in the area. *Antirrhinum majus* is a short-lived, self-incompatible, hermaphroditic perennial herb that has seed banks with most parents recorded 3-4 years prior (Field, unpublished data). Through extensive field work, the Barton lab has collected positional and genetic data of more than 20000 plants over the last 14 years (2009-2022) in Vall de Ribes, Spain. The area considered for this work stretches along a transect roughly 16km long. Little genome-wide neutral divergence is seen for

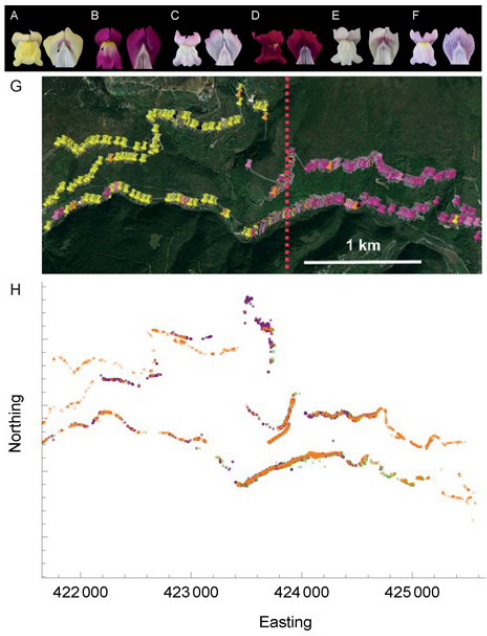

Figure 2 : A: *A. majus striatum*, B: *A. majus pseudomajus*, C-F: Different A. majus hybrids, G:Locaitons of recorded flowers in the field GPS-View 2009; H: Locations of recorded flowers 2009-2011,color coded by year

plants within this study area (F_{st} ~ 0.02), with only limited regions, associated with flower color, showing higher $F_{st\ (ST)}$ and steep clines (Whibley 2006, Tavares 2018). The distribution of plants in this study area is not uniform but heterogeneous and patchy. This makes this dataset ideal to look at the effect of patchy distribution and long-range dispersal on cline parameters and shape.

Aims

I aim to understand how different parameters of population structure - specifically, long-range dispersal and heterogeneous density - affect genetic variation. Cline analysis is a good tool for this: by comparing cline shape and parameters against theoretical predictions we can shed light on how genetic barriers and selection strength shape genetic variation. Existing theoretical models on the effect of population structure are often confined to discrete or uniformly distributed populations. Surendranadh et al (2022) developed one of the first continuous models, in order to look at the effect of a heterogeneous population structure on the distribution of heterozygosity and relatedness. However, the effect of population structure on selected loci has remained relatively unexplored so far in theoretical models. I therefore extended Surendranadh et al's (2022) simulation to incorporate the effect of selection at a single locus in a continuous and heterogeneous population structure, adding: a single locus under selection and positive frequency dependent selection acting on fathers to their initial basic assumptions. Leading to my model assumptions being:

- a continuous population
- a single locus under selection
- positive frequency-dependent selection on fathers
- discrete and non-overlapping generations

I simplify the actual reality of the *A. majus* hybrid zone by ignoring seed-dormancy and overlapping generations as well as linkage between multiple loci under selection.

The simulation works in two main steps: (1) generating the position for each individual and (2) finding the corresponding parents for each individual and assigning their genotype. All steps are described in their different simulation regimes in more detail below.

By choosing parents independently conditional on their location in this model, the simulation allows for any spatial structure and can be used to analyze the interaction between different demographic processes. In this thesis, I aim to untangle individual processes leading to population structure and their effect on cline properties by addressing the following aims:

Aim 1: Validation of my model

As a first step, it is necessary to validate my model by comparing it to theoretical predictions, comparing with Eq 1 above. For this, I simulate different population sizes (N) and different selection coefficients (s) with a uniform distribution of plants and a Gaussian dispersal kernel. I then compare the predicted widths for different selection coefficients with the realized widths from the simulations. I furthermore investigate how density affects cline parameters. This is the first step to find the necessary parameter range in which the simulation approaches the theoretical assumptions.

Aim 2: Analyzing the effect of heterogeneous population structure in form of low density at the center of the hybrid zone

As a second intermediate validation step, I introduce heterogeneity in population structure by simulating a low-density area at the center of the hybrid zone. According to the theoretical predictions discussed above, this will create a step in the cline proportional to the width of the barrier and to the density gradient, with the barrier strength being $B = x(k^2 - 1)$, as described above.

Aim 3: Analyzing the effect of leptokurtic dispersal

Thirdly, I will look at the effect of leptokurtic dispersal (i.e. an excess of long-range dispersal). This generates wider clines and is often underestimated, leading to underestimation of selection strength. Therefore, including long-range dispersal values based on inference from the field allows us to get more realistic estimates of the selection strength at hand. To this end, I simulate a uniformly distributed population with dispersal values sampled from the actual empirically calculated dispersal kernel. I then compare the generated cline shape and parameters to the clines produced with a Gaussian dispersal kernel. We would expect clines to get wider with long range dispersal.

Aim 4: Analyzing the effect of patchy distribution with normal and leptokurtic dispersal

Lastly, I want to compare the cline parameters and shape when simulating conditioned on the actual heterogeneous distribution of *A. majus* in the field, leading to smaller-scale variations in density over time and space across the entire hybrid zone. This brings us closer to answer whether local population structure produces a significant change in cline width and shape.

In addition to being able to analyze the effect of different demographic processes on cline parameters and shape, this model is furthermore the first step to be able to look at the interplay of patchy population structure and LD with selected loci, deepening our understanding of how demography can help strengthen genetic barriers, possibly leading to speciation.

Methodology

1. Validation of the model

Conditioning a model on known locations comes with some initial difficulties. Since only a fraction of the plants have been sampled over the 10 years, it is necessary to create a model that can continuously displace plants conditioned on the sampled locations in order to be able to simulate with a larger population size (N) than sampled from the field.

Before conditioning some parameters in the model on observed value ranges from the field, I validate the model by comparing a null model with uniform density and Gaussian dispersal for different s and different densities (i.e. different N), against theoretical predictions. I therefore fit the simulated clines and compare the cline parameters and shape with those theoretically predicted. I also compare proposed and realized dispersal to validate the simulation algorithm.

1. Simulation of a null model: Uniform distribution of plants and Gaussian dispersal kernel

Step 1a) Simulating a uniform distribution

In order to validate the model it is sufficient to simulate a smaller area than sampled from the field. I therefore assigned N locations by randomly sampling N points from a uniform distribution for each generation over 1000 generations in the rectangular area between {418 800, 426750} and {4684 700, 4687 025 } (i.e. 8kM x 2,3kM).

Step 2a) Finding parents with a Gaussian dispersal kernel

- Creating a Gaussian dispersal kernel:

In order to have a non-negative Gaussian dispersal kernel, I transform the proposed seed and pollen distribution from a Cartesian to a polar coordinate system. For each generation, I randomly sample N pairs of numbers from a normal distribution, with the standard deviation of 28m for seeds and a standard deviation of 89m for pollen (a third of the observed values adapted to the smaller range). These variables correspond to the xy-coordinates in a Cartesian coordinate system. I then take the square root of the sum of both squared to give the radial distance moved.

- Assigning parents:

After generating each plant location for a given generation g+1, we then assign corresponding parents in generation g, to each individual kid j in generation g+1, given the seed and pollen dispersal-kernel.

- Finding mothers:

    a. Taking the individual kid j, we draw a circle with radius r, set by a randomly chosen value out of the seed dispersal distribution and find 6 points on the circle in order to have multiple potential mothers to choose from, without needing excessive computational time.

    b. Next, we find the nearest plant to each point in the parent generation g.

    c. Finally, we choose the nearest plant with the closest position to the 6 points on the circle rim.

- Finding fathers:

    d. Firstly, for each individual in the parental generation g we calculate their relative fitness w, based on their surrounding closest 20 neighbors and their genotypes using a linear model (Mallet and Barton 1989):

$$WAA = 1 - s(p-0.5)$$

$$WAa = 1$$

$$Waa = 1+s(p-0.5)$$

    b. *W* being the relative fitness of the genotype, *p* being the frequency of one allele and *s* the selection coefficient. For each found mother m, we draw a circle with radius r, set by a randomly chosen value out of the pollen dispersal distribution and find 12 points on the circle to ensure that in most cases we have more than 5 potential fathers to choose from, without excessive computational time.

    c. Next we then find the nearest plant to each point in the parent generation g.

    d. Finally we randomly chose one of the nearest plants weighted by their relative fitness W.

- Assigning offspring genotype:

    a. After finding the parents of each kid, we assign the offspring genotype by randomly choosing one allele per parent.

2. <u>Checking realized vs. proposed dispersal</u>

To check this algorithm, I calculate the realized dispersal by calculating the distance between mother and father and mother and kid of all simulated generations. I then compare the probability distributions of the realized vs. the proposed distribution.

3. <u>Cline fitting and comparison of cline parameters</u>

In order to fit clines we clump individual plants into demes with a radius of 250 meters for N = 8000 and N = 5262 and 350 meters for N = 15000. In a uniform distribution, we then transform the mean XY position of each deme into a 1D coordinate by subtracting an arbitrary reference point from the X-coordinate and neglecting the Y-position. This then gives us the distance of each deme from the reference point as 1D coordinate. For each deme, I then calculate the total number of alleles and the number of alleles of one type. Next, I run a number of Metropolis trials in order to estimate the three or six parameters to describe a sigmoid or stepped cline. The Metropolis algorithm generates a random walk through the parameter space. After beginning with a random set of initial parameters, they are changed by a random amount, one by one. If these changes increase the likelihood they get accepted and otherwise rejected. Parameters are uniformly perturbed by a factor λ or λ^{-1} on acceptance or rejection. The estimates used here are giving runs of 3000 iterations. The parameters estimated for a sigmoid cline are width, center and F_{st} (the fixation index, the proportion of genetic variance of a subpopulation compared to the total genetic variance, a measure of genetic variability). For a stepped cline we furthermore need to estimate the size and rate of decay of the tails and rates of introgression at each side.

2. Modifications to the null model

After validation of the null model, I continue to separately simulate a low-density barrier in the center of the hybrid zone and leptokurtic dispersal. Finally, I add the local heterogeneous population structure from the field to condition plant locations in the simulation.

2.1. Simulating a low density at the center of the hybrid zone

Step 1b) Simulating a low-density center with uniform distribution

In order to create a low-density barrier I generate a center area between of +/- 80 meter of the midpoint where the two genotypes meet. This is roughly the middle of the x-transect. The width of the low-density center is chosen based on it being <w/5 and greater than the dispersal distance of 82,5 meters. I then reduce plant density within this center area by a factor of 10 (Figure 3). Finding the corresponding parents to each offspring then happens in the same manner as described above (Step 2a).

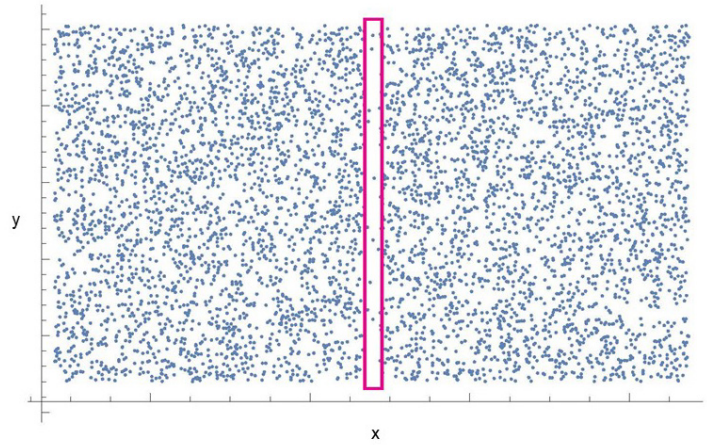

Figure 3 : The red rectangle marks the low-density center area were plants are uniformly distributed in a 10-fold lower density compared to the outer areas.

1. Simulating with a leptokurtic dispersal kernel

Step 2b) Finding parents, using a leptokurtic dispersal kernel

In order to simulate leptokurtic dispersal I use estimated seed and pollen dispersal distributions empirically generated from the field in order to assign parents to offspring.

2. Simulating the heterogeneous plant distribution observed in the field

Step 1c) Distributing plants according to the heterogeneous distribution found in the field

Since only a fraction of plants were sampled over only 11 years in the *A. majus* hybrid zone the conditioning of plant distribution on the spatial and temporal heterogeneity found in the field needs to be extrapolated. We therefore have to define further positions of plants by reiterating the observed distribution over the 11 years over the longer timespan of the simulation. In the years between 2017 and 2018 there were fewer individuals than usual, so they are combined, leaving us with positional data for 10 time points. In order to maintain a stable population size (N), we either:

a) Randomly sample from the given locations if the desired N is smaller than the recorded locations at the given time point,

or

b) If N is greater than recorded locations by k plants, we first assign the given locations to the first k plants, and then resample N-k locations randomly and assign positions displaced at a random angle of a circle with 3m radius from the chosen locations.

Tying the distribution of additional plants to the exact localities of found plants in the field allows us to simulate the realistic spatial structure, at least over small spatial and temporal scales.

Results

Frequency Dependent Selection vs. Selection against Hybrids

Simulations with frequency dependence need more computational power in order to find and store all necessary information of each individual and for its neighbors. Therefore, I ran only three replicates under the premise of frequency-dependent selection and compared them to simulations with selection against hybrids, to be able to continue with these in the interest of time. In cases of weak selection pressures theory predicts that clines maintained by frequency dependent selection are similar in shape and width to clines maintained by selection against hybrids with a uniform distribution and a Gaussian dispersal kernel. Both should create clines shaped like a basic sigmoid, and we can see that this is the case and both forms of selection create basic sigmoid shaped clines, even when fitted for stepped clines. However, when comparing the average clines produced by the two different forms of selection directly we can see that cline width is significantly greater with frequency dependent selection (Fig.4, TTest, $p = 0.02$).

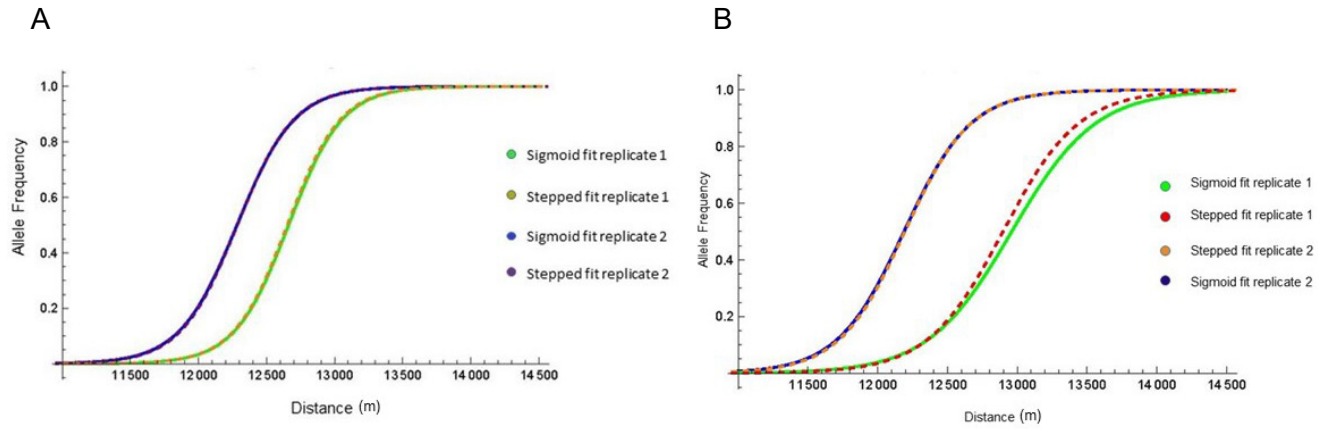

Figure 4: A: 2 replicates of clines from simulations with selection against hybrids and s = 0.1 and N = 8000, each fitted once against the basic sigmoid and stepped cline. B: 2 replicates of clines from simulations with frequency dependent selection and s = 0.1 and N = 8000, each fitted once against the basic sigmoid and stepped cline.

A

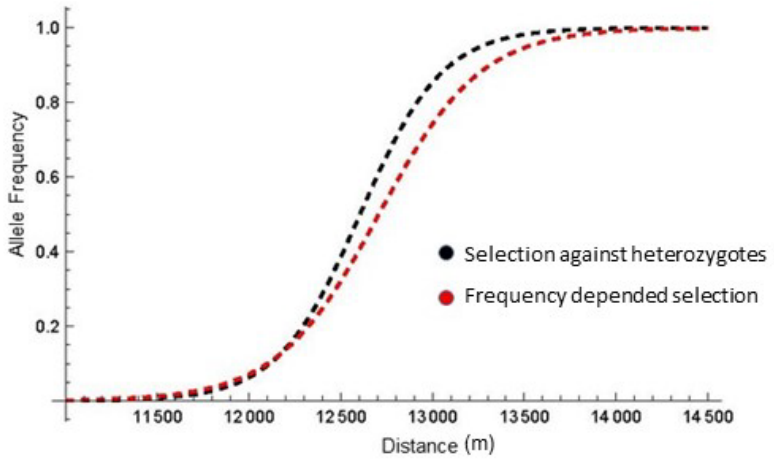

Figure 5: A: Average clines with frequency dependent selection and selection against hybrids, s = 0.1, N = 80000.

Realized vs. Proposed Dispersal

Looking at the proposed dispersal in comparison to the realized dispersal of simulations, we can see that both distributions are approximately Gaussian. While simulated pollen dispersal maps well onto the proposed probability distribution, simulated seed dispersal is slightly greater (Fig. 6).

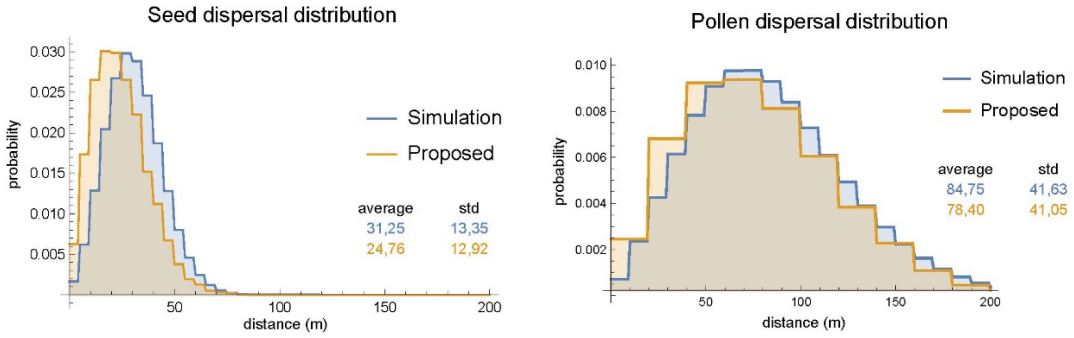

Figure 6: Realized vs. proposed seed and pollen dispersal distribution of single simulation with N =8000 and s = 0.1.

Width vs. selection

When comparing width vs. selection from simulations and theory we find that simulated widths of different N qualitatively follow theoretical predictions, with clines becoming narrower with stronger selection. However, quantitatively we get wider clines than predicted from theory (Fig. 7). Cline width of the simulations with N = 8000 are not significantly wider than clines (s = 0.1 p = 0.7). Comparing the width between the two different density paradigms, higher plant density creates overall narrower width (Fig. 7)

A

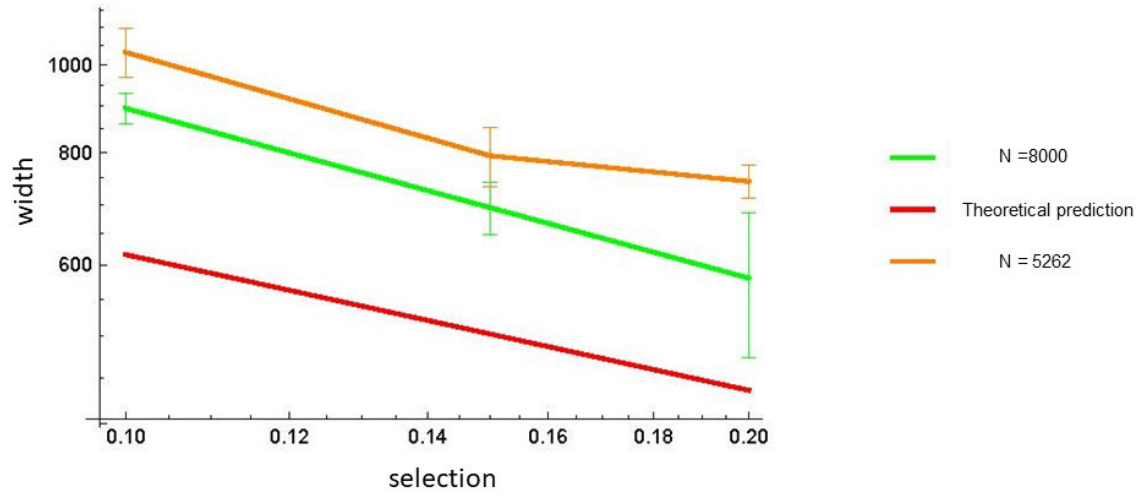

Figure 7 :Width plotted against selection on a log-log scale for N = 8000, and s =0.1, 0.15, 0.2, for 5 replicates each, as well as the theoretically predicted width at all three s.

Positional Changes in 2D

Comparing two different two dimensional plots of individuals across space colored by their genotype, of N = 8000, with s = 0.1, one can see that cline width could be affected by slight positional changes of heterozygote positions along the x-Axis (Fig.8).

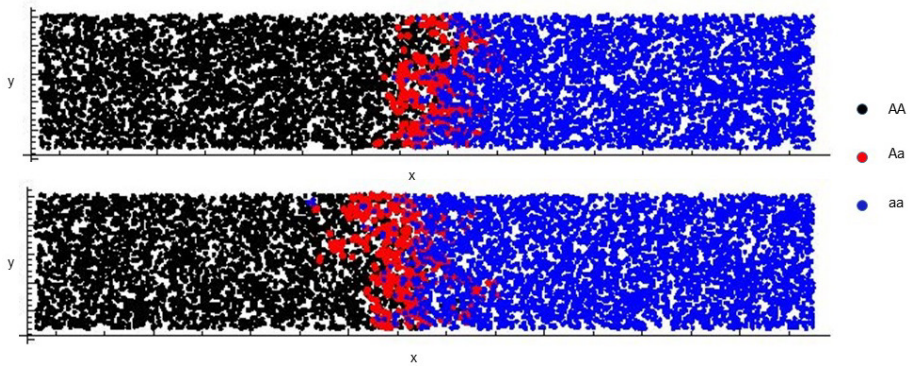

Figure 8: 2 replications of individual locations colored by genotype after 1000generations, N = 8000 , s = 0.1.

Density Barrier

For a low-density center running for 200m, with 10% of the overall plant density, with N = 8000 and s = 0.15, we can see a steeper cline and slightly but not significant narrower width (TTest, p = 0.5), with longer tails of introgression. Some replicates form a step, while others remain in a sigmoid shape, yet the likelihood estimates remain slightly better when fitted to a stepped shape. (Fig.9). Furthermore, the cline is tied more tightly with its center to the low-density sink at x = 12 475 with the mean center point being c = 12 445 and a standard deviation of 39,2m.

In comparison, the mean center point of a uniformly distributed simulation with same N and s is c = 12 324, with a standard deviation of 303,42m.

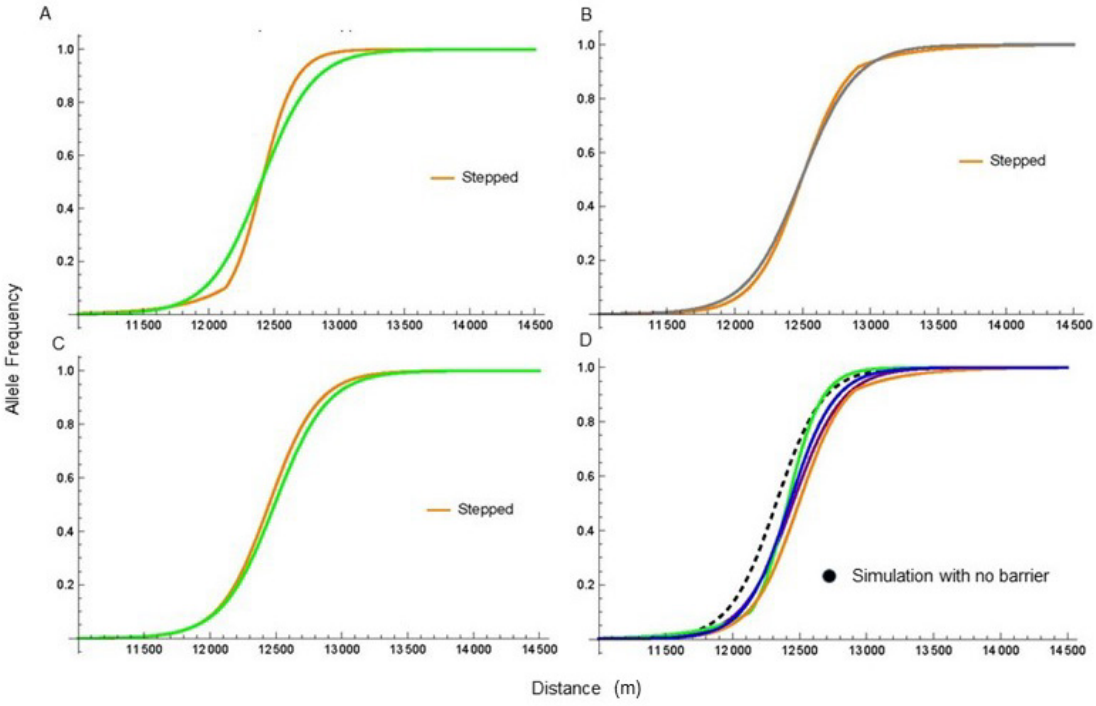

Figure 9 : A: One replicate of cline with density barrier N = 8000 , s =0.15, once fitted for basic sigmoid and once for stepped cline, B: One replicate of cline with density barrier N = 8000 , s =0.15, once fitted for basic sigmoid and once for stepped cline, C; One replicate of cline with density barrier N = 8000 , s =0.15, once fitted for basic sigmoid and once for stepped cline, D: All replicates fitted for stepped cline with basic sigmoid of average uniform distribution without density barrier.

Long Range Dispersal

In Figure 10 fitting one simulation with N = 8000 and s = 0.1 with long-range dispersal from the found distribution in the field. As expected we see a stepped cline, with greater width than compared to the same simulation with a Gaussian dispersal kernel with the same standard deviation.

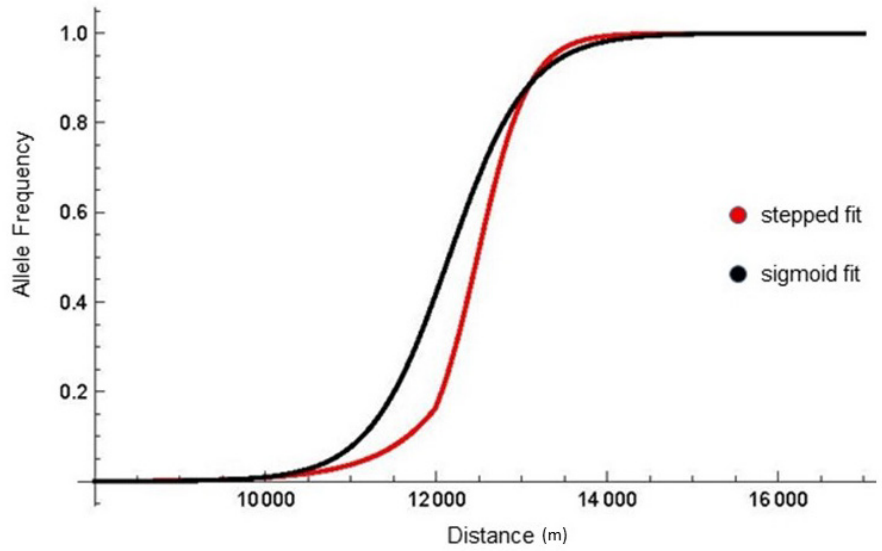

Figure 10: One simulation with long-range dispersal based on the dispersal distribution generated from the field fitted against basic sigmoid and stepped cline

Discussion

Dispersal Ranges

Across different densities, we see that our realized seed-dispersal values vary slightly to the proposed seed-dispersal. This is probably an artifact created by us only searching for six potential mothers. This finds a limited number of individuals and we might not find an individual corresponding to the exact chosen seed value. Choosing the amount of points we search potential mothers from will always be a compromise between optimizing computational time and efficiency and best model fit. Increasing plant density could lead to better results since with higher numbers of individuals the probability is higher to have multiple ones in a 360 degree seed dispersal range from the kid and finding one with only six points.

Greater width of clines from simulation of null model

Simulated cline widths from simulations with uniform distribution and Gaussian dispersal are broader than theoretical predictions at all selection strengths (Fig.7). Looking at the two different densities, this might be a problem of finite population size, with higher density of individuals approaching theoretical predictions, since these are based on infinite population sizes. In order to confirm this intuition, more simulations with even higher density would need to be conducted.

However, next to finite population size a further theoretical assumption we are breaking with is one dimensionality. This could also be causing the deviation in realized vs. predicted cline width. Theoretical cline models are based in one dimension, our simulation works in a two dimensional space. Here, the first step is to make sure that our dispersal is transformed into its one dimensional value, by dividing it by two before inputting it into formula (1). Looking at the two dimensional distribution we can see that when slicing the area into x-transects the width of heterozygote genotype occurrences seems to remain relatively stable while slightly shifting its position here and there (Fig.8). These positional shifts could be the reason why the one dimensionally fitted and aligned cline width is overestimated. Polechova and Barton (2011) have discussed effects of positional change on the width of clines due to genetic drift. Here, looking at individual clines drift narrows expected width, yet when looking at multiple the opposite occurs, showing that width estimations might depend on scale. However, Polechova and Barton were focusing mostly on clines created by near one dimensional habitats (e.g. along a coastline or rivers) (Polechova and Barton 2011). We can see that it is therefore necessary to generate a better understanding of the effect of positional fluctuation on cline width and shape in two dimensions. One way of analyzing this effect could be to take local width averages along x-transects and therefore account for positional shifts.

Density barrier

Simulating 10-fold density reduction in a 160m center zone was not enough to create a stable barrier to gene flow. However, simulated clines have better likelihood estimates when fitted against stepped clines than sigmoid and 2 out of 4 produce a slight step. As next steps it would be good to try for larger low-density ranges and different scales of density reduction in order to find the range in which a consistent barrier to gene flow starts. We can however clearly see that a density barrier ties the cline to the low-density center and therefore immobilizes it, which is also theoretically predicted (Barton 1979). One could try to use this effect to create clines with more stable positions. For this one could play around with different density gradients, which would tie the cline to a center point without creating a barrier for gene flow. As further steps leading us closer to simulating patchy distributions across the entire area one could envision simulating multiple density sinks, which are not placed in the center of the simulated area. This would allow us to get a better theoretical understanding and approximation of the effect of density reduction on cline position, movement, width and shape.

Long-range dispersal

Long-range dispersal seems to create a stepped cline as expected, however more replicates are needed to see how consistent this pattern is, since we expect high fluctuation. Long-range

dispersal can cause not only wider clines due to migration of foreign genotypes into the center and past, but also create a step when leaving offspring. However, since long range migration involves a lot of chance, because it can happen in any direction, more replicates would show how robust this step can appear and if it would, as is expected to happen at both ends ~50 percent of its occurrence.

Future Directions

This thesis is the first step towards theoretically untangling the effects of different elements of population structure on cline parameters and shape. We can now start filling the gap of current theoretical work on clines, by firstly understanding the effects of long-range dispersal and low-density patches separately. This allows us to create theoretical predictions for each scenario and then compare them in a combined condition, which is most likely to be commonly found in reality. Before being able to take this step and condition my model on spatial and temporal varying local population structure from the field, it is necessary to understand why this model does not reproduce theoretical predictions. After this issue is solved one can continue to condition the locations of simulated individuals on the actual locations found in the field, as described above, in order to get a realistic simulation of existing patchy distributions. This allows us then to compare clines produced by simulations with differing selection strength to the descriptive clines of the *A. majus* zone (Barton Group, unpublished), leading us to a better theoretical understanding of the effects of populations structure on selected loci.

Literature

Barton, N. H. (1979). The dynamics of hybrid zones. *Heredity*, *43*(3), 341-359.

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual review of Ecology and Systematics*, *16*(1), 113-148.

Barton, N. H. (1986). The effects of linkage and density-dependent regulation on gene flow. *Heredity*, *57*(3), 415-426.

Barton, N. H., Gale, K. S., & Harrison, R. G. (1993). Genetic analysis of hybrid zones. *Hybrid zones and the evolutionary process*, 13-45.

Bazykin, A. D. (1969). Hypothetical mechanism of speciaton. *Evolution*, *23*(4), 685-687.

Bridle, J. R., Baird, S. J., & Butlin, R. K. (2001). Spatial structure and habitat variation in a grasshopper hybrid zone. *Evolution*, *55*(9), 1832-1843.

Buno, I., Torroja, E., López-Fernádez, C., Butlin, R. K., Hewitt, G. M., & Gosálvez, J. (1994). A hybrid zone between two subspecies of the grasshopper Chorthippus parallelus along the Pyrenees: the west end. *Heredity*, *73*(6), 625-634.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, *49*(4), 327-335.

Haldane, J. B. S. (1948). The theory of a cline. *Journal of genetics*, *48*, 277-284.

Mallet, J. (1986). Hybrid zones of Heliconius butterflies in Panama and the stability and movement of warning colour clines. *Heredity*, *56*(2), 191-202.

Mallet, J., & Barton, N. (1989). Inference from clines stabilized by frequency-dependent selection. *Genetics*, *122*(4), 967-976.

Mallet, J., Barton, N., Lamas, G., Santisteban, J., Muedas, M., & Eeley, H. (1990). Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in Heliconius hybrid zones. *Genetics*, *124*(4), 921-936.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087-1092.

Moore, W. S. (1977). An evaluation of narrow hybrid zones in vertebrates. *The Quarterly review of biology*, *52*(3), 263-277.

Nagylaki, T. (1976). Clines with variable migration. *Genetics*, *83*(4), 867-886.

Polechová, J., & Barton, N. (2011). Genetic drift widens the expected cline but narrows the expected cline width. *Genetics*, *189*(1), 227-235.Stankowski, S., Shipilina, D., & Westram, A. M. (2021). Hybrid zones. *eLS*, *2*, 1-16.

Surendranadh, P., Arathoon, L., Baskett, C. A., Field, D. L., Pickup, M., & Barton, N. H. (2022). Effects of fine-scale population structure on the distribution of heterozygosity in a long-term study of Antirrhinum majus. *Genetics*, *221*(3), iyac083.

Szymura, J. M., & Barton, N. H. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads, Bombina bombina and B. variegata, near Cracow in southern Poland. *Evolution*, *40*(6), 1141-1159.

Szymura, J. M., & Barton, N. H. (1991). The genetic structure of the hybrid zone between the fire-bellied toads Bombina bombina and B. variegata: comparisons between transects and between loci. *Evolution*, *45*(2), 237-261.

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., ... & Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences*, *115*(43), 11006-11011.

Wang, L., Luzynski, K., Pool, J. E., Janoušek, V., Dufková, P., Vyskočilová, M. M., ... & Tucker, P. K. (2011). Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. *Molecular ecology*, *20*(14), 2985-3000.

Westram, A. M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., ... & Butlin, R. (2018). Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evolution Letters*, *2*(4), 297-309.

Whibley, A. C., Langlade, N. B., Andalo, C., Hanna, A. I., Bangham, A., Thébaud, C., & Coen, E. (2006). Evolutionary paths underlying flower color variation in Antirrhinum. *Science*, *313*(5789), 963-966.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*(2), 97.