Modelling the effects of Gene Flow and Selection on the Canalisation of an established Genetic Network

Author: Matthew Campos
Submitted August 2020

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science/Research at Imperial College London
Formatted in the journal style of Evolution & Development
Submitted for the MSc in Computational Methods in Ecology and Evolution

1 Declaration

All raw data collected were from the simulation I created for my project. The simulation requires the input of genetic model systems, and mathematical equations used to derive output. The model system and equations used in the simulation were sourced from the work of *Omholt et al* (2000) and *Gjuvsland et al* (2007). I was reponsible for data processing, cleaning and analysis. All analyses presented in the paper are from the simulations, with the help of my supervisor.

2 Acknowledgements

I would like to firstly thank Dr. Scott Rifkin for being a wonderful supervisor and guiding me throughout the project. This includes understanding background knowledge, results and overall research purposes. Secondly, thank you to Dr. Thomas Bell for agreeing to be my internal supervisor, making sure I am aware of the process of the project and ensuring my safety during such difficult times. Also to Dr. Samraat Pawar and Dr. James Rosindell for co-leading a challenging but rewarding CMEE course.
Finally I would like to thank the laboratory of Dr. Rifkin- Antonia Darragh, Jessica Bloom, Alexis Cugini, Yang Bing and Rachel Goodridge for being very welcoming and having wonderful and insightful weekly meetings. Good luck with everything!

3 Abstract

Studies have shown how the role of genotypic evolution on phenotype. How simple alterations in the pathways can lead to pronounced differences in species morphology. Using the genetic model of *Omholt et al*(2000) and *Gjuvsland et al*(2007), I investigate the interplay between gene flow and adaptive processes in the development of a genetic network showing a balance between selection, a strong driver towards canalisation of the network, and gene flow which attempts to homogenise the populations, with the expectation that long temporal periods of evolution with gene flow should lead to a more canalised network compared to without gene flow. Canalisation is represented by fitness values, which are determined by trait values. Heterozygosity in the population has been shown to drive fitness and quickly lead to canalisation while homogeneity relies on chance beneficial mutations, recombination or migrant allele values to become robust. The results show a possible effect that migratory pattern has on the canalisation of a network, and how it is network development could be context dependent. Outcomes from this study can be used to better understand phenotype-genotype relationships and the patterns that drive the development of species.

4 Introduction

Species migration can result in the following: (i) it leads to colonisation of new habitats and create new subpopulations spatially separate from the main population. Over time, vicariant events can cause the dispersed species to become reproductively isolated. This is known as allopatric speciation and it leads to the splitting of lineages, differentiating both related populations long-term (Barber 1999, Coyne 1992). This differentiation becomes more pronounced with geographic distance. Restricting gene flow allows both sets of species are able to rapidly evolve in their local optimums (García-Ramos & Kirkpatrick 1997). Overall, species that are reproductively isolated have more pronounced modifications which can be observed phenotypically and genotypically (Pongratz, Gerace & Michiels 2002, Sato, Isagi, Sakio, Osumi & Goto 2006). (ii) Migration can allow once isolated species to enter each other's habitats. If species are capable of interbreeding, this introduces new sets of alleles into an environment and interspecies reproduction passes on varying heritable genes which changes the developing genetic makeup of local species. The latter case is gene flow which helps to maintain the genetic diversity in an area but homogenize populations over long periods of time, through the recombination of genes (Sato et al. 2006, García-Ramos & Kirkpatrick 1997).

Advancements in genotypic techniques now makes us capable of furthering our understanding of phenotype-genotype relationship. As organisms evolve, phenotypic evolution is assisted with genotypic evolution. Collective expression of genes through pathways influence the morphologies we observe in species (Hinman, Yankura & McCauley 2009). Hereditary genome alterations through random changes in molecular mechanisms change varying aspects of the species (Chandrasekaran & Betrán 2008) These molecular changes induced by mutation and recombination lead to the variation of descending species (Chandrasekaran & Betrán 2008, Ohno 1999, Brown 2002). Over time, evolutionary forces involving genetic drift and selection acts on these polymorphisms and those most fit passes their variant genes and phenotype as a result, to future generations. This is the foundation of Darwin's theory natural selection.

Observation of the genome network development can give insight into species evolution. Orr showed that there is variation with respect to genetic differences or gene influence on phenotype. The effects of adaptive and non-adaptive processes vary among species where there is no common set of genes involved, nor is the effects and interactions of the genes similar for species (Orr 1998) Although long temporal period has shaped a myriad of genetic function and interactions, what can be investigated is the pattern at which these genetic processes develop over time. Genetic network simulations can be used to understand these patterns of evolution and the effect on phenotype-genotype relationships. Long temporal periods allow genetic interactions within a network to robustly develop, canalising the network (Orr 1998, Lynch 2007). *Lynch et al*(2007) highlighted the significance of non-adaptive processes as well in shaping genetic networks. The study showed that networks can still evolve its architecture and become redundant even without the influence of natural selection (Lynch 2007). Robustness can evolve from the effects of epistasis, additivity and dominance, all of which are connected (Omholt, Plahte, Øyehaug & Xiang 2000).

Species evolution is non-linear but this continuous process over long temporal periods results in the accumulation of optimal genetic adaptations that results in a robust network structure

that are adaptive and resistant to perturbations (Hinman et al. 2009). There is a balancing act as selection aids to propagate fitter variants in a population, while mutation and environmental change limits such propagation (Burt 1995). When migration is included, a balance between migration and selection will influence gene frequencies of future generations (Brown & Pavlovic 1992). As species evolve, studies have shown that pathways have a safety margin, that make them resistant to deleterious changes (Bourguet 1999). Species best suited to their environment will evolve to their local optima, which we can represent as a quantitative trait value. The further apart these values are, what I label as *environmental distance*, the greater the variance of the two species. I will consider the effects of varying migration rates, variants of a genetic network, the environmental distance between two populations and patterns of migration.

Ecological events can eliminate barriers and allow species to migrate into new environments, introducing new sets of genes in an environment. The presence of variant genes and network structures from gene flow hinders local adaptation and fixation of adaptive genes (Burt 1995). Previous research has looked at the effects of gene flow, selection and mutation at generating local adaptation at the phenotypic level, showing how maintenance of alleles and linkage is important in adaptation (Yeaman & Whitlock 2011). Even with random perturbations, there are bounds for which selection for canalization can act on, through the aid of genetic modifiers. They also revealed that under migration selection balance, selection for robustness increases with the migration rates (Proulx & Phillips 2005).

This paper investigates how these forces of gene flow and selection affect the development of genes and the genetic network. Even once a robust structure is reached, if it can resist change and maintain its network despite disruptions from gene flow, focusing on the regulatory interactions that are modified during the networks evolution and how these changes affect trait values (Hinman et al. 2009). As a genetic network evolves, there exists a threshold which is actively regulating these homeostatic genes (Gjuvsland, Plahte & Omholt 2007). As selection for robustness occurs within the local population, it can give insight into the change in architecture and reveal statistically significant interaction (Gjuvsland, Hayes, Omholt & Carlborg 2007). Using a multilocus system, I will construct a genetic network and simulate the effects over many generations and see how the output of the network changes, specifically looking at allelic interactions and tracking the fitness over time. Variance in fitness should decrease as a genetic network becomes robust, meaning after migration network should have lower variance than before migration. The expectation is that after migration, a more robust network is formed when compared to before migration. At the start allowing new alleles to enter the population hinders the network development, but other evolutionary forces including selection should counteract these perturbations and result in a robust network (García-Ramos & Kirkpatrick 1997). Especially when the migrant network is a different structure, gene flow will allow maladaptive alleles to enter and should these persist, will impose a fitness cost to individuals (Tigano & Friesen 2016).

5 Methods

I wrote a R script that constructs a genetic network, and a variant form, and simulates its evolutions, allowing migration to occur between two populations. All functions to perform adaptive and non-adaptive processes were written from scratch and implemented in the simulation. The following functions are:

- Population: initialises the starting populations of specified size where each individual (row) contains 12 allele sites (4 per gene). Since it is a di-allelic model, it is a 2-dimensional array.
- Fitness: determines the fitness value of each individual based on their trait values and used as a probability for offspring contribution. A heavy tailed Cauchy distribution is used to determine fitness value from trait values. Each individual has a probability of passing on their genotype to the next generation and function randomly samples from the distribution to select parents, representative of genetic drift.
- Mutation: produces an array same dimensions as the population and random uniformly distribution of values to determine which sites undergo mutation based on inputted mutation rate. Generates a new value using a normal function with current value as the mean and a standard deviation of 0.001.
- Recombination: randomly chooses which site, and if any consecutive sites downstream, to switch allele values for each individual.
- Migration: using a uniform distribution, randomly generates values for each individual in the migrant population to determine which individuals will migrate and replace those in the main population. Population is kept constant in both populations, representing a balanced dispersal between immigration and emigration (Rice & Papadopoulos 2009, W. Morris, E. Diffendorfer & Lundberg 2004).

5.1 The model

A di-allelic interlocus model from the research of *Omholt et al*(2000). In this case, all the genes are hereditary, representing only the regulatory and coding region which determine protein expression and rate of expression. Studies has shown that mutations along the coding region are known to cause morphological variation within species (Stern & Orgogozo 2009). This model structure evolves dominance through epistatic interactions and autoregulatory effects. Using a system of equilibrium solutions and solved ordinary differential equations (ODE), simulated protein concentrations corresponding to phenotype are measured over time (Omholt et al. 2000). Here I consider the loci as quantitative factors of protein function, and trait value is determined by protein concentrations. The greater the amount of protein expressed, the larger the trait value. The model consists of three genes, X_1, X_2 and X_3. Let j represent the genes where $j = 1, 2, 3$, each gene X_j consists of two alleles, X_{j1} and X_{j2}. This leads to the formula:

$$y_j = X_{j1} + X_{j2} \tag{1}$$

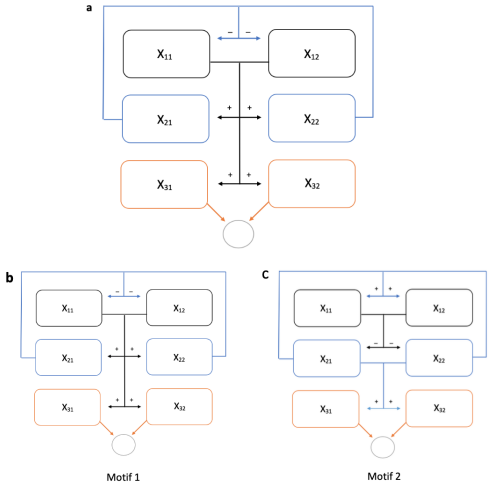


Figure 1: Diagram showing the genetic model and two variants used to represent the migrant population. (a) Interlocus model of the population in focus. Lines labelled with mathematical symbols showing the interactions between genes. Gene X_1 interacts with both gene X_2 and X_3, positively regulating both of them. To limit site values below infinity, gene X_2 is reponsible for negatively autoregulating X_1. There is an output for each gene where $j = 1, 2$ and $y_j = x_{j1} + x_{j2}$. Gene X_3 contains the trait values for each individual, which is the output. Circle represents phenotype which is determined from trait values using a Cauchy Distribution. (b) and (c) represent models for the migrant population. (b) is the same pathway and regulation as (a) however (c) is switched where gene X_1 negatively autoregulates X_2, and gene X_2 positively regulates X_1 and X_3. Again, output of gene X_3 are the trait values used to derive fitness.

Where y_j is the total protein concentration at each gene. There are four sites which represent the different factors affecting protein production. These are α, γ, ϑ and P. α is the protein production rate while γ is the degradation rate (Omholt et al. 2000). For both sets of populations, a single gene, X_3 determines the trait value for individuals and quantifiably differentiates the populations in terms of morphology (Orr 2001). For the population in focus, gene X_1 positively regulates gene X_2 and gene X_3, and gene X_1 is negatively regulated by gene X_2. This is to regulate trait value and prevent the value from exceeding to infinity. As gene X_2 increases in expression, it decreases X_1 expression, negatively autoregulating the system and limiting its value. Let $j = 1, 2, 3$ and $i = 1, 2$, from the separate researches of *Omholt et al*(2000), and *Gjuvsland et al*(2007), R_j is a regulatory Hill Function representing a Michaelis-Menten mechanism, where $S(y_j, \theta, P) = \frac{y_j^P}{y_j^P + \theta^P}$. The Hill Function explains the relationship between regulator and producer, where $\vartheta$is the amount of regulator needed for 50% production rate and P affects the steepness of the curve (Gjuvsland, Hayes, Omholt & Carlborg 2007, Omholt et al. 2000). Should the network be negatively regulated, it leads to the following equation:

$$R_j(y) = 1 - S(y, \theta_j, P_j), j = 1, 2 \tag{2}$$

And if positively regulated:

$$R_j(y) = S(y, \theta_j, P_j), j = 1, 2 \tag{3}$$

Again, letting $j = 1, 2, 3$, as gene X_1 positively autoregulates gene X_2 and gene X_3, and gene X_2 negatively autoregulates gene X_1, this results in the following equations:

$$R_{1j}(y_2) = 1–S(y_2, \theta_{2j}, P_{2j}), \tag{4.1}$$

$$R_{2j}(y_1) = 1–S(y_1, \theta_{1j}, P_{1j}), \tag{4.2}$$

$$R_{2j}(y_1) = 1–S(y_1, \theta_{3j}, P_{3j}) \tag{4.3}$$

μ is the ratio of αand γper locus. Using the equilibrium solutions, total protein concentration is

calculated by the following equations:

$$y_1 = \mu_{11}(1–S(y_2, \theta_{21}, P_{21})) + \mu_{12}(1–S(y_2, \theta_{22}, P_{22})) \tag{5.1}$$

$$y_2 = \mu_{21}(S(y_1, \theta_{11}, P_{11})) + \mu_{22}(S(y_1, \theta_{12}, P_{12})) \tag{5.2}$$

$$y_3 = \mu_{31}(S(y_1, \theta_{31}, P_{31})) + \mu_{32}(S(y_1, \theta_{32}, P_{32})) \tag{5.3}$$

5.2 Migrant network

For the first motif, the genetic network will be the same as the main population, just evolving to a different local optimum trait value of either 65 or 80. For the second motif however, the difference is that gene X_1 negatively regulates gene X_2, while gene X_3 and gene X_1 are positively regulated by gene X_2. The formulas used to derive y_1, y_2 and y_3 values for the migrant population are as follows:

$$y_1 = \mu_{11}(S(y_2, \theta_{21}, P_{21})) + \mu_{12}(S(y_2, \theta_{22}, P_{22})) \tag{6.1}$$

$$y_2 = \mu_{21}(1–S(y_1, \theta_{11}, P_{11})) + \mu_{22}(1–S(y_1, \theta_{12}, P_{22})) \tag{6.2}$$

$$y_3 = \mu_{31}(S(y_2, \theta_{31}, P_{31})) + \mu_{32}(S(y_2, \theta_{32}, P_{32})) \tag{6.3}$$

This is to represent the concept of differentiated species but can still integrate in the other population and interbreed.

5.3 The simulation

A total of 44 permutations based on conditions in *(see Appendix A)* of environmental distance, genetic network structure, migration rates and migration patterns were simulated for 1,200 generations each run. For the effect of genetic drift and to account for the large deviations of values, a Cauchy distribution is used to generate fitness probabilities per generation. Since the Cauchy distribution is characterized for its heavy tails. The values entered in the Cauchy distribution are the desired trait values. It is important to note that environment is kept constant. Both populations were kept constant at 500 individuals. The main population evolved to a trait value of 50 with a standard deviation 8, while the migrant population alternated between 65 and 80 with standard deviation 10. The large standard deviations characterise the varying forms of morphology that can be noticed in species. The trait values represent the environments of both populations and the local optimums they evolve to.

For the simulation we assume that both populations have the same size and stay constant, with migrants replacing individuals. There is no spatial structure and all individuals have an equal chance of being replaced. Both populations undergo divergent selection, stabilising in their own environments to different specified trait values, thus differentiating the populations over time (Sato et al. 2006). Alleles for each individual can either be homogenous, using a uniform distribution to determine starting value between 0.1 and 0.3 for both populations, or heterogenous, using a uniform to randomly generate the starting allele values, again between 0.1 and 0.3. If the population is homogenous, each individual in the population starts with the same value at each locus, otherwise have differing values if heterogenous. Recombination is equal chance at any locus and interchanges the alleles and everything downstream. Mutation

can occur at each locus by randomly deviating from the current value. The probability is the same constant for both populations where each locus has an equal chance of mutating. Mutation probability is kept constant at 0.0011 per site. A mutation in the second gene will have trans-regulatory effects as gene y_2 negatively autoregulates gene y_1, while the effect of gene y_1 will affect the expression values of gene y_3. Since all genes in the model represent regulatory and coding regions, mutations in any site can be considered to affect phenotype, for its pleiotropic effects (Rice & Rebeiz 2019, Landry, Lemos, Rifkin, Dickinson & Hartl 2007).

Fitness is reproductive success, or the probability of being a parent and passing on their allele values which is determined by phenotypic value. Each individual per generation has no limit as to how many times they can be a parent, however the standard deviation of 8 and 10 in the Cauchy distribution attempts to produce varying combination of parents. Migration rates varied between 1%, 3% and 5%. As migrant individuals enter the population, they randomly replace individuals in the population. With constant population size, this represents immigration and emigration. Furthermore, low migration rates were used to prevent migration population from completely replacing the original population and allowing the network to be able to adapt to the new values. Both populations have a burn-in period of 80 generations to evolve in their own environments before migration can happen. Also, migration only occurs till the 700th generation. The remaining 500 generations are to assess how the network responds to the migration. Patterns of migration were also considered, varying between each generation, every 10 generations, every 5 generations and random (between 1% and 5% each occurrence) after the 80th generation.

5.4 Analysis

Analysis was done on the recorded fitness, trait values and population arrays. At the end of each simulation, fitness is normalised by dividing fitness probabilities with the median of Cauchy distribution for the local population, with 1.0 being the highest possible fitness value. Firstly, control conditions of no migration were simulated to see how rapidly isolated networks evolve. Without migration, I expect rapid evolution of allele values, especially for the heterogenous population due to varying alleles present (García-Ramos & Kirkpatrick 1997). To analyse robustness, I calculate a robustness ratio. The fittest individuals before and 10 generations after migration were recorded and replicated such that there were 4 separate mutations per site. Trait values were again inputted into the Cauchy distribution to determine fitness values. The robustness ratio is then the variance in fitness after migration divided by variance in fitness before migration. Ratios were then log transformed as to linearise and make it less skewed. A negative value is thus desired for robustness and analysis of variance test is done to see if any factors significantly contribute to robustness.

6 Results

6.1 Cauchy Distribution

A Cauchy distribution was used to determine the fitness values based on trait values. The plot on the right shows how the shape of the distribution is similar to a normal distribution as seen in *figure 2*. The peak represents the median, in this case the maximum probability a trait has of passing on to the next generation. Unlike a normal distribution, it is narrower within the curve. Even with large standard deviation, *figure 2* shows how small deviations from the desired trait values can significant decline in probability of an individual being selected for reproduction. These desired trait values represent the environment and the distance between them is the *environmental distance*. Individuals within their populations are evolving to their respective peaks. Long tail distribution allowed for the possibility of migrant genes values persisting for generations, allowing investigation into how the network develops with invading alleles.

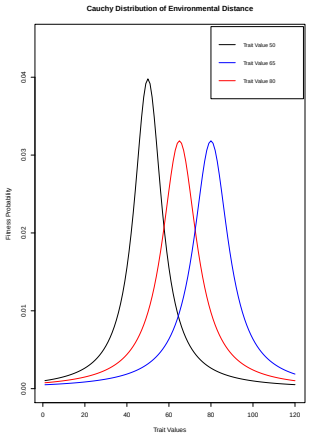


Figure 2: Plot of the Cauchy Distribution used to derive fitness probabilities (for reproduction) and fitness values (normalising) of individuals. Distribution is also representative of the environmental distances respective populations evolved to. The main population (black) evolved to a trait value of 50, where the peak trait value has a reproductive probability of 3.98%. Migrant populations either evolved to a trait value of 65 (blue) or 80 (red).

6.2 No Migration

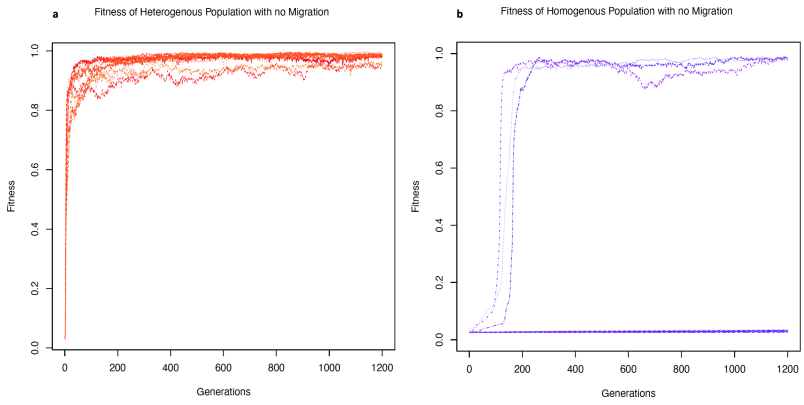


Figure 3: graphs showing the fitness evolution without migration for the populations that started as heterozygous or homozygous. (a) shows the evolution of a network that is homogenous network, where all starting allele values are the same per site per individual, randomly generated using a uniform distribution between 0.10 and 0.30. (b) shows how a heterogenous network evolves where starting values (between 0.10 and 0.30) for each site are different. Homogenous populations rarely evolved and had a mean fitness of 0.165, standard deviation of 0.035 while the heterogenous populations had an average fitness value of 0.955 with a standard deviation of 0.324.

Figure 3 plots the fitness values of starting homogenous and heterogenous population, they both show rapid evolution and have very similar patterns of sharp gradients in the early generations.

The graphs shows how genetic variation helps a population evolve rapidly while homogenous rely on chance beneficial scenarios to improve fitness. A population that began as heterogenous evolves faster as there was greater variation in the allele values, allowing for better combinations to be selected for. Gradients of the line at the start are steep and vary once it begins to plateau. The results show that when there is no migration happening, the genetic network does evolve rapidly to the desired trait values. In a very short time period, it reaches an average fitness of 0.955. Normalising the fitness reveals how rapidly the trait values rises to 50, plateauing around a fitness value of 1 well within 100 generations. Although there are cases of the fitness deviating, overall the rise to the average fitness is steep. The homozygous population had a lot more noise throughout the simulation. There were instances where the population did not even increase in fitness and stayed near within the confine of its starting fitness value. There are a few instances where fitness sharply rose to around 1.00 and stayed there.

6.3 Migration

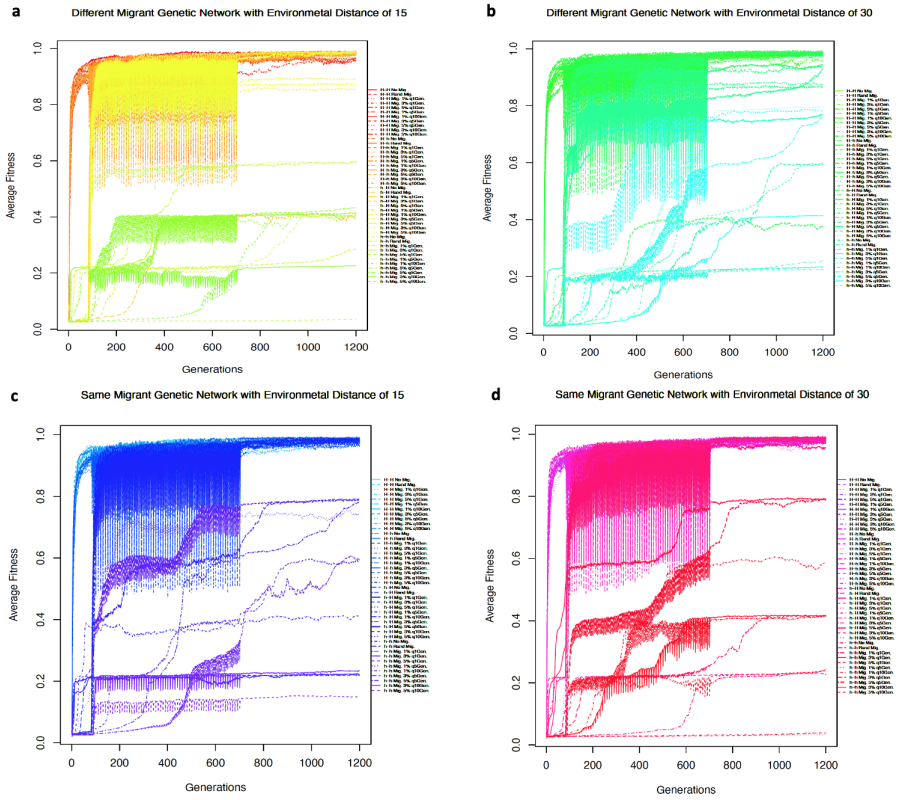


Figure 4: Plots showing the effect of migration on fitness over time. Migration occurs between generations 80 and 700, where the network gets 500 generations afterwards to try and recover. Each line is one of the 44 different starting conditions for the 4 combinations of genetic variant and environmental distance. 'H' means heterozygous while 'h' is homozygous, and 'q' means every. For example, 'H-h Mig. 1% q1Gen.' is heterozygous (main population), homozygous (migrant population), migration rate of 1% every generation. The graphs reveal the fluctuations in fitness values caused by gene flow where (a) shows the effect of migrant network with different regulations from the main population, evolving to a trait value of 65 (Environmental Distance of 15). It is a variation where gene y_1 negatively regulates gene y_1, and gene y_2 positively regulates the other genes, including gene y_3 which is the trait values. Mean fitness was 0.743, with a standard deviation of 0.375. (b) is a migrant network same as in graph (a) where the regulations are different, instead evolving to 80 (Environmental Distance of 30). Mean fitness value was 0.753, standard deviation of 0.364. (c) shows the effect of a migrant population but with the same regulations as in the main population. Therefore, gene y_2 is negatively regulated by gene y_1, and gene y_1 positively regulates the other genes. Similar to graph (a) it is evolving to a trait value of 65. Mean fitness of 0.769 and standard deviation of 0.359. (d) same regulation as the migrant network in graph (c) however evolves to a trait value of 80. Mean fitness of 0.743 and standard deviation of 0.381.

With migration, the genetic network was still able to evolve at time to high fitness value but there was greater variation. *Figure 4* shows gene flow can disrupt the development of a genetic network. The average fitness for the different environmental distance and variant networks were lower compared to without migration. For a population that is starting as heterozygous, it evolves quickly at the start but gene flow causes large deviations away from the average fitness, dropping to almost half the average fitness. Constant fluctuations, dropping below 0.50 highlights how distant the trait values are in the Cauchy distribution. As mentioned before, it is a narrower distribution in the centre so deviations away from it can lower the fitness value by a significant amount.

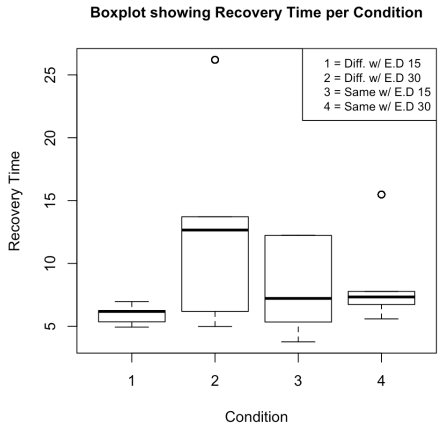


Figure 5: Boxplot showing the recovery times of the two variant migrant genetic regulatory makeups, evolving to a trait value of 65 or 80. Two variants of the genetic regulation where one variant is the same as the main population, thus $gene_1$ positively regulates $gene_2$ and $gene_3$, and $gene_2$ negatively regulates $gene_1$, and the other variant is $gene_2$ positively regulating the other genes, while $gene_1$ negatively regulates $gene_2$. Average recovery time different variant evolving to a trait value of 65 is 5.94 generations (standard deviation 0.72 generations), and for the other evolving to 80, mean of 12.75 generations with standard deviation of 7.57 generations. For the same migrant regulatory network, recovery time average of 8.16 generations (standard deviation of 3.51 generations) for environmental distance of 15 and mean of 8.58 generations with standard deviation 3.54 generations for environmental distance 30.

After migration, the network shows the ability to recover and stabilise very quickly. Populations that evolved slowly at the start are likely populations starting with homozygous individuals. Rarely did the populations reached a fitness value of 1.00 however showed the ability to increase over time. The variation between the different genetic structures and environmental distances was quite high which is to be expected from the visualised deviations. As expected, average fitness is lower compared to without migration.

Although no statistical test was conducted to see effects of environmental distance and or genetic variant on recovery time, boxplot highlights the possibility alleles were able to persist longer, possible because they are beneficial, showing that developmental pattern could be context-dependent. The network recovery time for the different migrant regulatory network evolving to a trait value of 80 (environmental distance 30) did have the highest average of 12.75 generations but a standard deviation of 7.57, which could possibly be due to homogenous scenarios and slower evolution. For the same autoregulatory network motif, when the migrant population is evolving to a trait value of 15 (environmental distance of 65), it took an average of 8.16(standard deviation 3.51) for the fitness value to equal or rise above the average fitness. The migrant network that evolved to trait value of 80 also showed consistent behaviours as it took a similar 8.58 generation recovery average, fluctuating between 3.54 generations.

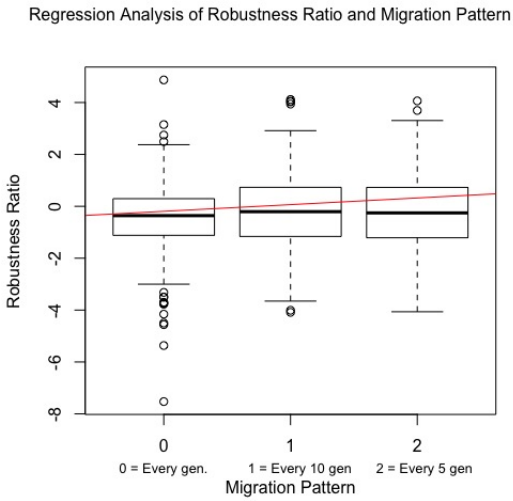


Figure 6: Graph showing the plots of the regression analysis of the log of the variance ratio and migration pattern and genetic makeup of both populations. Analysis of variance results *(see Appendix C)* show statistical significance of migration pattern on the log of the ratio. However, post hoc Tukey Tukey test *(see Appendix C)* shows no statistical significance with a particular migration pattern and robustness ratio.

6.4 Robustness

Multiple linear regression was done to see if any of the independent variables had a significant effect on the log robustness ratio. The data was log transformed to make it less skewed and linearise it. There were 4 factors: main population starting genetic makeup, migrant population starting genetic makeup, migration rate and migration pattern which were considered, as well as the interaction of starting genetic makeup of both populations (homozygous or heterozygous), and the interaction of migration rate and pattern. Migration rates were 1%, 3% and 5%, while migration patterns were every generation, every 5 generations and every 10 generations. Analysis of variance result *(see Appendix C)* reveals that the log ratio is statistically different for the migration patterns compared to the other factors. Linear regression computed an F-statistic of 3.25 with a corresponding p-values of 0.04, below the αlevel of 0.05. Before doing a post hoc Tukey test, Bartlett test was done to see homogeneity of variances (p-value of 0.32) and normal quantile plots created to see if data distribution was relatively normal *(see Appendix C)*. Ends of the normal QQ-plot deviated a bit suggesting a symmetric distribution that is fat-tailed. This is to be expected as due to the range of values present when migration is occurring. Post hoc Tukey analysis *(see Appendix C)* revealed no significant effect of any particular migration pattern however that could be due to small sample sizes.

7 Discussion

7.1 No Migration

Previous studies showed how a reproductively isolated population would evolve rapidly (García-Ramos & Kirkpatrick 1997). Without gene flow, populations did evolve rapidly to their environment due to a balance between selection and drift (García-Ramos & Kirkpatrick 1997, Tigano & Friesen 2016, Barber 1999). Especially for a heterogenous genetic network, which would evolve a lot quicker which was the case. Compared to a population that started as homogenous,

heterozygous individuals had variation in allele values at the start allowing for better combinations of values for selection to act on. This allowed such values to propagate and spread in the population. Swift rise (within 100 generations) can be attributed to the fitness distribution and the non-limitations in reproduction of the simulation. Following the fitness distribution, the most fit individuals per generation were likely to have been selected for reproduction many times allowing their combination of alleles to be passed on at a high frequency. Thus, the slopes of the heterogenous plots in *figure 3* are steep, suggesting selection was strongly acting on the population. Deviations seen are likely due to mutation and recombination events that removed favourable allele values and combinations.

In the homogenous case, the instances where fitness rose close to 1 is likely due to chance beneficial mutations and recombination to create better combination of allele values, and epistasis which drive the adaptiveness (Tigano & Friesen 2016). A beneficial mutation in $gene_1$ for example can have positively consequences for the values in $gene_2$ and $gene_3$. With selection and non-limiting reproduction acting on the system, it allows for such values quickly spread in short time, raising the fitness of the population. Additionally, *figure 2* shows the magnitude in difference of deviating towards from the median of the distribution. For the desired trait value of 50 with a standard deviation of 8, the maximum probability an individual has of being chosen is 3.98%. Although this is a small value, an individual that is 10 trait values away will have a probability of 1.55%, a three-fold difference between them. So, selection and limitless parental contribution allows even a small number of fitter individuals to be parents many times and pass on their values.

7.2 Migration and Robustness

To begin with genetic regulatory pathway of a population is seen to effect recovery time. Boxplot in *figure 5* shows that recovery time for a migrant population with a different regulatory network, one that evolves to a trait value of 65, and same regulatory system evolving to trait value 80 are quickly diminished from the population. However, a different migratory network that evolves to 80 or the same migratory network evolving to 65 persists in the population longer. Other than beneficial mutation or recombination events, there could be a possibility that gene flow helped to improve fitness. That patterns caused by evolutionary forces could be context-dependent. For a homogenous population that evolves slowly in, gene flow could lead to variant alleles closer to the peak of the distribution, hence a greater number of rising peaks in *figure 4b and 4c*. Further runs and tests must be conducted to statistically observe if gene flow has a beneficial effect depending on environmental distance and genetic makeup.

For heterogenous populations, *figure 4* shows how gene flow has an expected negative effect on fitness. It leads to large deviations in the fitness. Poorly adaptive traits to the environment would delay the system ability to evolve to local optimum as the values deviate far from it (García-Ramos & Kirkpatrick 1997). This was to be expected as periods where the foreign alleles far from the peak of the distribution. trait value and fitness to deviate from the local optimum.

Analysis of variance test revealed statistical significance of migration pattern on the log ratio of robustness. Although post-hoc Tukey test revealed no significance within the combinations of migration pattern, this is likely due to small sample sizes. Further testing will likely reveal

significant patterns between migration rate and canalisation of a network. There was still great insight on how the genetic network was able to overcome disruptions caused by gene flow and prevent homogenizing the populations and resisting perturbations. While gene flow caused large deviations in the fitness (as seen in *figure 4*), selection counteracts gene flow and maintains the adaptiveness of the network (Feder, Egan & Nosil 2012, Burt 1995). Moreover, I could investigate the patterns at which migration, selection and drift affect population growth, analysing effect of dominance on robustness (Rice & Papadopoulos 2009, Otto & Bourguet 1999). When migrant alleles enter the population, recombination gives them the opportunity to persist (Feder et al. 2012). Not just migrant allele values, but also complementary combination of migrant alleles that could allow it to persist (Feder et al. 2012). Yet despite the presence of migrant alleles, divergent selection heavily favours local alleles in particular and is quick to remove them (Tigano & Friesen 2016). Selection favouring locally fit individuals with better combination of alleles. *Feder (2012)* study highlighted a friction that exists between divergent selection, and gene flow and recombination. Results in this study show how when selection force is greater, it acts to swiftly remove migrant alleles in the population.

7.3 Shortcomings and Future Work

Due to time constraints and difficult circumstances, major assumptions were made in the model and simulation. For example, the fitness distribution in *figure 2* shows how variation was quickly eliminated from the model due to the narrow shape of centre of the distribution, allowing for selection to act strongly on the network. Few trials as well of the different conditions and design of the simulation prevented significant results from appearing. Realistically, the model did not take into account spatial or temporal aspects. Species are distributed along a habitat and the environment itself can change due to perturbations i.e. vicariant events (García-Ramos & Kirkpatrick 1997). With more time, I would have designed a more realistic model, taking into account dispersal habits, species distribution, environment perturbations, and migration limits. Species dispersal along with varying environmental conditions would have created a more complex system and delayed the quick homogenization of the network to allow for investigation of gene flow effects on fitness (García-Ramos & Kirkpatrick 1997, Barber 1999, Sato et al. 2006). This also would better maintain the variation in the population to better analyse the impacts. This would prevent the population being completely replaced in the current model even with higher migration rates. With regards to environmental spatial dynamics, I would have made it so that different areas would have different local optimums.

The current simulation assumed that there is no limit to parental contribution. Since there was no limit, the fitter individuals quickly pass on their trait values in the simulation. This rapidly canalised the system and decreased the variation in the environment. To improve this, I would limit parental contribution i.e. individuals reproducing can only be parents to five next generation individuals. This would prevent the system from stabilising so rapidly and allowing for more diverse alleles to continue to propagate in the population. There were no ecological barrier restricting migration and individuals were always replaced by migrants. Random environmental perturbations that would alter local optimums would further delay network evolution allowing for longer periods of variance. Thus, the effects of selection would not so quickly evolve the

system and can better investigate how network responds to disturbances. This would create more varying network structures and allow further investigation into network development with varying motifs.

Overall, further runs must be conducted to see effects of migration pattern on robustness. Although the post-hoc Tukey test showed no significance, it likely due to small sample size of runs per starting condition *(see Appendix A)*. In addition, another aspect to research is the possible benefits of gene flow on an evolving network. More runs can give statistical insight as to whether this is the case, by tracking lineages and allele values. Moreover, I can also consider dominance and investigate how the migration, selection and drift affect population growth, analysing effect of dominance on robustness (Rice & Papadopoulos 2009, Otto & Bourguet 1999). Lastly, it would also be interesting to investigate the opposite case as to how migration patterns are influenced by these evolutionary forces (W. Morris et al. 2004).

8 Concluding remarks and looking forward

In just a few centuries, anthropogenic changes largely due to societal development has disrupted many habitats and altered the evolutionary path of many species, all of which is likely to have long-term ecological consequences. This study aimed to investigate the scenario where gene flow allows related species, that could have originally been reproductively isolated, to interbreed, and the long-term impacts on species morphology and development. This can give insight into possible patterns that could emerge with future novel interactions between species to understand how development of both phenotype and genotype are altered. Although results did reveal an effect of migratory patterns on the evolution of a species network, further studies must be conducted to truly observe the interactions of environmental forces and the interplay between molecular genetic mechanisms and macro evolutionary processes. Improvement to the current simulation design and repeated trials can contribute to such studies.

9 Data Code and Availability

All code can be found in the GitHub Repository:
https://github.com/matthewcampos/CMEECourseWork.git

References

Barber, P. H. (1999), 'Patterns of gene flow and population genetic structure in the canyon treefrog, hyla arenicolor (cope)', *Molecular Ecology* **8**(4), 563–576.

Bourguet, D. (1999), 'The evolution of dominance', *Heredity* **83**(1), 1–4.

Brown, J. S. & Pavlovic, N. B. (1992), 'Evolution in heterogeneous environments: effects of migration on habitat specialization', *Evolutionary Ecology* **6**(5), 360–382.

Brown, T. A. (2002), How genomes evolve, *in* 'Genomes. 2nd edition', Wiley-Liss.

Burt, A. (1995), 'The evolution of fitness', *Evolution* **49**(1), 1–8.

Chandrasekaran, C. & Betrán, E. (2008), 'Origins of new genes and pseudogenes', *Nature Education* **1**(1), 181.

Coyne, J. A. (1992), 'Genetics and speciation', *Nature* **355**(6360), 511–515.

Feder, J. L., Egan, S. P. & Nosil, P. (2012), 'The genomics of speciation-with-gene-flow', *Trends in genetics* **28**(7), 342–350.

García-Ramos, G. & Kirkpatrick, M. (1997), 'Genetic models of adaptation and gene flow in peripheral populations', *Evolution* **51**(1), 21–28.

Gjuvsland, A. B., Hayes, B. J., Omholt, S. W. & Carlborg, Ö. (2007), 'Statistical epistasis is a generic feature of gene regulatory networks', *Genetics* **175**(1), 411–420.

Gjuvsland, A. B., Plahte, E. & Omholt, S. W. (2007), 'Threshold-dominated regulation hides genetic variation in gene expression networks', *BMC Systems Biology* **1**(1), 57.

Hinman, V. F., Yankura, K. A. & McCauley, B. S. (2009), 'Evolution of gene regulatory network architectures: examples of subcircuit conservation and plasticity between classes of echinoderms', *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1789**(4), 326–332.

Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. & Hartl, D. L. (2007), 'Genetic properties influencing the evolvability of gene expression', *Science* **317**(5834), 118–121.

Lynch, M. (2007), 'The evolution of genetic networks by non-adaptive processes', *Nature Reviews Genetics* **8**(10), 803–813.

Ohno, S. (1999), Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999, *in* 'Seminars in cell & developmental biology', Vol. 10, Elsevier, pp. 517–522.

Omholt, S. W., Plahte, E., Øyehaug, L. & Xiang, K. (2000), 'Gene regulatory networks generating the phenomena of additivity, dominance and epistasis', *Genetics* **155**(2), 969–980.

Orr, H. A. (1998), 'The population genetics of adaptation: the distribution of factors fixed during adaptive evolution', *Evolution* **52**(4), 935–949.

Orr, H. A. (2001), 'The genetics of species differences', *Trends in ecology & evolution* **16**(7), 343–350.

Otto, S. P. & Bourguet, D. (1999), 'Balanced polymorphisms and the evolution of dominance', *The American Naturalist* **153**(6), 561–574.

Pongratz, N., Gerace, L. & Michiels, N. K. (2002), 'Genetic differentiation within and between populations of a hermaphroditic freshwater planarian', *Heredity* **89**(1), 64–69.

Proulx, S. R. & Phillips, P. C. (2005), 'The opportunity for canalization and the evolution of genetic networks', *The American Naturalist* **165**(2), 147–162.

Rice, G. & Rebeiz, M. (2019), 'Evolution: How many phenotypes do regulatory mutations affect?', *Current Biology* **29**(1), R21–R23.

Rice, S. H. & Papadopoulos, A. (2009), 'Evolution with stochastic fitness and stochastic migration', *PloS one* **4**(10), e7130.

Sato, T., Isagi, Y., Sakio, H., Osumi, K. & Goto, S. (2006), 'Effect of gene flow on spatial genetic structure in the riparian canopy tree cercidiphyllum japonicum revealed by microsatellite analysis', *Heredity* **96**(1), 79–84.

Stern, D. L. & Orgogozo, V. (2009), 'Is genetic evolution predictable?', *Science* **323**(5915), 746–751.

Tigano, A. & Friesen, V. L. (2016), 'Genomics of local adaptation with gene flow', *Molecular ecology* **25**(10), 2144–2164.

W. Morris, D., E. Diffendorfer, J. & Lundberg, P. (2004), 'Dispersal among habitats varying in fitness: reciprocating migration through ideal habitat selection', *Oikos* **107**(3), 559–575.

Yeaman, S. & Whitlock, M. C. (2011), 'The genetic architecture of adaptation under migration–selection balance', *Evolution: International Journal of Organic Evolution* **65**(7), 1897–1911.

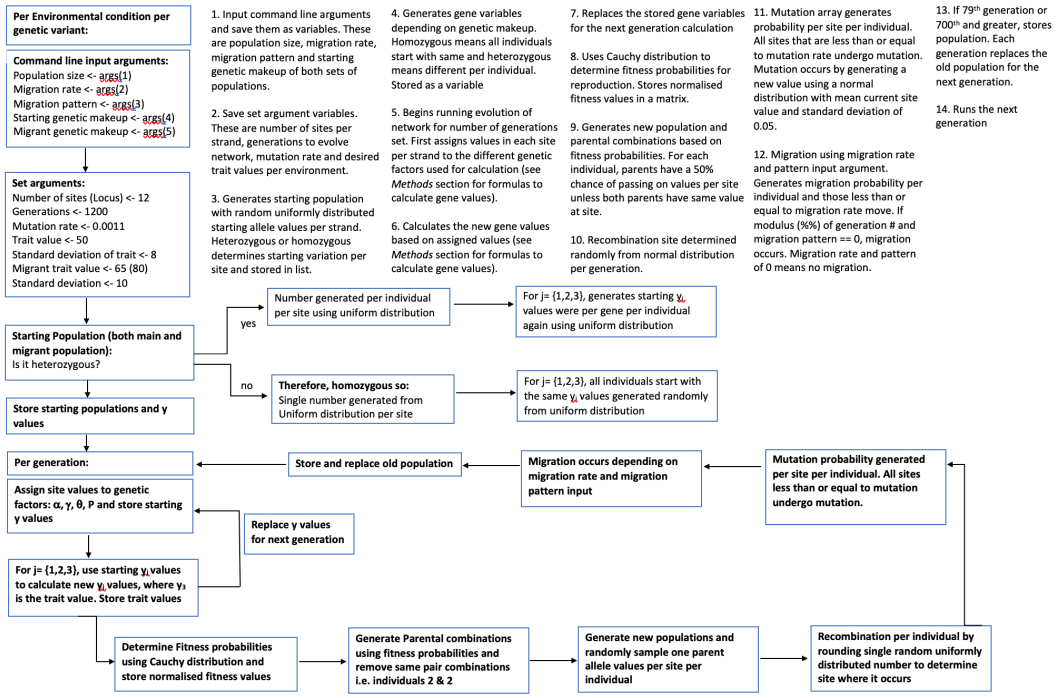
Appendix A Permuations of Conditions

Description: The different combinations of conditions used for data sorting and simulations. Data folders are permutations of Genetic structure and Environmental distance i.e. Same and 30 (Main population 50 and Migrant population 80). Within each folder are 44 combinations of starting genetic makeup of populations, migration rate and migration pattern for the simulations. For example, within Same and 30 is: homogenous - heterogenous, 3%, every generation.

<table>
<tr><th colspan="2">GENETIC STRUCTURE</th><th>MIGRATION RATE (%)</th><th>MIGRATION PATTERN</th><th colspan="2">ENVIRONMENTAL DISTANCE (TRAIT VALUE)</th></tr>
<tr><td colspan="2">SAME</td><td rowspan="3">0</td><td rowspan="3">Every generation</td><td>MAIN</td><td>50</td></tr>
<tr><td>MAIN</td><td>MIGRANT</td><td rowspan="4"></td><td rowspan="4"></td></tr>
<tr><td>Homogenous</td><td>Homogenous</td></tr>
<tr><td>Heterogenous</td><td>Heterogenous</td><td rowspan="3">1</td><td rowspan="3">Every 5 generations</td></tr>
<tr><td>Homogenous</td><td>Heterogenous</td><td rowspan="4">MIGRANT</td><td rowspan="4">65
80</td></tr>
<tr><td>Heterogenous</td><td>Homogenous</td></tr>
<tr><td colspan="2">DIFFERENT</td><td rowspan="3">3</td><td rowspan="3">Every 10 generations</td></tr>
<tr><td>MAIN</td><td>MIGRANT</td></tr>
<tr><td>Homogenous</td><td>Homogenous</td></tr>
<tr><td>Heterogenous</td><td>Heterogenous</td><td rowspan="3">5</td><td rowspan="3">Random</td></tr>
<tr><td>Homogenous</td><td>Heterogenous</td></tr>
<tr><td>Heterogenous</td><td>Homogenous</td></tr>
</table>

Appendix B Computer Program Workflow

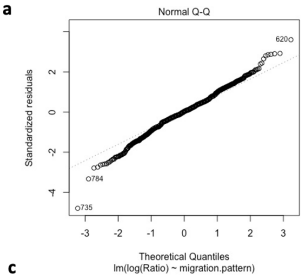
Description: Workflow of R script of simulation program and the logic behind the design of the simulation.



Appendix C Tests for One-way Analysis of Variance on Regression of the log Robustness Ratio and Migration Pattern

Description: Multiple regression was conducted to see if any of the independent variables (and interactions among them) had a statistically significant effect on the log Robustness ratio. Graph (a) is a normal Q-Q plot to check if the distribution of the data was normal, with no large outliers. Distribution shows that it is a normal around the centre of the distribution however deviates a little at the ends, suggesting it is fat-tailed. This is to be expected with migrant allele values that are evolving far from the trait value. (b) is a Bartlett test conducted to see if there is homogeneity in the variances. A p-value of 0.32 so variances are equal. After conducting analysis of variance test of log Robustness ratio and Migration pattern, (c) is the post hoc Tukey test which shows no statistical significance of any particular migration pattern and the log Robust ratio.

a



b

statistic.Bartlett's K−squared	2.29474186452542
parameter.df	2
p.value	0.317470324149718
data.name	log(Ratio) by migration.pattern
method	Bartlett test of homogeneity of variances

c

migration.pattern	diff	lwr	upr	p adj
1-0	0.25689994	-0.03959993	0.5533998	0.1046874
2-0	0.22657133	-0.06992854	0.5230712	0.1721194
2-1	-0.03032861	-0.34730030	0.2866431	0.9725580