Effect of population structure on neutral genetic variation and barriers to gene exchange

by

Parvathy Surendranadh

November, 2024

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
Doctor of Philosophy

Committee in charge:

Carrie Bernecky, Chair
Nicholas Barton
Beatriz Vicoso
Peter L. Ralph

The thesis of Parvathy Surendranadh, titled *Effect of population structure on neutral genetic variation and barriers to gene exchange*, is approved by:

Supervisor: Prof. Nicholas Barton, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Prof. Beatriz Vicoso, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Prof. Peter L. Ralph, University of Oregon, Eugene, USA

Signature: _____

Defense Chair: Prof. Carrie Bernecky, ISTA, Klosterneuburg, Austria

Signature: _____

Signed page is on file

ISTA Thesis, ISSN: 2663-337X

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 dissertation.

I accept full responsibility for the content and factual accuracy of this work, including the data and their analysis and presentation, and the text and citation of other work.

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: _____

Parvathy Surendranadh
November, 2024

Signed page is on file

Abstract

Understanding the role of evolutionary processes in shaping genetic variation has been a primary goal in evolutionary genetics. In this regard, a key question is how genetically distinct populations evolve in the face of gene flow, thereby generating genetic and phenotypic divergence and reproductive isolation (RI). This requires quantifying the role and relative contributions of prezygotic and postzygotic isolating mechanisms on the reduction of gene exchange between populations, and identifying regions in the genome that mediate RI, which is often polygenic. Further, this needs distinguishing neutral and selected regions in the genome, and discerning how selection influences patterns of neutral divergence.

Population structure, defined as any deviation from panmixia, such as geographic distribution, movement and mating patterns of individuals, influences how genetic variation is structured in space and shapes the neutral null model. Availability of large scale spatial genomic datasets now enables us to detect signatures of population structure in genetic data and infer population genetic parameters. Such inferences are crucial and have wide applications in biodiversity, conservation genetics, population management and medical genetics. However, inferences are based on assumptions that do not always match the complex reality, thus leading to erroneous conclusions. Moreover, the role and interaction of heterogeneous population density and dispersal, which are ubiquitous in nature, has been challenging to study owing to their mathematical complexity. In such scenarios, feedback between theory, data and simulations can prove to be useful.

In this thesis, I examine the effect of population structure on neutral genetic variation and barriers to gene exchange in hybridising populations, thereby bridging together the fields of spatial population genetics and speciation.

Despite being a key concept in speciation, reproductive isolation (RI) lacks a quantitative definition and has been used and measured differently across different fields. Chapter 2 gives a quantitative definition of RI, in terms of the effect of genetic differences on gene flow. We give analytical predictions for RI in a range of scenarios, in terms of effective

migration rates for discrete populations and barrier strength for continuous populations. In addition to this, we discuss current measures of RI and their limitations, and propose the need for new measures that combine organismal and genetic perspectives of RI.

In chapter 3, I examine the combined effect of assortative mating, sexual selection and viability selection on RI. For this, we consider a polygenic 'magic' trait under a mainland-island model. We obtain novel theoretical predictions for molecular divergence in terms of effective migration rates, which bears a simple relationship to measurable fitness components of migrants and various early generation hybrids. We explore the conditions under which local adaptation can be maintained despite maladaptive gene flow and quantify the relative contributions of viability and sexual selection to genome-wide barriers to gene flow.

The next two chapters of the thesis focus on a hybrid zone of *Antirrhinum majus* that consist of two subspecies- the magenta flowered *A. m. pseudomajus* and the yellow flowered *A.m. striatum*. Previous studies have suggested that flower colour is target of pollinator mediated selection and is influenced only by few genes. While these regions show high genetic differentiation between the subspecies, the rest of the genome is seen to be well mixed. Chapter 4 examines the effects of heterogeneous population density and leptokurtic dispersal on isolation by distance and the distribution of heterozygosity by focusing on non-flower colour markers.

Chapter 5 analyses cline shapes and associations among 6 focal flower colour markers to understand how selection and dispersal maintain this hybrid zone. We see sharp coincident stepped clines at all loci and positive associations throughout the hybrid zone, contrary to the expected patterns from diffusive gene flow. With a novel scheme of inferring dispersal combined with multilocus simulations, we show that stepped clines do not reflect genetic barriers to gene flow, but are rather a result of long-distance migration. This framework allows us to get realistic estimates gene flow and selection and shows how traditional cline analysis may lead to inaccurate conclusions when assumptions of the theory are not met.

Overall, this thesis investigates how different features of population structure leave detectable signatures in genetic variation, namely in patterns of isolation by distance, linkage disequilibrium and genetic divergence. It also highlights how effective migration rates provide useful way of analysing polygenic architectures and shed new light into hybrid zones. In doing so, I identify scenarios when simple models become insufficient and suggest possibe directions by combining genetic data with simulations.

Acknowledgements

First and foremost, I am grateful to Nick for being my supervisor. Your passion, dedication and depth of knowledge have inspired me tremendously and I believe that I have grown as a researcher over the years through all our scientific discussions. More importantly, I cannot thank you enough for being kind, understanding, and supportive, especially during the tough times and when I lacked confidence.

I extend my deepest gratitude to my institute ISTA, to the members of the graduate school, administrative staff, HR department, the grant office, IT department and to Vlad for making the technical and bureaucratic aspects smooth.

I also acknowledge the funding agencies Marie Curie COFUND Doctoral Fellowship, Austrian Science Fund FWF (grant P32166) and ERC (grant PR1000ERC02) for financially supporting my research over the years.

I am also grateful to my committee members Beatriz and Peter for being in all the progress reviews, giving me valuable feedback and encouragement.

I am also very thankful to all the 'Bartonoids' for making the group a happy space and for gifting some great memories. Funmi, I have really enjoyed our time together at ISTA and thank you for being a great friend through all the happy and tough times (and particular for listening to all my rants). Himani, from being my first rotation project co-supervisor to a close friend, I am really grateful for your friendship, for the meetups and for parathas. I have learnt a lot from you both scientifically and personally. Sean, thank you for your unwavering support and motivation and for always rooting for me. I also thank Louise, Sofia, Hila, Gemma, Laura and Rohit for the fun meetups and coffee sessions.

I am also deeply indebted to my family for all their support and encouragement. Amma and Achan and Munni, you have been my constant cheerleaders and I would not have been here today if not for you. Though you didn't fully understand the ups and downs of the PhD life, thanks for being there for me all throughout. Paapu, thank you for being my de-stressor and source of happiness. I also thank my partner Ate for being my strongest

support system and a great listener. Thanks for believing in me and lifting me up every single time.

This 6-year long journey would not have been possible without my second family. PP, Chill, Pakki, Thaps, Mus and Divs, thank you for the fun filled sleepless nights, the movie and food tours, the stupid games, accepting the child in me and for making Vienna home. Special thanks to 'blum' for being my core and strength, and pp, for 'Pochlarn'. I also thank Harithech, Resh, Sweetu, Nayana, Miki, Shino, Ammoni, Adi, Tobs, Cath and Helena for always being a call away. Your friendship kept me going when things looked down.

And finally, to Vienna for the amazing experience and making me grow into a better human being.

About the Author

Parvathy Surendranadh completed BS-MS dual degree in Mathematics from the Indian Institute of Science Education and Research (IISER) Mohali before joining ISTA in September 2018. Her research interests include the evolution of genetic variation in structured populations, with a focus on speciation. She is also keen on developing new approaches to better understand data from natural populations and making realistic inferences. While at ISTA, she primarily worked on the hybrid zone of snapdragons and was actively involved in the annual fieldwork in the Spanish Pyrenees. During her PhD, she worked on several collaborative projects one of which has been published in Genetics, and a target review published in the Journal of Evolutionary Biology. She has also presented her research at several conferences like PopGroup, ESEB, and GRS and was invited to a department seminar in Lille. She also co-supervised rotation students and one master student in the Barton Group. Outside of science, she is a dancer and enjoys travelling and reading.

List of Collaborators and Publications

Chapter 2 (Published)

Title: What is reproductive isolation?

Journal: *Journal of Evolutionary Biology*

url: https://doi.org/10.1111/jeb.14005

Westram, A. M., Stankowski, S., **Surendranadh, P.**, & Barton, N. (2022). What is reproductive isolation?. *Journal of Evolutionary Biology*, 35(9), 1143-1164.

Associated commentary: Westram, A. M., Stankowski, S., **Surendranadh, P.**, & Barton, N. H. (2022). Reproductive isolation, speciation, and the value of disagreement: A reply to the commentaries on 'What is reproductive isolation?'. *Journal of Evolutionary Biology*, 35(9), 1200-1205.
https://doi.org/10.1111/jeb.14082

Author contributions

Westram, A: Conceiving the idea, conceptualization of the study, writing the manuscript

Stankowski, S: Conceiving the idea, conceptualization of the study, visualization, writing the manuscript

Surendranadh, P: Conceptualization of the study, performing simulation, visualisation, writing the manuscript (SM).

Barton, N: Conceptualization of the study, analytical results, writing the manuscript

Chapter 3 (Submitted)

Title: Effect of assortative mating and sexual selection on polygenic barriers to gene flow

Surendranadh, P., & Sachdeva, H. (2024). Effect of assortative mating and sexual selection on polygenic barriers to gene flow. *bioRxiv*, 2024-07.

url: https://biorxiv.org/cgi/content/short/2024.07.30.605898v1

Author contributions

Surendranadh, P: Designing the study, mathematical analysis, simulations, visualization, writing the article

Sachdeva, H: Designing the study, mathematical analysis, writing the article

Chapter 4 (Published)

Title: Effects of fine-scale population structure on the distribution of heterozygosity in a long-term study of *Antirrhinum majus*

Journal: *Genetics*

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.

url: https://doi.org/10.1093/genetics/iyac083

Author contributions

Surendranadh, P: Designing the study, data analysis and simulations, visualization, writing the manuscript

Arathoon, L: Designing the study, data analysis, visualization, writing the manuscript

Baskett, C: Writing the manuscript

Field, D: Generating SNP panel the study, feedback

Pickup, M: Conceiving the idea, feedback

Barton, N: Designing the study, data analysis, guidance and feedback

[0]* These authors contributed equally to this work.

Chapter 5 (In preparation)

Title: Genetic analysis of clines in an *Antirrhinum majus* hybrid zone

Author contributions

Surendranadh, P: Designing the study, data analysis and simulations, visualization, writing the article

Stankowski, S: Writing the article, visualization

Field, D: Generating the SNP panel for the study

Barton, N: Designing the study, data analysis and inference, writing the article, guidance and feedback

Table of Contents

Chapter 1

General Introduction

1.1 Motivation

Genetic diversity in populations is shaped by the interplay of evolutionary processes such as selection, mutation, genetic drift and migration. Moreover, individuals mostly occupy habitats that are spatially continuous and heterogeneous. Population structure, defined as any deviation from panmixia, includes the distribution, movement and mating patterns of individuals in space. This plays a central role in shaping the geographic distribution of genetic variation by influencing patterns of migration and drift, and defines the neutral null model. Ecological and evolutionary processes interact to shape genetic variation at loci influencing fitness differences between or within populations, thereby influencing patterns of adaptation and speciation. Moreover, strong genome-wide selection influences neutral divergence due to the pervasive effects of selection on neutral loci (Barton et al., 2013; Allman and Weissman, 2018; Bierne et al., 2013). Thus, a long-standing question in evolutionary biology is understanding how population structure contributes to the patterns of genetic variation we see in nature.

A key question that has been puzzling researchers, specifically in the field of speciation, is how genetic variation across populations evolves in the face of gene flow (Mayr, 1942; Dobzhansky, 1937; Coyne and Orr, 2004). Understanding how speciation occurs with gene flow despite homogenizing genetic differences between populations, has attracted much interest and has been debated widely over the last 50 years (Dopman et al., 2024; Smadja and Butlin, 2011). Divergent selection between populations can be driven by environmental differences, habitat choice, genetic incompatibilities, sexual selection, assortative mating, etc. which can act either independently or together (and may themselves structure populations). These processes reduce genetic exchange and generate both genetic and phenotypic divergence, and thereby reproductive isolation (RI) between populations. Thus, analysing the role and relevance of multiple reproductive barriers to gene flow

has been a key focus among speciation researchers. In this context, assortative mating and sexual selection are of particular interest due to their prevalence in nature. Despite being widely studied, the role of sexual selection in speciation with gene flow remains unsolved; does sexual selection promote/ or hinder speciation? Does sexual selection work synergistically with ecological selection to impede gene flow? (Kirkpatrick and Ravigné, 2002; Rundle and Nosil, 2005; Maan and Seehausen, 2011; Servedio and Kopp, 2012). To answer these questions, we need to be able to quantify the effects of assortative mating and sexual selection on RI, accounting for the genetic architecture of traits. However, RI has lacked a quantitative definition and has been used and measured differently across the field (Stankowski and Ravinet (2021), see chapter 2). With better sequencing technologies and whole genome datasets, we can in principle measure RI from genetic data. However, this requires distinguishing neutral vs selected regions in the genome and discerning how selection influences patterns of neutral divergence along the genome (Feder et al., 2012).

Geography also plays a key central role in shaping patterns of genetic variation, but quantifying the impact of heterogeneous population structure has been quite challenging both theoretically and empirically (Hey and Machado, 2003; Bradburd and Ralph, 2019). With the abundance of genetic datasets, the field of evolutionary biology has advanced with the development of new inference methods, making it feasible to better estimate population genetic parameters such as selection, gene flow, effective population size, etc from features of genetic variation. Such inferences are crucial and have wide applications in biodiversity, conservation genetics and population management. Nevertheless, we often use fairly simple models of population structure whose assumptions are not met in reality (Guillot et al., 2009). Furthermore, the two fields of spatial population genetics and speciation evolved rather independently until recently: inferences in spatial population genetics mainly assume neutrality while speciation studies often ignore the effects of realistic density and dispersal, except in the study of hybrid zones. Even most hybrid zone studies infer selection and gene flow from fairly simple cline models that assume uniform density and diffusive gene flow, which might not hold in nature. With powerful computational capacity and large-scale spatial genomic datasets, the field can progress further by developing new theory that can be tested against existing methods and genetic datasets. Simulations can further inform us on expected patterns in complex scenarios, paving the way for new theory and methods. This constant feedback is not only powerful but unavoidable: we can ask whether complex models are more efficient and whether we can detect signatures of population structure from genetic data, developing better summary statistics accounting for these effects.

In this thesis, I seek to fill these gaps by studying how population structure shape genetic variation at neutral and selected loci. Specifically, I first look at the combined effects of assortative mating, sexual selection and viability selection on genetic barriers to gene flow,

considering a polygenic trait. While a lot of commonly studied traits are mediated through large numbers of loci, their effect has been largely ignored in theoretical developments in speciation, due to their mathematical complexity. Next, I shift my focus to hybrid zones (of snapdragons), which being spatially explicit populations and 'natural crossing experiments' are perfect systems to study the effect of realistic population structure. Before giving an overview of the chapters, I will first briefly introduce these concepts and mention some of their limitations. Specifically, I will discuss below the models of population structure, and commonly used methods to infer demography from neutral genetic variation. I will also briefly introduce RI (with details in Ch.2) and hybrid zones.

1.2 Models of population structure

Population structure, defined as any deviation from panmixia, is ubiquitous in nature and is formed as an interplay between population density, dispersal and mating patterns. In spatially structured populations, dispersal and population density are typically non-symmetric. Dispersal, the displacement between parents and offspring, varies among species and landscapes. When migrants successfully mate to produce offspring, there is gene flow. Dispersal also directly affects population density, ρ (the number of individuals per unit length or area), and neighbourhood size (the area from which parents come) given by $2\sqrt{\pi}\sigma\rho$ in 1D and $4\pi\rho\sigma^2$ in 2D respectively (Wright, 1943). In turn, varying population density and variation in reproductive fitness amplify genetic drift and affects patterns of dispersal. Thus, in spatially structured populations, genetic variation is shaped by underlying demographic processes.

Theoretical understanding of the effects of population structure stems from various spatial genetic models. The most commonly used models of discrete population structure are the island (Wright, 1931) and stepping stone (Malécot, 1948). The most basic island model assumes demes of equal effective size N_e that exchange migrants randomly at an equal rate m (see figure 1.1A). The stepping stone model considers more realistic dispersal but still assumes demic structure with equal or varying deme size (figure 1.1B). Kimura and Weiss (1964) studied stepping stone models with nearest neighbour migration in one, two and three dimensions. They calculated an approximate formula for how correlation of allele frequencies decreases with distance at equilibrium. While these models assume populations to be divided into demes, models of continuous population structure were studied by Wright (1943); Malécot (1948); Maruyama (1972). They show patterns of isolation by distance, where individuals closer in space are more genetically related than individuals farther apart, i.e identity by descent decays with distance (figure 1.1C). These models give analytical predictions for how equilibrium probability of identity decreases with distance for homogeneous migration, assuming uniformly distributed populations

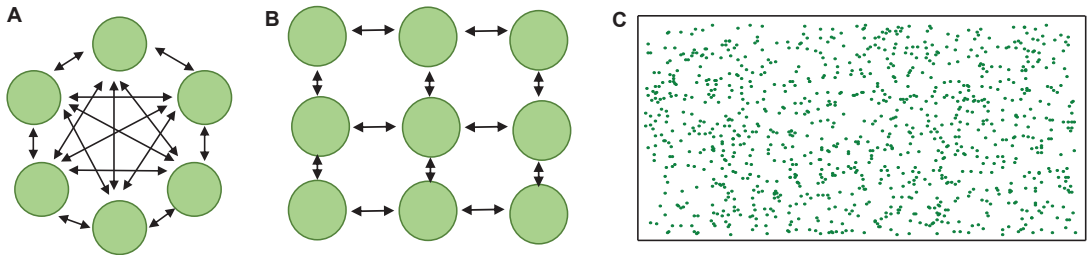


Figure 1.1: A: Island model B: 2D stepping stone model with nearest neighbour migration C: Individuals uniformly distributed in a continuous 2D habitat.

in 2D with constant population density and independent movement and reproduction of genes. However, Felsenstein (1975) found an inconsistency in the model, known as 'pain in the torus', which arises due to the lack of density dependent regulation which is inevitable in truly continuous populations. This leads to increased clumping of individuals with time, and consequently violates the assumption of uniform density. The Spatial-Lambda-Fleming-Viot (SLFV) model (Barton et al., 2010) addresses this issue by considering a model where all reproduction and dispersal happens via extinction/ recolonisation events (a fraction of individuals go extinct and are replaced by offspring of a small number of nearby parents every generation) generating correlated movements of genes. Isolation by distance models for the rate of decay of genetic similarity with distance have been extended to include temporal patterns (Duforet-Frebourg and Slatkin, 2016), physical barriers (Nagylaki, 1988) and non-Gaussian dispersal (Smith and Weissman, 2023). In reality, populations are neither assorted into demes nor have a uniform distribution, but rather have a patchy distribution in a continuous habitat. Broquet et al. (2013) examined this for an infinite island model where genetic patchiness arose from the combined effects of drift and dispersal. Moreover, in continuous populations patches can go extinct and get recolonised, further enhancing heterogeneity and generating correlated movements of individuals and thus relatedness. Thus, extending theory to account for heterogeneous density population structure together with population regulation are indispensable to understand genetic variation in spatially extended populations. This is also crucial not only in generating a baseline null model but also for developing inference methods and finding whether population structure leave detectable signatures in spatial genetic data.

1.3 Neutral demographic inference from genetic data

Inferring population structure from genetic variation has been a long-standing problem in evolutionary biology. Advances in genotyping technology now give us access to whole genome sequences, together with spatial information. Availability of such data led to the development of novel methods that aim to infer population structure from genetic variation. Alongside advances in population genetics, landscape genetics has developed

with a focus on demographic inference (Manel et al., 2003; Holderegger and Wagner, 2006; Storfer et al., 2007; Schoville et al., 2012). It aims to identify environmental factors that affect spatial genetic variation and gene flow among populations, thus bridging the gap between ecology and evolutionary biology. It also finds direct applications in conservation biology and management studies (by identifying populations that need rescue, devising management strategies, analyzing the spread of diseases, etc). With the rapid development of such inference methods, the field of spatial genetics has also expanded at pace (Guillot et al., 2009). Below, I briefly discuss some of the commonly used inference methods and their pitfalls.

The most widely used method dates back to Wright (1931), which uses F_{ST}, defined as the average probability of identity by descent within a subpopulation, F_S, relative to the total population F_T, $F_{ST} = \frac{F_S - F_T}{1 - F_T}$ to infer the number of migrants. However, this is based on the discrete island model where F_{ST} at equilibrium is given by $F_{ST} \sim \frac{1}{1+4N_e m}$. While F_{ST} itself is a good measure of genetic differentiation between populations, translating it into levels of gene flow is not appropriate since $N_e m$ gives the relative strength of migration and genetic drift, and not the actual gene flow, m (Whitlock and McCauley, 1999). Another popular approach utilizes isolation by distance to show that the slope of linear regression between pairwise genetic diversity estimates $F_{ST}/(1 - F_{ST})$ and pairwise distances at equilibrium gives a measure of neighbourhood size (Rousset, 1997) and has been been incorporated into software such as spagedi (Hardy and Vekemans, 2002) and GENPOP (Rousset, 2008). This method accurately estimates neighbourhood size for longer distances and is robust to different dispersal distributions in homogeneous populations. Robledo-Arnuncio and Rousset (2010) further extended these results for heterogeneous populations with fluctuating density.

Alongside isolation by distance models, models of admixture try to estimate proportions of shared ancestry between populations by utilising signals from linkage disequilibrium or correlations across loci (Sankararaman et al., 2008; Patterson et al., 2012; Green et al., 2010). Such models are widely used to determine admixture in humans (Ralph and Coop, 2013; Lohse and Frantz, 2014; Yair et al., 2021). Complementary to admixture models, Isolation with Migration (IM) models try to estimate the time since divergence, gene flow or effective population size under a model of population splitting with migration (Hey and Nielsen, 2004). However, both these models work only with discrete populations.

Unlike the methods mentioned above, which are based on population genetic models, several methods rely on descriptive statistics. Principal Component Analysis (PCA) (Menozzi et al., 1978), STRUCTURE (Pritchard et al., 2000) and ADMIXTURE (Falush et al., 2003) infer genetic structure by identifying clusters without utilizing spatial information. However, they fail to identify continuous populations with a spatially heterogeneous distribution (see Novembre and Stephens (2008) on PCA). A second

category of methods utilizes landscape information and models its effects. These include Mantel tests (Legendre and Fortin, 2010) that finds associations between two or more distance matrices to see if the geographic distribution affects the genetic distances, and isolation by resistance models (McRae, 2006) that predicts genetic structuring in complex landscapes while accounting for habitat heterogeneity. However, they produce an inaccurate null model when correlations are ignored (Guillot and Rousset, 2013; Lundgren and Ralph, 2019) and when gene flow is asymmetric respectively.

While the above methods have been useful in inferring demographic parameters from genetic data, they are built on strict assumptions which are often violated when applied to complex reality. The implications of gene flow and genetic structure are often misinterpreted and conclusions often make general inferences that are not warranted by the data and not validated by replicate studies (Richardson et al., 2016; Meirmans, 2012). Methods that rely on purely statistical models do not model the evolutionary processes that generate the genetic structure. Thus, interpreting the outcomes from inferences and understanding their biological relevance is challenging. Additionally, inaccurate sampling schemes and failure to include population structure could lead to falsely ascribing discrete structure when genetic variation is continuously distributed, leading to erroneous outcomes of population genetic summary statistics or in GWAS (Battey et al., 2020). Accurate inference requires spatially explicit null models that can be determined only from models of realistic spatial structure. Efficient use of simulations for modelling realistic demography (for example SLIM; Haller and Messer (2019)) or for validating inference methods (by comparison to known direct estimates) can help identify their utility and accuracy. Recent years saw more such studies combining simulations and statistical likelihood-based approaches together with population genetic models to infer demography from spatial genetic data (Bradburd et al., 2016, 2018; Ringbauer et al., 2017).

1.4 Reproductive isolation and barriers to gene flow

Speciation requires the evolution and maintenance of multiple barriers to gene flow, i.e., the evolution of traits or processes that reduce genetic exchange between hybridizing populations (Coyne and Orr, 2004; Maan and Seehausen, 2011; Kulmuni et al., 2020; Servedio et al., 2011). These include traits that are under divergent ecological selection or sexual selection, or genetic incompatibilities, that cause a reduction in fitness of hybrids and thereby generate postzygotic isolation, together with traits that underlie assortative mating, thus contributing to prezygotic isolation (Kirkpatrick and Ravigné, 2002). However, understanding the role of different processes in speciation requires quantifying their strengths, relative contributions and interactions, as well as the the genetic architecture of speciation traits and the genomic distribution of loci mediating

influencing them (Smadja and Butlin, 2011). While this is often described in terms of the build up of reproductive isolation (RI), the term lacks a clear quantitative definition and is often used differently across speciation researchers. Chapter 2 gives a detailed review of what we mean by reproductive isolation, defined in terms of reduction in gene exchange due to the effect of genetic differences between populations. We also provide analytical predictions for RI in specific scenarios of divergent selection, and discuss challenges to measuring it in practise.

While the role of divergent selection is well studied in speciation, disagreement centres around the role of sexual selection (and assortative mating) and how it interacts with natural selection (Safran et al., 2013; Servedio and Boughman, 2017; Kopp et al., 2018). Assortative mating occurs when there is positive correlation between phenotypes of mating pairs (Lewontin et al., 1968; Jiang et al., 2013). These can occur either as a by-product of spatial or temporal isolation, or due to mate choice. When assortative mating is based on phenotype matching, it can lead to sexual selection due to differential mating success of phenotypes. However, whether sexual selection promotes or hinders speciation depends on its interaction with natural selection and how they generate divergence between populations (Kirkpatrick and Nuismer, 2004; Servedio and Kopp, 2012). Specifically, this depends on how associations or linkage disequilibrium (LD) between traits influencing multiple reproductive barriers are generated and maintained across generations to produce a strong reduction in gene flow between populations (Felsenstein, 1981; Butlin and Smadja, 2018; Barton and De Cara, 2009; Dopman et al., 2024; Servedio, 2009).

Such descriptions of RI in terms of LD, often described as coupling, are useful because, when selection is mediated through large number of loci, divergence requires sets of locally favourable alleles to establish in the population despite maladaptive gene flow. This in turn depend on whether selected loci evolve independently under direct selection or whether changes in divergence depends on the effects of indirect selection due to LD between multiple loci. Thus, understanding how different processes favour speciation also depends on the ease with which associations can be maintained between different barrier traits: direct selection on polygenic traits, linkage between trait loci and pleiotropy (wherein the same loci influence multiple speciation traits) make it easier to maintain LD whereas indirect selection (i.e when a locus does not directly influence fitness, but is associated with traits that do so and thus experience selection) requires associations between sets of loci to be maintained despite recombination trying to break these favourable gene combinations. Moreover, strong LD between barrier loci can in turn generate associations between barrier and neutral loci (referred to as genome-wide LD), thereby strengthening neutral divergence between populations and generating a genome-wide reduction in gene flow. While LD plays a crucial role in speciation, analytical descriptions of LD are challenging as one must consider how many different multi-locus combinations co-evolve.

In this regard, a powerful way of describing allele frequency changes across multiple loci is to consider how introgressing alleles at any locus are transferred between different genetic backgrounds, defined as different combination of alleles at other trait loci. To establish in the recipient population, the allele must associate with the locally favourable genetic background through multiple recombination events over multiple generations. The rate at which an incoming allele gets transferred between alternate genetic backgrounds is described as the effective migration rate (Bengtsson, 1985). Strong LD between trait loci can thus reduce effective migration rate across the genome; this effect is stronger for tightly linked loci and weakest for unlinked loci. This reduction in effective migration rate between population increases genetic divergence between populations, thus strengthening RI. While the notion of m_e is valuable since it captures the effect of LD between a focal locus and all other loci with a single quantity and provides a way to quantify RI, it has been much less considered in theoretical studies of speciation. In this thesis, I demonstrate how m_e is a natural way of quantifying RI (see Ch. 2) and apply theory based on m_e to quantify RI in a scenario when both natural and sexual selection operate (see Ch. 3) and show its application in the context of hybrid zones (Ch. 5)

1.5 Hybrid zones

Hybrid zones are narrow regions where divergent populations meet, mate and produce recombinant individuals of shared ancestry (Endler, 1977). They are quite common in nature and are 'natural laboratories' that give important insights into the mechanisms of reproductive isolation, genetic basis of traits underlying adaptation and can be utilized to infer quantitative strength of the underlying processes that maintain them (Hewitt, 1988; Abbott et al., 2013). Hybrid zones are characterized by a set of clines - gradients in morphological, behavioural or cytological phenotypes or allele frequencies at one or more loci that mediate the traits differentiating the hybridising taxa. Most commonly studied hybrid zones are maintained by a balance between dispersal and some form of selection, either depending on how well individuals adapt to their local environment (exogeneous) or independent of the environment (endogenous): due to selection against hybrids, frequency dependent selection or genetic incompatibilities (Barton and Hewitt, 1985). In the former case, the cline is centred where the environment transitions, whereas the latter may move due to differences in individual fitness, asymmetric selection, dominance or due to differences in density and dispersal. However, such movement is rarely detected in nature as hybrid zones are expected to be trapped at local barriers due to heterogeneity of population structure (Barton and Hewitt, 1985; Barton, 1979a).

Early studies of clines and hybrid zones mathematically described the rate of change in allele frequency at a single selected locus (Haldane, 1948; Bazykin, 1969; Slatkin, 1973),

and showed that the expected cline shape is sigmoid; $p(x) = \frac{1}{1+\exp\left\{2\sqrt{s}(x+c)/\sigma\right\}}$, where $p(x)$ is the allele frequency at location x, s is the strength of selection against heterozygotes, c is the cline centre and sigma is standard deviation of parent-offspring dispersal distance per generation, along some axis.

This was further extended to multilocus clines, where more than one locus contributes to the differences between divergent populations (Slatkin, 1975; Barton, 1983). In such cases, associations between loci (even when unlinked), known as linkage disequilibria (LD), are generated since individuals carry with them gametes from their source population as they disperse (Barton and Gale, 1993; Li and Nei, 1974). For example, with two selected loci with allele frequencies p and u, linkage disequilibrium $D = \frac{\sigma^2}{r}\frac{\partial p}{\partial x}\frac{\partial u}{\partial x}$, where r is the rate of recombination between selected loci (Barton (1986), see details in Ch. 5). This increases the effective selection experienced by a focal locus, due to indirect selection from all other loci that it is associated with, which narrows the cline, causing a stepped cline shape; with long exponential tails on either side (Szymura and Barton, 1986; Barton, 1983).

Sharp clines at selected loci may also generate stepped clines at a linked neutral locus, with the strength of the step decreasing with increasing map distance to the selected locus (Barton, 1979b). Strong LD between neutral and selected loci can thus delay the flow of neutral alleles across the cline centre, thereby causing a genetic barrier to gene exchange, and thereby reproductive isolation (Barton and Bengtsson, 1986). The strength of the barrier to gene flow B is given by $\frac{\Delta p}{\frac{\partial p}{\partial x}}$, where Δp is the step in the allele frequency and $\frac{\partial p}{\partial x}$ is the allele frequency gradient at the edge and could delay the flow of neutral alleles by $T = (B/\sigma)^2$ generations.

In the genomic era, hybrid zones have been extensively studied to understand how evolutionary processes maintain distinct populations in the face of gene flow. Geographic cline analysis is the commonly used approach wherein the cline shape and its parameters - cline width (which is inversely proportional to the selection strength) and cline centre - are inferred by fitting allele frequency data to sigmoid and stepped cline models. Combining these with patterns of LD enables researchers to infer the strength of selection and gene flow (Mallet, 1986; Mallet et al., 1990). Furthermore, comparing cline centres across different loci can reveal the underlying genetic mechanisms; for example, indirect selection through strong LD can generate coincident clines whereas genetic drift (Nürnberger et al., 1995) or epistasis (Gavrilets, 1997) can pull them apart. In addition, dominance can cause asymmetries in cline shape as recessive alleles are less visible to selection causing higher introgression on one side of the hybrid zone (Mallet and Barton, 1989).

Inferences based on cline models have proven very useful to infer the strengths of underlying evolutionary processes. However, natural populations are more complex than simple cline models, making it necessary to check whether their assumptions hold true and test against alternate methods (for example pedigree- based fitness or dispersal measures, or using

simulations). One main caveat with current models is the approximation of gene flow as diffusion, justified in the limit of weak selection. However, dispersal distributions are most often leptokurtic, consisting of a small fraction of individuals that travel longer distances (Nathan et al., 2012), and therefore cannot be modelled as a diffusion process when selection is moderate or strong. Secondly, the assumption of uniform population density might not hold true, calling for further studies that examine the interaction of heterogeneous density with non-diffusive dispersal on cline shape and LD. Finally, non-genetic processes like a physical barrier or long-range dispersal can generate stepped clines, making the direct inference of barrier strength from cline shape an overestimate. Thus, identifying the causes of the step and its relation to barriers to gene flow is necessary.

1.6 The model system *Antirrhinum majus*

Antirrhinum, commonly known as snapdragon, is a genus of flowering plants native to the western Mediterranean. They have been used as a model system since the mid 1800's in studies of genetics, plant development, flower colour and pollination biology (Schwarz-Sommer et al., 2003). In this thesis, I focus on the species *Antirrhinum majus*, which is an outcrossing, self-incompatible, hermaphroditic, short lived perennial plant. *A. majus* is diploid with 8 chromosomes and a genome size of $\sim$ 500 Mb. It includes two subspecies: the magenta-flowered *A.m. pseudomajus* and the yellow flowered *A.m striatum* (figure 1.2B). When these populations come into contact, they form narrow hybrid zones with wide range of flower colour hybrid phenotypes (figure 1.2B). However, apart from the flower colour differences that separate the two subspecies, they share the same environment, ecology and pollinators. Around four independent hybrid zones have been identified so far in the Pyrenees, around Spain and France, in Planoles, Cadi, St.Marsal and Avalannet.

Previous studies have done substantial work regarding the molecular genetics of flower colour. These showed that magenta colour is mainly regulated by three genes, tightly linked (0.5 cM apart) loci *Rosea* and *Eluta* (Tavares et al., 2018) and *Rubia* (Field et al in prep.), whereas *Sulfurea* (Bradley et al., 2017) together with *Flavia* and *Cremosa* (Bradley et et al in prep.) control yellow colouration. A preliminary analysis showed steep clines for both flower colours, suggesting that these traits are under selection (Whibley et al., 2006). One hypothesis is that the clines could be maintained by positive frequency dependent selection by bees that pollinate the two subspecies where bees prefer the commonest phenotype or by selection against hybrids as they could be inherently less attractive to the bees. Analysis of genomic divergence between the two subspecies showed F_{ST} peaks only at a few genomic regions coinciding with known location of flower colour genes (Tavares et al., 2018). Thus, there is little divergence between the subspecies in this population

across nearly all of the genome, except for limited regions involved in floral pigmentation, and no genome wide barriers to gene flow have been detected (Ringbauer et al., 2018). A second study confirmed these findings by analysing clines across the genome using 6 pools of individuals that span the hybrid zone (Field et al in prep.). While all these findings suggest that flower colour might be a target of selection, we still lack compelling evidence and quantitative understanding of how selection and dispersal maintain the hybrid zone.

In this thesis, I focus on the hybrid zone in Val Di Ribes, near Planoles (figure 1.2A). This hybrid zone has been extensively sampled every year since 2009 (still ongoing) during the flowering months (between May and August). We record GPS positions using Trimble and collect leaves from each plant which are dried using silica gel and used for DNA extraction. Colour scoring for magenta and yellow components of each flower was also done in the field. KASP SNP genotyping from $\sim$ 22500 individuals was used to design a panel of 248 SNPs at polymorphic and divergent loci. Of these, 91 markers showed no evidence of selection (see details in Ch. 4) and thus represented 'neutral' SNPs. Extensive sampling over a decade showed that this population has a patchy and transient distribution with overlapping generations. In a parallel study, this dataset was used to obtain a pedigree, giving us direct estimates of dispersal and fitness (Field et al in prep.). About 2500 parent-offspring trios were identified using the software SNPPIT, which showed that both distribution of distance between parents (which gives the pollen dispersal distance) and distance between either parent and offspring are leptokurtic (figure 1.2C). Moreover, comparing the years of when the offspring was present vs when its parents were identified suggested that there is a seed bank, with an average seed dormancy of around 3 years. More details regarding the hybrid zone and the model system are given in chapters 4 and 5.

1.7 Thesis Outline

The overarching aim of this thesis is to study how different features of population structure shape genetic variation when populations evolve in the face of gene flow.

Reproductive isolation (RI) is a key concept in speciation. However, RI lacks a unified definition and measures of RI are divided between organismal measures, that focus on the reduced fitness of hybrids and genetic measures, that focus on the reduction in gene flow. Chapter 2 proposes a definition for RI: "a quantitative measure of the effect of genetic differences on gene flow. RI compares the flow of neutral alleles from one population to another population, given a set of genetic differences that reduce gene flow, with the flow expected without any such differences". The exact definition of RI will depend on the spatial and genomic context. We provide analytical predictions for how RI can be measured in different scenarios, in terms of effective migration rates (m_e) and strength of

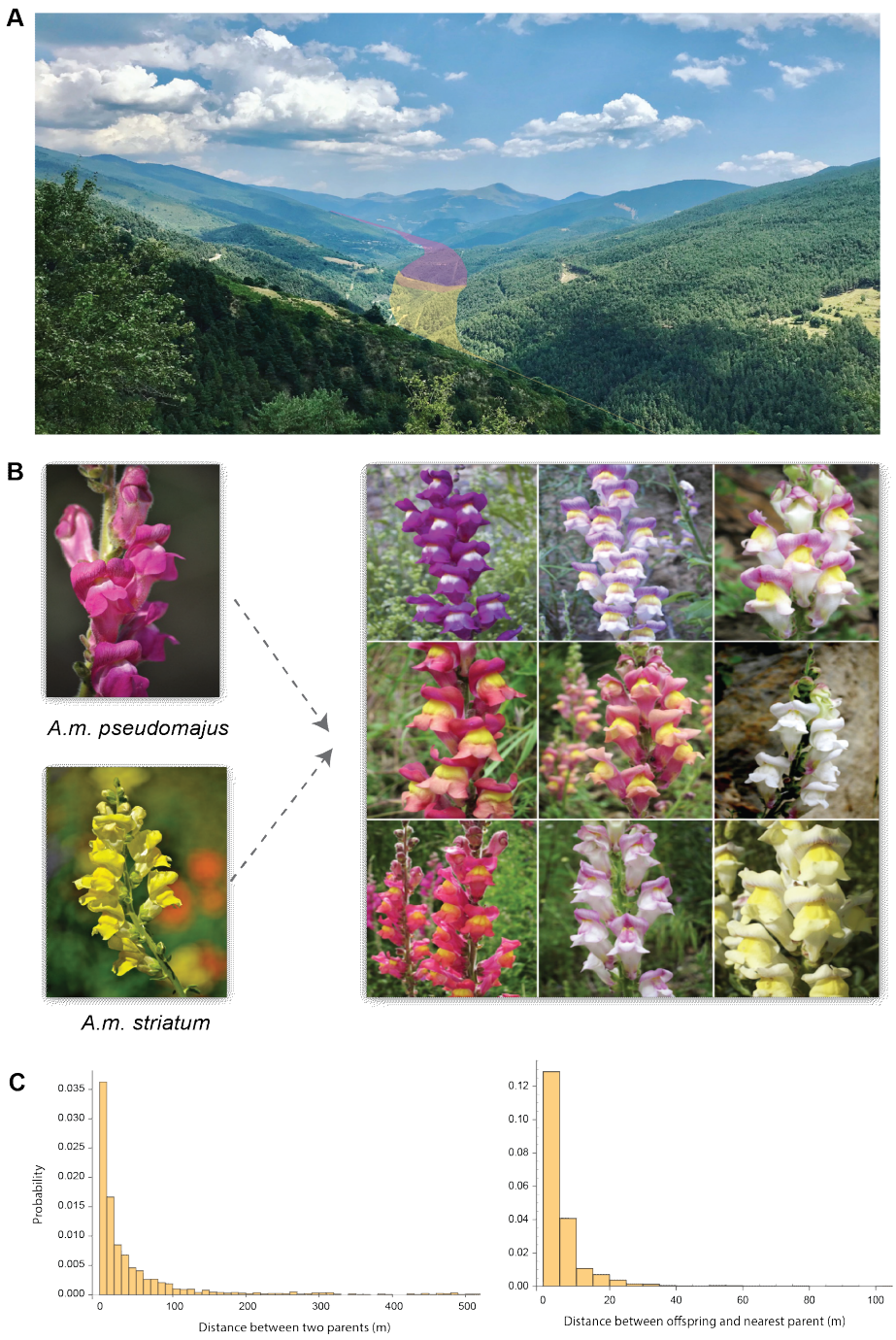


Figure 1.2: A: The Val di Ribes with shaded areas roughly representing the spread of *A.m. pseudomajus* (magenta), *A.m. striatum* (yellow) and hybrids (orange). B: Representative images for the two parental subspecies on the left and wide array of hybrid phenotypes seen in the hybrid zone. C: The left and right plot shows distribution of distances between the parents (denoting the pollen dispersal distance) and distance between the offspring and the nearest parent respectively obtained from 2500 parent-offspring trios.

the barrier to gene flow (B) for discrete and continuous populations respectively. Chapter 2 discusses the need for new measures of RI that combines both organismal and genetic approaches.

In chapter 3, I focus on assortative mating based on self-referencing, which generates population structure due to deviations from random mating expectations. Chapter 3 analyses the effects of assortative mating and sexual selection together with viability selection on RI. For this, I consider a polygenic 'magic' trait, mediated through large number of small effect loci, under a mainland-island model. The same loci influence both divergent selection and mediate assortative mating on the island, thereby generating sexual selection. We obtain novel theoretical predictions for molecular divergence on the island, in terms of effective migration rates, which bears a simple relationship to measurable fitness components of migrants and various early generation hybrids. We also quantify the relative contributions of viability and sexual selection to genome-wide barriers to gene flow. This is especially relevant given long-standing interest in understanding how multiple reproductive barriers (non-)independently contribute to speciation.

In the next two chapters, I shift my focus into studying hybrid zones, specifically using the long-term spatial genetic data of snapdragons from Val Di Ribes, Spain. Chapters 4 and 5 focuses on genetic variation at neutral and selected regions in the genome respectively. In Chapter 4, I examine the effect of heterogeneous density and leptokurtic dispersal on the distribution of heterozygosity. Using spatially explicit individual based simulations, we find that patterns of isolation by distance (which relates to the mean heterozygosity) and the variance in inbreeding (which relates to the (co)variance in heterozygosity) in snapdragons is shaped by their patchy population structure and leptokurtic dispersal distribution.

In the final chapter, I analyse clines at 6 focal markers that mediate flower colour differences in the hybrid zone. We find sharp coincident stepped clines at all loci, suggesting that flower colour is under selection. Further, we show that long-range dispersal causes cline shape and linkage disequilibrium to deviate from those expected from simple cline models. Using multilocus simulations, we show that the stepped cline shape is caused by both the dispersal and moderate LD due to multilocus selection, enabling us to quantify the genetic barrier to gene flow between the hybridizing populations.

To conclude, in this thesis, I show how demography structures genetic variation in the genome and how it leaves distinct and novel patterns in the data. I then combine theory with data and simulations to infer population genetic parameters and identify scenarios when simple models become insufficient.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., et al. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26(2):229–246.

Allman, B. E. and Weissman, D. B. (2018). Hitchhiking in space: Ancestry in adapting, spatially extended populations. *Evolution*, 72(4):722–734.

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

Barton, N. and Bengtsson, B. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57:357–376.

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

Barton, N. H. (1979b). Gene flow past a cline. *Heredity*, 43(3):333–339.

Barton, N. H. (1983). Multilocus clines. *Evolution*, 37:454–471.

Barton, N. H. and De Cara, M. A. R. (2009). The evolution of strong reproductive isolation. *Evolution*, 63(5):1171–1190.

Barton, N. H., Etheridge, A. M., Kelleher, J., and Véber, A. (2013). Genetic hitchhiking in spatially extended populations. *Theoretical Population Biology*, 87:75–89.

Barton, N. H. and Gale, K. S. (1993). Genetic analysis of hybrid zones. In Harrison, R. G., editor, *Hybrid Zones and the Evolutionary Process*, pages 13–45. Oxford University Press, Oxford.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, 16:113–148.

Barton, N. H., Kelleher, J., and Etheridge, A. M. (2010). A new model for extinction and recolonization in two dimensions: quantifying phylogeography. *Evolution*, 64(9):2701–2715.

Battey, C. J., Ralph, P. L., and Kern, A. D. (2020). Space is the place: effects of continuous spatial structure on analysis of population genetic data. *Genetics*, 215(1):193–214.

Bazykin, A. (1969). Hypothetical mechanism of speciaton. *Evolution*, 23:685–687.

Bengtsson, B. (1985). The flow of genes through a genetic barrier. In *Evolution: Essays in honour of John Maynard Smith*, pages 31–42. Cambridge University Press, Cambridge.

Bierne, N., Roze, D., and Welch, J. J. (2013). Pervasive selection or is it...? why are f_{ST} outliers sometimes so frequent? *Molecular Ecology*, 22(8):2061–2064.

Bradburd, G. S., Coop, G. M., and Ralph, P. L. (2018). Inferring continuous and discrete population genetic structure across space. *Genetics*, 210(1):33–52.

Bradburd, G. S. and Ralph, P. L. (2019). Spatial population genetics: It's about time. *Annual Review of Ecology, Evolution, and Systematics*, 50(1):427–449.

Bradburd, G. S., Ralph, P. L., and Coop, G. M. (2016). A spatial framework for understanding population structure and admixture. *PLoS Genetics*, 12(1):e1005703.

Bradley, D., Xu, P., Mohorianu, I. I., Whibley, A., Field, D., Tavares, H., and Coen, E. (2017). Evolution of flower color pattern through selection on regulatory small rnas. *Science*, 358(6365):925–928.

Broquet, T., Viard, F., and Yearsley, J. M. (2013). Genetic drift and collective dispersal can result in chaotic genetic patchiness. *Evolution*, 67(6):1660–1675.

Butlin, R. K. and Smadja, C. M. (2018). Coupling, reinforcement, and speciation. *The American Naturalist*, 191(2):155–172.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates, Sunderland, MA.

Dobzhansky, T. (1937). *Genetics and the origin of species*. Columbia University Press, New York, 1st edition.

Dopman, E. B., Shaw, K. L., Servedio, M. R., Butlin, R. K., and Smadja, C. M. (2024). Coupling of barriers to gene exchange: causes and consequences. *Cold Spring Harbor Perspectives in Biology*, page a041432.

Duforet-Frebourg, N. and Slatkin, M. (2016). Isolation-by-distance-and-time in a stepping-stone model. *Theoretical Population Biology*, 108:24–35.

Endler, J. A. (1977). *Geographic variation, speciation, and clines*, volume 10 of *Monographs in Population Biology*. Princeton University Press.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

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

Felsenstein, J. (1975). A pain in the torus: some difficulties with models of isolation by distance. *The American Naturalist*, 109(967):359–368.

Felsenstein, J. (1981). Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution*, 35(1):124–138.

Gavrilets, S. (1997). Hybrid zones with Dobzhansky-type epistatic selection. *Evolution*, 51(4):1027–1035.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., et al. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722.

Guillot, G., Leblois, R., Coulon, A., and Frantz, A. C. (2009). Statistical methods in spatial genetics. *Molecular Ecology*, 18(23):4734–4756.

Guillot, G. and Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, 4(4):336–344.

Haldane, J. (1948). The theory of a cline. *Journal of Genetics*, 48:277–284.

Haller, B. C. and Messer, P. W. (2019). Slim 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, 36(3):632–637.

Hardy, O. J. and Vekemans, X. (2002). Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, 2(4):618–620.

Hewitt, G. (1988). Hybrid zones - natural laboratories for evolutionary studies. *Trends in Ecology & Evolution*, 3:158–167.

Hey, J. and Machado, C. A. (2003). The study of structured populations—new hope for a difficult and divided science. *Nature Reviews Genetics*, 4(7):535–543.

Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167(2):747–760.

Holderegger, R. and Wagner, H. H. (2006). A brief guide to landscape genetics. *Landscape Ecology*, 21(6):793–796.

Jiang, Y., Bolnick, D. I., and Kirkpatrick, M. (2013). Assortative mating in animals. *The American Naturalist*, 181(6):E125–E138.

Kimura, M. and Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561–576.

Kirkpatrick, M. and Nuismer, S. L. (2004). Sexual selection can constrain sympatric speciation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1540):687–693.

Kirkpatrick, M. and Ravigné, V. (2002). Speciation by natural and sexual selection: models and experiments. *The American Naturalist*, 159(S3):S22–S35.

Kopp, M., Servedio, M. R., Mendelson, T. C., Safran, R. J., Rodríguez, R. L., Hauber, M. E., Scordato, E. C., Symes, L. B., Balakrishnan, C. N., Zonana, D. M., and van Doorn, G. S. (2018). Mechanisms of assortative mating in speciation with gene flow: Connecting theory and empirical research. *The American Naturalist*, 191(1):1–20.

Kulmuni, J., Butlin, R. K., Lucek, K., Savolainen, V., and Westram, A. M. (2020). Towards the completion of speciation: the evolution of reproductive isolation beyond the first barriers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1806):20190528.

Legendre, P. and Fortin, M.-J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, 10(5):831–844.

Lewontin, R., Kirk, D., and Crow, J. (1968). Selective mating, assortative mating, and inbreeding: definitions and implications. *Eugenics Quarterly*, 15(2):141–143.

Li, W.-H. and Nei, M. (1974). Stable linkage disequilibrium without epistasis in subdivided populations. *Theoretical Population Biology*, 6(2):173–183.

Lohse, K. and Frantz, L. A. F. (2014). Neandertal admixture in eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, 196(4):1241–1251.

Lundgren, E. and Ralph, P. L. (2019). Are populations like a circuit? comparing isolation by resistance to a new coalescent-based method. *Molecular Ecology Resources*, 19(6):1388–1406.

Maan, M. E. and Seehausen, O. (2011). Ecology, sexual selection and speciation. *Ecology Letters*, 14(6):591–602.

Malécot, G. (1948). *The mathematics of heredity*. Masson & Cie, Paris.

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. and 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., and 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.

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18(4):189–197.

Maruyama, T. (1972). Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics*, 70(4):639–651.

Mayr, E. (1942). *Systematics and the origin of species*. Columbia University Press, New York.

McRae, B. H. (2006). Isolation by resistance. *Evolution*, 60(8):1551–1561.

Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular Ecology*, 21(12):2839–2846.

Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792.

Nagylaki, T. (1988). The influence of spatial inhomogeneities on neutral models of geographical variation: I. formulation. *Theoretical Population Biology*, 33(3):291–310.

Nathan, R., Klein, E., Robledo-Arnuncio, J. J., and Revilla, E. (2012). *Dispersal kernels*, volume 15. Oxford University Press, Oxford, UK.

Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649.

Nürnberger, B., Barton, N., MacCallum, C., Gilchrist, J., and Appleby, M. (1995). Natural selection on quantitative traits in the *Bombina* hybrid zone. *Evolution*, 49(6):1224–1238.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics*, 192(3):1065–1093.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across Europe. *PLoS Biology*, 11(5):e1001555.

Richardson, J. L., Brady, S. P., Wang, I. J., and Spear, S. F. (2016). Navigating the pitfalls and promise of landscape genetics. *Molecular Ecology*, 25(4):849–863.

Ringbauer, H., Coop, G., and Barton, N. H. (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351.

Ringbauer, H., Kolesnikov, A., Field, D. L., and Barton, N. H. (2018). Estimating barriers to gene flow from distorted isolation-by-distance patterns. *Genetics*, 208(3):1231–1245.

Robledo-Arnuncio, J. J. and Rousset, F. (2010). Isolation by distance in a continuous population under stochastic demographic fluctuations. *Journal of Evolutionary Biology*, 23(1):53–71.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. *Genetics*, 145(4):1219–1228.

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for windows and linux. *Molecular Ecology Resources*, 8(1):103–106.

Rundle, H. D. and Nosil, P. (2005). Ecological speciation. *Ecology Letters*, 8(3):336–352.

Safran, R. J., Scordato, E. S., Symes, L. B., Rodríguez, R. L., and Mendelson, T. C. (2013). Contributions of natural and sexual selection to the evolution of premating reproductive isolation: a research agenda. *Trends in Ecology & Evolution*, 28(11):643–650.

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303.

Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., and Manel, S. (2012). Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution, and Systematics*, 43(1):23–43.

Schwarz-Sommer, Z., Davies, B., and Hudson, A. (2003). An everlasting pioneer: the story of *Antirrhinum* research. *Nature Reviews Genetics*, 4(8):655–664.

Servedio, M. R. (2009). The role of linkage disequilibrium in the evolution of premating isolation. *Heredity*, 102(1):51–56.

Servedio, M. R. and Boughman, J. W. (2017). The role of sexual selection in local adaptation and speciation. *Annual Review of Ecology, Evolution, and Systematics*, 48(1):85–109.

Servedio, M. R. and Kopp, M. (2012). Sexual selection and magic traits in speciation with gene flow. *Current Zoology*, 58(3):510–516.

Servedio, M. R., van Doorn, G. S., Kopp, M., Frame, A. M., and Nosil, P. (2011). Magic traits in speciation: 'magic' but not rare? *Trends in Ecology & Evolution*, 26(8):389–397.

Slatkin, M. (1973). Gene flow and selection in a cline. *Genetics*, 75(4):733–756.

Slatkin, M. (1975). Gene flow and selection in a two-locus system. *Genetics*, 81(4):787–802.

Smadja, C. M. and Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, 20(24):5123–5140.

Smith, T. B. and Weissman, D. B. (2023). Isolation by distance in populations with power-law dispersal. *G3: Genes, Genomes, Genetics*, 13(4):jkad023.

Stankowski, S. and Ravinet, M. (2021). Defining the speciation continuum. *Evolution*, 75:1256–1273.

Storfer, A., Murphy, M. A., Evans, J. S., Goldberg, C. S., Robinson, S., Spear, S. F., Dezzani, R., Delmelle, E., Vierling, L., and Waits, L. P. (2007). Putting the 'landscape'in landscape genetics. *Heredity*, 98(3):128–142.

Szymura, J. and Barton, N. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina Bombina* and *B. Variegata*, near cracow in southern poland. *Evolution*, 40:1141–1159.

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., and 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.

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

Whitlock, M. C. and McCauley, D. E. (1999). Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. *Heredity*, 82(2):117–125.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2):97–159.

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114–138.

Yair, S., Lee, K. M., and Coop, G. (2021). The timing of human adaptation from neanderthal introgression. *Genetics*, 218(1):iyab052.

Chapter 2

What is reproductive isolation?[1]

Abstract

Reproductive isolation (RI) is a core concept in evolutionary biology. It has been the central focus of speciation research since the modern synthesis and is the basis by which biological species are defined. Despite this, the term is used in seemingly different ways, and attempts to quantify RI have used very different approaches. After showing that the field is divided over the precise meaning of the term, we attempt to clarify key issues, including what RI is, how it can be quantified in principle, and how it can be measured in practice. Following other definitions with a genetic focus, we propose that RI is a quantitative measure of the effect of genetic differences on gene flow. RI compares the flow of neutral alleles from one population to another population, given a set of genetic differences that reduce gene flow, with the flow without any such differences. We show how RI can be quantified in a range of scenarios. A key conclusion is that RI depends strongly on circumstances—including the spatial, temporal and genomic context—making it difficult to quantify it in a way that is directly comparable across systems. After reviewing methods for estimating RI from empirical data, we conclude that it is difficult to measure in practice. We discuss our findings in light of the goals of speciation research and encourage the use of methods for estimating RI that integrate organismal and genetic approaches.

Keywords: adaptation, genomics, natural selection, population genetics, speciation, theory

2.1 Introduction

A biological species is defined as a group of interbreeding natural populations that are reproductively isolated from other such groups (Mayr, 1942; Coyne and Orr, 2004). The

[1]This work can be found online at https://doi.org/10.1111/jeb.14005

notion of reproductive isolation (RI) is thus central to our understanding of species and speciation. But what, exactly, do we mean by 'reproductive isolation'? Despite being deeply embedded in the language of speciation, the term is used in seemingly different ways, usually without a precise definition; attempts to quantify it have used very different approaches that measure different things. Bridging different perspectives and approaching a general definition is important for our conceptual understanding of speciation and for efforts to quantify RI empirically. The main aim of this article is to contribute to progress in this respect.

In Box 1, we briefly summarise the history of the term and present the results of a recent survey on RI among evolutionary biologists working on speciation. Both the survey and the historical overview suggest that researchers tend to focus on different aspects when they explain what RI means for them, mainly falling into two groups: a reduction in the production or fitness of hybrids ("organismal focus") or a reduction in gene flow ("genetic focus"). Despite these differences in focus, these perspectives of RI are not contradictory or mutually exclusive: many proponents of a genetic focus point out that gene flow is reduced because there is a reduction in the production and fitness of hybrids (Gavrilets, 2004). Similarly, some proponents of the organismal focus emphasise that the reduced production of hybrids is relevant because it restricts gene flow and leads to the formation of genetically distinct clusters (Sobel and Chen, 2014).

Thus, while different perspectives highlight different aspects of RI, they are clearly related to one another: Reproductive isolation refers to a situation involving a pair of populations; genetic differences between them lead to a reduction in hybrid formation or fitness (e.g. different adaptations, different mating preferences, or intrinsic incompatibilities), which in turn leads to a restriction in gene flow (Fig. 2.1).

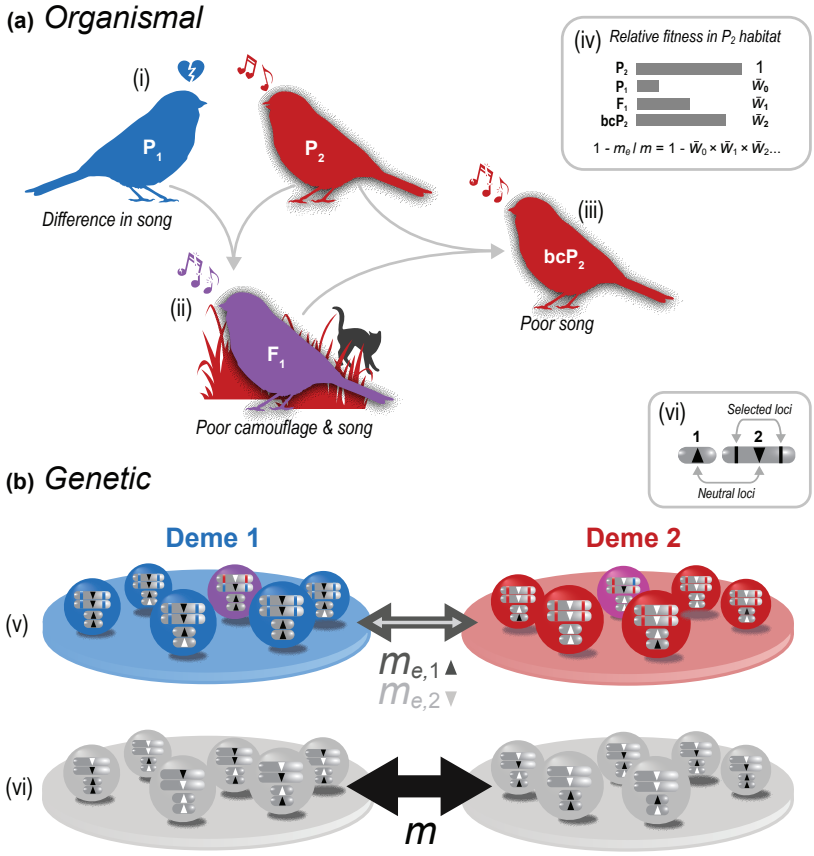


Figure 2.1:
Different but complementary perspectives of reproductive isolation. A) The 'organismal' perspective of RI tends to focus on the reduction in successful interbreeding between taxa. In this example, two bird populations (P_1 and P_2) have diverged for colour and song. Reduced attractiveness and camouflage cause immigrants, F_1s, and backcrosses (bc) to have lower mean fitness relative to the resident type (a-c). (d) Example relative fitnesses are given for a P_1 immigrant ($\overline{W}_0$), F_1 ($\overline{W}_0$) and P_2 backcross ($\overline{W}_2$), relative to a resident P_2 individual. In a 2-deme model with low migration, the gene flow for an unlinked neutral locus relative to that expected without any barriers to gene flow (m_e/m) is the product of the mean fitnesses of successive hybrid classes ($\overline{W}_0\overline{W}_1\overline{W}_2\ldots$). B) The genetic perspective tends to focus on the reduction in gene flow between populations due to selection acting on genetic differences. This is illustrated for a 2-deme model with a genetic barrier (e), contrasted with a neutral scenario with no barrier (f). In both scenarios, diploid individuals carry $n = 2$ chromosomes. (f) Both chromosomes (1 and 2) carry a neutral locus (up and down facing triangles), each with two alternative alleles (black and white). In the barrier scenario (e), the neutral locus on chromosome 2 is flanked by a pair of loci that affect fitness; blue and red alleles at these loci maximize fitness in demes 1 and 2, respectively, but severely reduce fitness in the other deme. Although individuals migrate between demes at the same rate in both scenarios, the rate of gene flow at neutral loci—which is proportional to the width of the arrows and evident from the amount of neutral allele sharing between the demes—is lower than expected in the barrier scenario due to their association with selected loci. Note that this 'effective' migration rate (m_e) differs between the two neutral loci, and is more strongly reduced for neutral alleles linked to the selected loci.

Despite this common conceptual understanding of RI, there is much need for clarification, which we attempt in this article. The main points that need to be addressed include
1. What, precisely, is reproductive isolation? Should we define and measure it in terms of patterns of reproduction, or should we instead focus on the corresponding reduction in gene flow?
2. RI as verbally defined above, and as often used in the literature, refers to a feature of populations or species, rather than a quantity. In the survey (Box 1), almost none of the respondents described RI as a quantitative measure. To compare different barriers to gene flow or different study systems, to study speciation over time, and to study genomic patterns we need to be able to quantify RI (e.g. Stankowski and Ravinet (2021)). There is an understanding in the literature that RI should measure the range from gene flow unrestricted by genetic differences at one extreme, to complete speciation (when no fertile hybrids are produced and no gene flow occurs) at the other extreme. Yet, how to quantify RI between these two extremes is unclear.
3. Divergence occurs in situations often much more complex than depicted in Fig. 2.1. For example, divergence often happens in continuous space, making a 2-deme model unrealistic. How do we define RI in continuous or complex space, and what are the entities (i.e., populations) between which we aim to measure it in the first place? And how do we integrate the fact that RI also varies over time (e.g. when two populations have come into secondary contact after divergence in allopatry)?
4. To be useful empirically, we must be able to estimate RI from field or experimental data. There are several very different approaches to measuring RI in the literature. Importantly, when trying to quantify RI, the precise focus on organismal vs. genetic aspects of RI becomes relevant, as these are associated with very different methods (e.g. using sequencing data vs. lab crosses). Approaches also differ in whether there is a single estimate for a given population pair, or a series of estimates reflecting variation along the genome. Given a quantitative definition of RI, how do we best measure it from empirical data, and how do existing approaches perform?
In this article, we will provide a general definition of RI based on patterns of gene flow and show how it can lead to a quantitative measure in different spatial scenarios, and explain its relationship with the organismal focus. We first illustrate RI in simple scenarios (2-deme and continuous space), and then in more complex examples. Finally, we discuss how to estimate RI from empirical data, asking whether, and under which circumstances, existing measures of RI reflect our quantitative definition.

2.2 Towards a general definition of RI

After some consideration and debate we have settled on a general definition of RI not all that different to one posed by (Dobzhansky, 1937) or other gene flow-based definitions

(Gavrilets, 2004; Stankowski and Ravinet, 2021)(For the definition of RI and other terms also see the Glossary (2.8); Table S1 (2.8)). We chose to define RI in terms of gene flow, because the level of gene exchange is what ultimately determines the extent to which populations can evolve independently. We propose that RI is a quantitative measure of the effect of genetic differences on gene flow. RI compares the flow of neutral alleles from one population to another population, given a set of genetic differences that reduce gene flow, with the flow without any such differences (Fig. 2.2). The exact definition depends on the spatial context. By "population", we simply mean a set of individuals. This could be defined by spatial position, but this is not necessarily so (see below). By "genetic differences that reduce gene flow", we mean any genetic differences contributing to traits (at the organismal level) that restrict gene flow between groups of individuals, including both intrinsic incompatibilities and adaptations to local environments. These loci are often under divergent selection, but can also include loci contributing to assortative mating or habitat choice.

Importantly, RI is defined only for neutral loci, not for the barrier loci (i.e. the loci contributing directly to the genetic differences that reduce gene flow, such as loci under divergent selection) themselves. Neutral loci can, however, be perfectly linked to loci contributing to barriers. We limit the definition of RI to neutral loci because we wish to separate selection on specific alleles from its effect on gene flow at linked and unlinked neutral loci.

While we define RI explicitly in terms of gene flow, it is directly connected to the "organismal focus", as genetically-based barriers are barriers to the production or fitness of hybrids. Genetically-based barriers include, for example, intrinsic incompatibilities. If two populations contain incompatible alleles, fewer hybrids will be produced and thus the flow of neutral alleles between these populations or areas will be reduced. Genetically-based barriers also include those leading to assortative mating or habitat choice, if these barriers have a genetic basis and lead to a reduced probability of alleles moving between populations or areas. Importantly, genetically-based barriers also include environment-dependent barriers, i.e. loci contributing to local adaptation in a heterogeneous environment. For example, if large individuals are favoured by selection in one population and small individuals are favoured in another population, selection against migrants and hybrid individuals reduces the effective gene flow, generating RI.

Geographic isolation, e.g. geographical distance or a physical barrier (e.g. a mountain range or a river), reduces gene flow but does not contribute to RI because it is not a genetically-based feature of the individuals, but rather entirely environmental. However, physical barriers and genetic barriers can interact to modulate RI (see below).

In any given situation, three things are needed to formulate a concrete definition of RI: i) the two populations, ii) the timescale over which the allele movement is to be considered,

and iii) the genomic position of the focal neutral allele.

How to define the two populations is not necessarily obvious. First, groups can be defined spatially. Below we describe RI in a simple system of two demes connected by gene flow. In that case, it is obvious that the two demes correspond to the two populations. In arbitrarily complex spatial settings, it may be necessary to specify two areas between which we want to determine RI; the estimate of RI may differ substantially depending on which areas we choose. Second, populations could be defined by traits. For example, if two hybridizing taxa occur in sympatry, a trait that typically diagnoses them (e.g. plumage colour) could be used. Third, populations could be genetic clusters. For example, if two cryptic species occur in sympatry, genetic markers could be used to identity two genetic groups (e.g. using Principal Component Analysis or STRUCTURE; Pritchard et al. (2000)). It is important to note that any grouping will be somewhat arbitrary: the fact that there is gene flow between the two "populations" means that there are not actually two clearly separated groups. Reflecting the same issue, different groupings would typically lead to different results. For example, as alleles underlying traits can introgress, and as an individual in a deme might be a migrant, space- or trait-based clusters may not be identical to genetic clusters. Many environments are complex, local adaptation to various environmental factors occurs, and different environmental transitions or gradients do not necessarily coincide in space. In such cases, there are multiple possible populations between which we could measure RI, and a single value is certainly not sufficient to generally describe patterns of RI.

Similarly, the temporal scale might not be straightforward to define. In scenarios where selected loci are at migration-selection equilibrium (see examples below), the rates of gene flow stay constant over time, and no specific timescale needs to be defined. However, when selected loci are not at equilibrium, e.g. after a secondary contact between divergent populations, allele frequencies at selected loci change over time (e.g. while incompatibilities are purged or uniformly favoured alleles introgress) and thus the rate at which neutral alleles are exchanged may change as well.

Finally, it is crucial to define the focal genomic region for which the reduction in gene flow is determined. In the literature, RI is mostly discussed as a genome-wide concept, with a single measure for a pair of populations or a hybrid zone (e.g. Lowry et al. (2008); Schluter (2009); Rabosky (2016)). However, RI is also sometimes described as related to the effective migration rate, m_e, which varies along the genome, suggesting that RI varies along the genome as well (Barton and Bengtsson, 1986). RI as a single genome-wide concept must reflect the general barrier to gene flow experienced by a locus without any specific features – that is, a neutral locus unlinked to any selected loci. In contrast, RI along the genome reflects how effective gene flow varies along the genome depending on the association with particular loci contributing to reproductive barriers. We refer to

these different concepts of RI as "genome-wide RI" and "local RI", respectively.

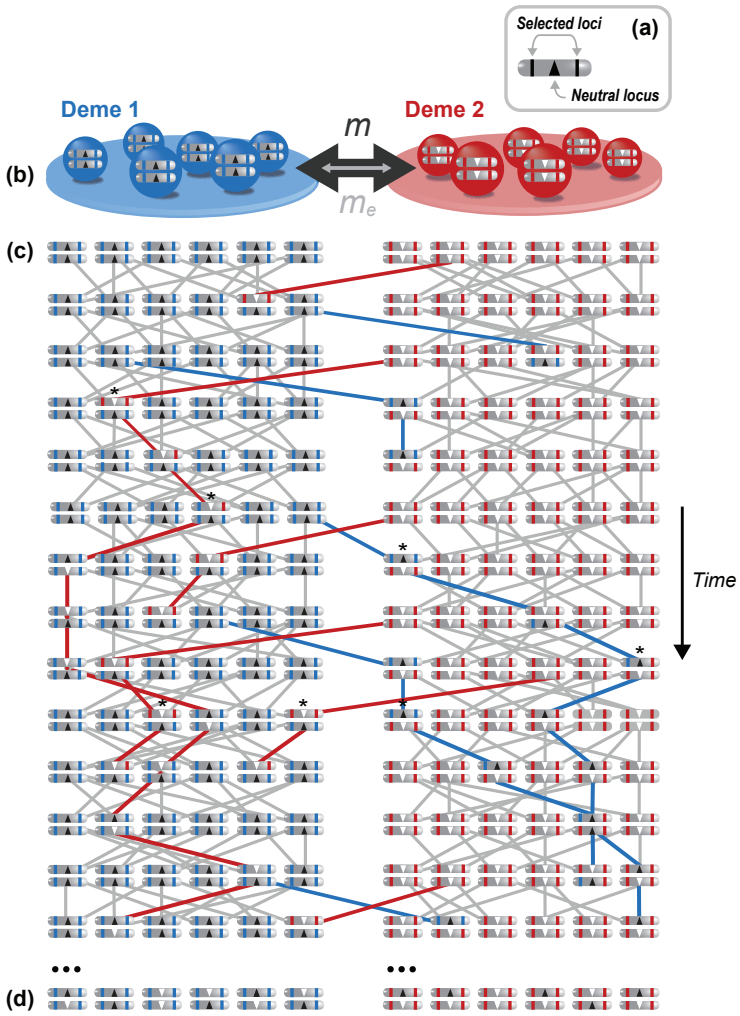


Figure 2.2: A pedigree depicting the movement of neutral alleles between populations and genetic backgrounds in the presence of a barrier. A) and B) Diploid individuals carry one chromosome with a single neutral marker flanked by two loci divergently selected between two demes. C) At time 0 (top row), the two demes are fixed for alternative alleles at the neutral and selected loci. The blue and red selected loci maximize fitness in demes 1 and 2, respectively, but severely reduce fitness in the other deme. Panel C depicts the pedigree for the 2 demes across multiple generations. Lines connect parents (row i) and their offspring (row $i+1$). Coloured lines trace the passage of immigrant haplotypes through the pedigree. Because divergent selection is strong, immigrant alleles can only persist for a short time in the foreign deme. Because they are associated with selected sites, the movement of neutral alleles between demes is also restricted, but they can persist when they fall onto the local genetic background. In this example, escaping their association with selected loci requires two recombination events (recombination events are indicated by the asterisks, while the line style indicates the strength of the association between foreign neutral and functional alleles). D) Eventually, the allele frequency difference at neutral loci will be homogenized, but this takes much longer than expected in the absence of a barrier.

In the following, we first provide simple examples where an exact quantitative definition of RI can be given. How this can be measured in reality is a separate question and will be discussed in a later section ('Estimating RI from empirical data').

2.3 Example scenario 1: Gene flow into a single deme

Assumptions

The simplest case is a situation with two populations (demes) with unidirectional gene flow from the source population into the focal recipient population. RI is generated by a set of selected loci that are fixed different between the source and the recipient population. We assume that the rate of gene flow is low, such that complex hybrids are rarely generated and can be neglected, that our assumption of fixed differences at selected loci is a good approximation, and that the processes under unidirectional gene flow (assumed for simplicity) approximate those under bidirectional gene flow.

Quantitative definition of RI under this model

Gene flow between demes is described by a migration rate, m, which reflects the proportion of migrants in the focal deme after migration. This migration rate describes migration in the absence of reproductive barriers between demes. However, if reproductive barriers do exist, the actual rate at which neutral alleles from the source reach the recipient population is reduced, e.g. because they are associated with alleles that are selected against in the recipient population (Fig. 2.1). They will in that case be more likely to be removed from the recipient population than in the absence of barriers, which is mathematically equivalent to a lower effective migration rate, m_e (Bengtsson, 1985; Barton and Bengtsson, 1986). Analogous to the baseline migration rate, m_e is simply the allele frequency change in a generation (δp), relative to the allele frequency difference between demes (Δp), $m_e = \delta p / \Delta p$.

Bengtsson (1985) describes the effective migration rate as "that rate of migration which would have the same evolutionary effect in a population with no genetic barrier as the actual migration rate (m) now has in the population with a barrier". The effective migration rate is more strongly reduced for neutral alleles closely linked to selected loci and increases with increasing distance from such loci.

In the described scenario RI (in this case labelled RI_{2D} for "2 deme") can be defined through the ratio between m_e and m:

$$\mathrm{RI}_{2D} = 1 - m_e/m \tag{2.1}$$

If a locus is not affected by a barrier to gene flow, $m_e = m$, and $\mathrm{RI}_{2D}= 0$. If gene flow at a locus is completely prevented, $m_e = 0$ and $\mathrm{RI}_{2D}=1$.

In the limit of low migration rates, and when the population is at a genetic and demographic equilibrium, this quantity depends only on the source and recipient genotypes. In the following, we describe how both genome-wide and local RI can be calculated under this model.

Genome-wide RI: RI due to unlinked loci

We first show how to calculate RI if all selected loci are unlinked. Genome-wide RI is defined by gene flow at a focal neutral locus that is not linked to any selected loci. Migrants enter the recipient population at a rate m, and alleles at the focal neutral locus enter the recipient population at the same rate. To find m_e, we need the probability that an allele coming from the source population recombines onto the native genetic background. As m_e reflects an average over time rather than a snapshot, we need to consider this probability not just for the first generation after a neutral allele enters the recipient population, but for all following generations as well. Thus, m_e/m is the probability that an allele in a newly arrived migrant will ultimately recombine onto the native genetic background. (This is essentially the reproductive value (Fisher, 1930; Grafen, 2006) of a fresh migrant, relative to that of native individuals).

When first entering the recipient population, the allele experiences selection because it is located in an individual with a complete source population genome, thus containing all divergent alleles selected against in the recipient population. This individual will have a fitness $\overline{W}_0 < 0$ (relative to pure individuals of the recipient population, which will have a fitness of 1). $\overline{W}_0$ includes reductions in fitness due to all possible genetic differences, including reductions in viability, mating success and fecundity. In the following generation, the allele will be located in F_1 hybrids between source and recipient genomes. These F_1 hybrids still carry alleles from the source population, but the number of such alleles is halved compared to first-generation migrants, so that selection is weaker; in addition, there might be heterosis effects. In all following generations, the focal allele (if still present) will be located in backcrosses with native individuals, with fewer and fewer source alleles. Thus, the negative selection experienced by the focal neutral allele weakens over time as it becomes progressively decoupled from other source alleles (Fig. 2.2).

As m_e describes the combined effect of the migration rate m and the probability to persist in the recipient population despite selection, it is simply $m_e = m\overline{W}_0\overline{W}_1\overline{W}_2\ldots$. RI is then

$$\mathrm{RI}_{2D} = 1 - m_e/m = 1 - \overline{W}_0\overline{W}_1 \prod_{k=2}^{\infty} \overline{W}_k \tag{2.2}$$

(Bengtsson, 1985). For all generations, it is important to note that the relevant fitnesses are those of the actual migrant or hybrid genotypes in the recipient population, and that hybrid genotypes are modified by selection (through progressive purging of locally deleterious alleles). Thus, these $\overline{W}_i$ are not necessarily identical to hybrid fitnesses determined e.g. in lab backcrosses.

If we do ignore this purging effect, we can assume that the average number of selected alleles halves in each generation. Further assuming that fitness is multiplicative across loci, we can then simplify (see Appendix A.2 Flow into a single deme):

$$\mathrm{RI}_{2D} = 1 - m_e/m = 1 - \overline{W}_0\overline{W}_1 \prod_{k=2}^{\infty} \overline{W}_k \approx 1 - \overline{W}_0\overline{W}_1\overline{W}_2^2 \tag{2.3}$$

In Appendix A.2: Flow into a single deme, unlinked loci, we show that this approximation is still roughly correct despite our neglect of purging, unless selection is very strong. Even with epistasis, this is still likely to be a good approximation.

In this simplest case, therefore, we can calculate the effective migration rate, and thus a measure of reproductive isolation, from the mean fitnesses of successive backcross generations. RI is determined primarily by the first few generations: the influx of genes is reduced by the same factor in the first backcross generation as in all subsequent backcrosses. We have described a two-deme situation, but a similar approach will work for two taxa in complete sympatry if the level of gene flow between them is low.

Local RI: RI along the genome

To understand local RI as it varies along the genome, we need to consider linkage. Compared to unlinked neutral loci (previous section), neutral loci linked to selected loci are more strongly affected by selection; to persist in the recipient population they must recombine onto the recipient genetic background before being eliminated by selection. Thus, the key parameter is the strength of selection, relative to recombination.

The simplest case is a focal neutral locus linked to a single selected locus with selection coefficient s, separated by a recombination distance r. Barton and Bengtsson (1986) show that in this case

$$\mathrm{RI}_{2D} = 1 - r/(s + r) \tag{2.4}$$

This equation demonstrates that a neutral locus tightly linked to the selected locus ($r \ll s$) experiences a strong barrier, and RI_{2D} approaches 1. Gene flow is reduced substantially within a region of size $r \approx s$ (Fig. 2.3), and RI_{2D} decreases with increasing distance from the selected locus. However, even unlinked loci, with $r = 1/2$, give $\mathrm{RI}_{2D} \sim 2s$, consistent with the previous section showing that even without linkage to selected loci, considerable RI can be produced.

Next, suppose that there is one selected locus on either side of the focal neutral locus, with selection and recombination rates on the left (s_1, r_1) and on the right (s_2, r_2), and the selective effect of the two alleles together $(s_{1,2})$. The effective migration rate is now the product of the effects of each locus, multiplied by a term which equals 1 if there is no epistasis (i.e. $s_{1,2} = s_1 + s_2$):

$$\text{RI\_2D} = 1 - \frac{r_1}{r_1 + s_1} \frac{r_2}{r_2 + s_2} \frac{r_1 + r_2 + s_1 + s_2}{r_1 + r_2 + s_{1,2}} \tag{2.5}$$

(from Eqs. A6, A7 of Barton and Bengtsson (1986)). This equation shows that, assuming a constant distance between the two selected loci, a neutral locus equidistant from both selected loci experiences the lowest levels of RI, while a neutral locus tightly linked to one of the selected loci can experience a much stronger barrier, even though it is further away from the second selected locus (Fig. 2.3).

Finally, there might be multiple selected loci on either side of the focal neutral locus. For a given set of selection coefficients and recombination rates, RI can readily be calculated, but there is no simple general equation describing RI when the number of selected loci is large. Also in this case, the ratio of selection to recombination is a crucial parameter. RI decreases the further the selected loci are apart and (as in the previous example) the further away the neutral locus is from the nearest selected locus (see Appendix A.2: Flow into a single deme, Linked loci, for further details).

In this section, we considered the effects of divergent selection on tightly linked loci; previously, we considered unlinked loci. The relative contributions of these two for an average neutral locus depend on the length of the genetic map, and on the number of loci. Consider the simplest case, where selection is spread uniformly over a map of length R, with total multiplicative selection $S = \theta R$. In this case, with relatively strong selection (i.e.,$S \gg R$), unlinked loci dominate; linked loci would dominate only for small or moderate θ (≤ 1, say), and for extremely large numbers of loci ($\log n > R$) (see Appendix A.2: Relative contributions of unlinked vs. linked loci). However, these two components of reproductive isolation act over different timescales: selection over the whole genome (which is mostly unlinked) acts to quickly reduce the contribution of migrants, by severely reducing the fitness of early generation backcrosses. Selection at tightly linked loci then acts over a much longer timescale, because it takes longer for recombination to break up the association between the selected and the neutral locus. Even if linked selection is responsible for only a small fraction of reproductive isolation, it may still cause "islands of divergence" around the selected loci. Conversely, the genome-wide RI cannot be found reliably by summing the estimated distant effects of the islands.

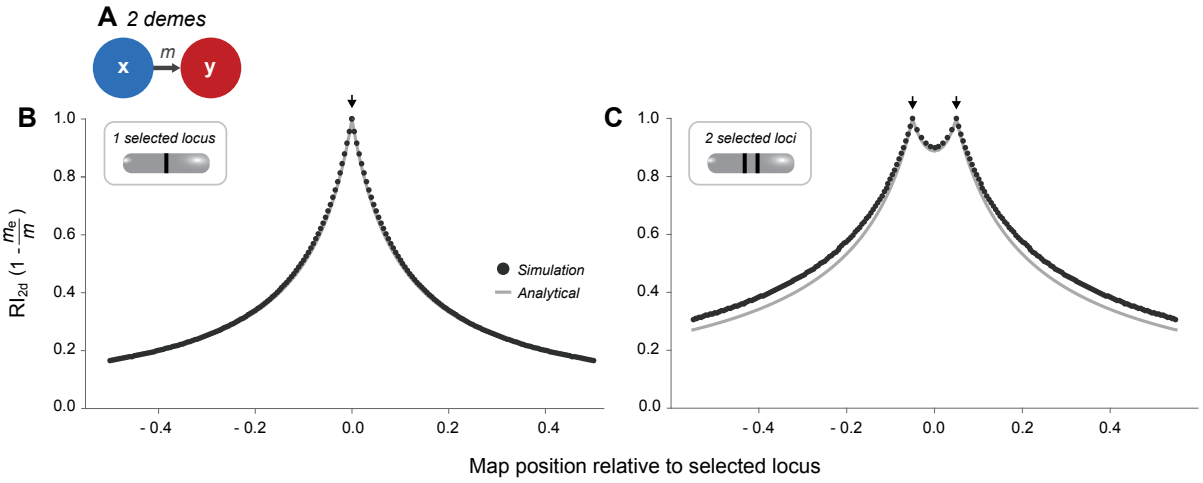


Figure 2.3: **Reproductive isolation in a two-deme scenario**. A) Two panmictic demes connected by unidirectional gene flow; the two locations between which RI is measured simply corresponds to the two demes. B) and C) RI along the genome under this model, with a single locus under selection at position 0 (B) or two loci under selection at positions -0.5 and 0.5. The x-axis gives the recombination rate of the neutral locus relative to position 0.0 (C). The black curve corresponds to deterministic simulations (details in Appendix A.1); the grey curve corresponds to Eq. 2.5. $s = 0.1$ for all selected loci.

2.4 Example scenario 2: Hybrid zones

Assumptions

Where divergent populations meet in continuous space, rather than in a 2-deme situation (previous section), they may be separated by hybrid zones, i.e. they locally interbreed and form hybrids, but remain distinct away from the hybrid zone centre (Bazykin, 1969; Barton and Hewitt, 1985; Stankowski et al., 2021). We here assume a hybrid zone in a single dimension. The reduction of gene flow across the zone can be caused by intrinsic as well as extrinsic components, but importantly, we assume that all genetic barriers to gene flow more or less coincide in space. We also again assume that an equilibrium between migration and selection has been reached for the selected (but not necessarily the neutral) loci.

Under these conditions, spatial clines form at selected and neutral loci (see Fig. 2.5 for examples from simulations). The steepest clines form for selected and closely linked loci. Each cline is steepest in the centre of the zone and flattens towards the sides.

Quantitative definition of RI under this model

If the population is distributed along a spatial continuum, we cannot unambiguously define two populations between which to determine RI. We could arbitrarily define two

areas and use the effective migration rate as above, but the result would then depend strongly on how the two populations are delimited. It is thus more natural to view gene flow as a diffusion across continuous space (Fisher, 1937; Wright, 1943; Haldane, 1948) rather than as a rate between two arbitrarily defined demes, and to measure reproductive isolation through its effect on the rate of diffusion across the hybrid zone itself.

In continuous space, the rate of diffusion of alleles depends only on the variance of the distance between parent and offspring (Nagylaki, 1976). To fit the general definition given in the Introduction, RI must describe how this diffusion process is impeded by genetic differences between the two sides of the hybrid zone. In a simple model of diffusion, the only factor that affects gene flow between two points is the spatial distance between these points. It is thus most natural to represent RI with the unit of a distance, where a greater distance implies a greater obstacle to gene flow.

For that, one can calculate the spatial distance that would be needed to generate the same allele frequency change if there was no barrier to gene flow. The allele frequency change without a barrier is represented by the gradient of allele frequency (or slope of allele frequency change), p', near the barrier (Fig. 2.5C). If there was no barrier, allele frequency would continue to change as a straight line with that slope. However, the local barrier - whether it be physical or genetic - causes an abrupt step in allele frequency, Δp (Fig. 2.5C). Thus, the strength of the barrier to gene flow can be defined as $B = \Delta p/p'$, i.e. the distance that would be required to generate the same allele frequency change as the step (Fig. 2.5C). Crucially, this ratio quickly settles to a constant value (Nagylaki, 1976). This distance, B (for barrier), for any neutral locus is approximately

$$B = \int ((\frac{\overline{W}(x)}{\overline{W}_0})^{-\frac{1}{r}} - 1)dx \tag{2.6}$$

where r is the harmonic mean recombination rate between the focal neutral locus and all the selected loci (Barton, 1986), and $\overline{W}(x)$ is the mean fitness in the hybrid zone, relative to the mean fitness outside, $\overline{W}_0$ (Barton, 1986). As one moves across the hybrid zone, the relative mean fitness decreases towards the zone centre, and the gradient in the linked neutral cline increases by a factor $(\frac{\overline{W}(x)}{\overline{W}_0})^{-\frac{1}{r}}$. Thus, the barrier, B, is found by integrating the increase in gradient across the hybrid zone. This is an approximation valid in the limit where selection is weak relative to recombination, but remains accurate for moderately strong barriers (provided that the distribution of hybrid index is unimodal; Kruuk et al. (1999); Barton and Shpak (2000)). As it depends on recombination rate, this equation can be applied for linked as well as unlinked neutral loci and can thus be used to calculate genome-wide as well as local RI. Fig. 2.4 shows an example of how B changes along the genome under fixed s. B increases with decreasing hybrid fitness and with proximity of the focal neutral locus to selected loci.

B describes the impediment to movement of neutral alleles from one side of the hybrid zone to the other, given divergent selection at some loci, as a spatial distance. The comparison to the movement without any such barriers is implicit, as in that case the corresponding spatial distance is just 0. B therefore fits our general definition of RI but differs from the definition of RI given for the two-deme model in that it can exceed 1 and has the unit of a spatial distance.

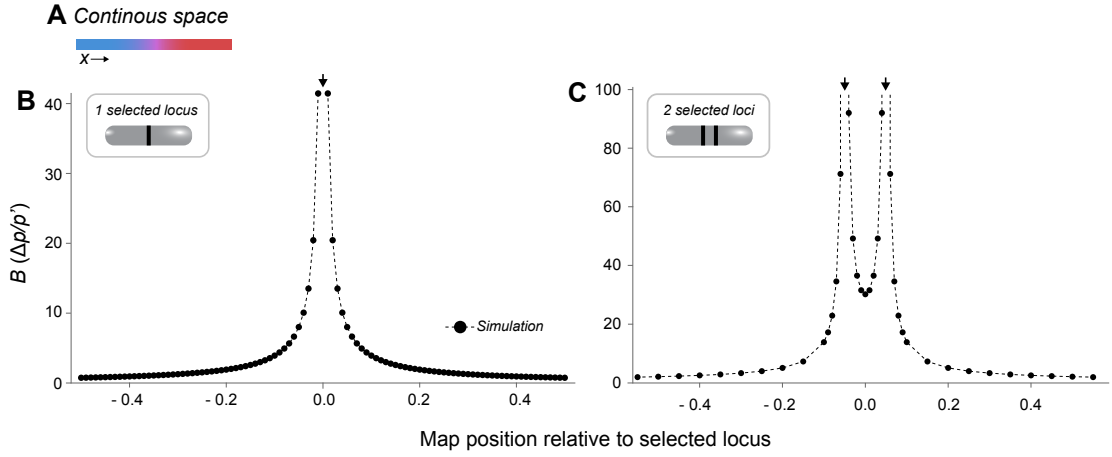


Figure 2.4: **: Reproductive isolation in a continuous-space scenario (across a simple hybrid zone).** A) Two divergent populations meet in a hybrid zone in continuous space. B) and C) The barrier B along the genome under this model, with a single locus under selection at position 0 (B) or two loci under selection at positions -0.5 and 0.5. The x-axis gives the recombination rate of the neutral locus relative to position 0.0 (C). Points show results from deterministic simulations, and the line connects these points (details see A.1). $s = 0.1$ for all selected loci.

RI between two demes vs. across a hybrid zone

What is the relation between measures of RI in a two-deme vs. a hybrid-zone situation, i.e. the relation between $\mathrm{RI}_{2D} = 1 - m_e/m$ and B?

We can determine m_e and relate it to B in a hybrid zone if we consider a zone that is limited in space (i.e. individuals to the left and right of the hybrid zone centre occupy a finite spatial range) and established a long time ago. Under these conditions, allele frequencies for neutral loci on each side of the central step are almost constant due to long-term mixing. The left and right side thus naturally form two areas similar to two demes, which are separated by a barrier to gene flow, and between which an effective migration rate can be determined (see Appendix A.2: Relation between m_e and B in one dimension): $m_e \approx \frac{\sigma^2}{2BX}$, where X is the length of the habitat on the focal side of the zone. Thus, under these conditions, m_e simply decreases inversely with barrier strength.

However, to determine RI analogous to a two-deme model, we not only require m_e but also m for the hybrid zone. How to define m for a hybrid zone is less clear: While two homogeneous areas, similar to two demes, can naturally emerge for neutral loci affected by a barrier to gene flow, this is not the case for neutral loci not affected by a barrier to gene flow (which would be required to determine m). Thus, it is unclear between which two areas m should be defined, and there is no natural direct equivalent to $\mathrm{RI}_{2D} = 1 - m_e/m$ for a hybrid zone (This issue is elaborated in Appendix A.2: Relation between m_e and B in one dimension).

Nevertheless, empirical studies frequently apply two-deme models and assumptions to systems that actually show divergence across a continuous hybrid zone (e.g. studies calculating F_{ST} along the genome for two samples from hybridizing populations). Samples are often taken at a somewhat arbitrary distance from the hybrid zone. A main reason for doing this is that hybrid zone analysis usually requires much larger sample sizes. It is important to notice that in this case, estimates of m_e or RI_{2D} reflect the reduction in effective gene flow between the two sampled areas, which is not necessarily the same as the reduction experienced directly in the hybrid zone. If the samples used are distant from the hybrid zone, it is possible that additional barriers to gene flow located in space between the hybrid zone and the sample contribute to the measure of m_e or RI_{2D}. RI_{2D} thus summarises the effects of all barriers between the two samples, and is not equivalent to B for the hybrid zone.

2.5 Other scenarios

Simulations can be used to determine RI in a multitude of scenarios Above we have discussed quantitative definitions of RI under two relatively simple scenarios, making simplifying assumptions. These assumptions may be violated in many empirical settings. The exact effects of these violations on RI can in some cases be explored analytically (e.g. effects of > 2 demes; see below). In other cases, this might not be currently easy (e.g. a two-deme scenario with high migration; see below). However, importantly, RI can be evaluated in any scenario that can be simulated. For discrete demes, both the change in allele frequency at a focal neutral locus in a focal deme (δp) and the allele frequency difference between demes (Δp) can be recorded in simulations, and provide $m_e = \delta p/\Delta p$ (Fig. 2.5A). As the migration rate m is set for the simulation, RI can be calculated. In continuous space, for a hybrid zone situation, B can be measured as $B = \Delta p/p'$ (Fig. 2.5B,C). In more complex settings in continuous space, one can arbitrarily define two areas between which RI can again be measured using δp and Δp.

We cannot cover all deviations from the simple two-deme and hybrid zone models in this paper. However, in the following, we discuss the effects of deviations that may be

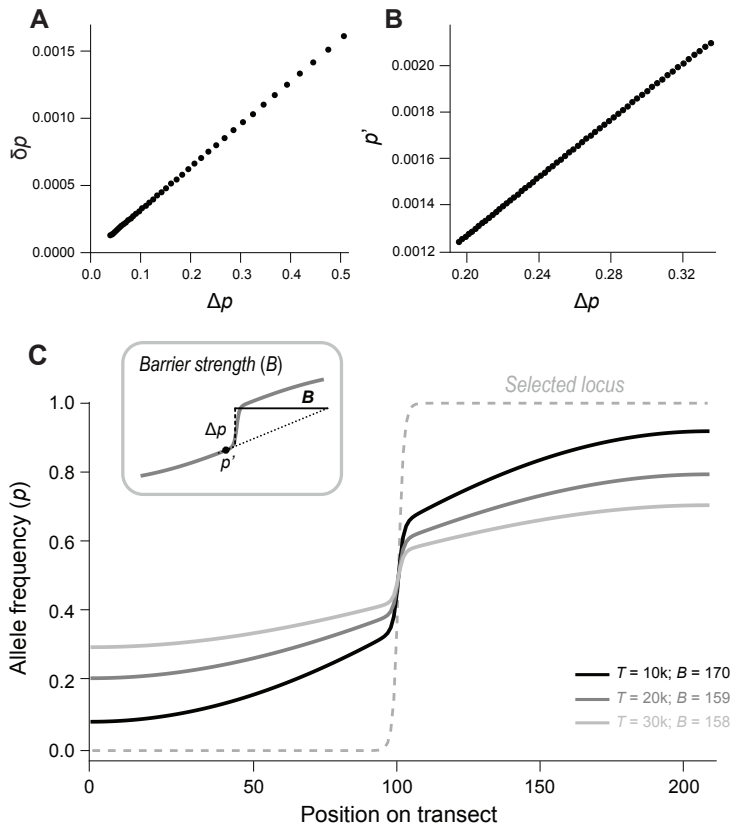


Figure 2.5: **Estimating RI from simulations.** A) Discrete demes: In contrast to most empirical situations, in simulations it is possible to observe not only the allele frequency differences between two demes in a given generation (Δp) but also the allele frequency change over a generation (δp). As the ratio between the two gives m_e, the slope of δp against Δp provides m_e. The different values in the plot can come either from following a single neutral locus over time or from looking at multiple loci with different allele frequencies across a single generation; but note that the linear relationship between δp and Δp from different time points will only appear when the selected loci are at equilibrium for the selected loci; under non-equilibrium, the time point to determine RI must be specified and all values must be sampled from that same time point. In either case, RI can then simply be calculated from m_e and the known m. B) Hybrid zone: The barrier B for a neutral locus can be measured as the ratio between the gradient (p') and the central step (Δp) (see panel C), which also stays constant over time. Again, the different values in the plot can come either from following a single neutral locus over time (as in panel C) or from looking at multiple loci with different allele frequencies in a single generation. C) Hybrid zone: RI can be measured from cline patterns. At the selected locus (grey dotted line), a steep spatial cline has reached equilibrium. At linked neutral loci, the barrier to gene flow generates weaker clines. We show a single linked neutral locus at three time points (different shades of grey). After an initial short period of stabilization (not shown), even though neutral allele frequencies are still changing, the barrier B measured for the linked neutral locus is approximately stable over time, reflecting a stable rate of flow across the zone. Simulations with nearest neighbour migration with $m = 0.5, s = 0.2$ for the selected locus and $r = 0.01$ for the neutral locus.

particularly common in nature.

Range overlap

The two focal populations may occupy overlapping ranges, living in (partial) sympatry. For this situation to be stable, there must be genetic differences that prevent the populations from mixing, and consequently, some degree of reproductive isolation. The effective migration rate and RI can be defined in the same way as for two demes in different locations. In Box 2, we consider this situation in more detail, and in particular, whether the fraction of range (non)overlap is an appropriate measure of RI.

Two-deme models with higher gene flow

If levels of gene flow between two demes are high (as opposed to our assumptions above), F_2 and complex hybrids may be common, and the outcome is unpredictable. We cannot simply focus on the fitness of successive backcrosses, but need to consider the fitnesses of all possible hybrid genotypes and their respective frequencies, making the calculation more complicated. There is also a more fundamental problem: with very high levels of gene flow, what are the two populations between which we want to measure RI? If most individuals cannot clearly be assigned to a population, a measure based on two distinct groups may not be appropriate.

Gene flow between two demes within a larger set of demes

If we define a set of demes each with its own set of selection coefficients and connected by a set of migration rates, we can still define the effective migration rate. For that, the analytical approach proposed by Barton and Bengtsson (1986) splits all possible individuals into a set of "pools", where each pool is a combination of a certain deme with a certain genetic background (genetic background = genotype at selected loci). Then, one can generate a (potentially very large) matrix that describes the gene flow between each possible pair of such pools for a neutral locus; the matrix values are dependent on m, the selection coefficient on each genotype in each given deme, and the recombination rates. m_e between pairs of demes at equilibrium can then be calculated from this matrix.

Non-equilibrium situations

We have assumed the selected loci to be at equilibrium (or fixed different) when defining RI in specific situations above. In that case, RI is stable over time, even though allele frequencies at neutral loci might change – their long-term rates of gene flow are constant, and thus RI is constant as well.

If selected loci are not at equilibrium, RI will change over time. A simple example to see this is a secondary contact between two demes containing multiple intrinsic incompatibilities of the form AAbb in deme 1 and aaBB in deme 2, with A and B being incompatible but otherwise not selectively different from a and b. These incompatibilities can be resolved by fixing the ancestral alleles (a and b) in both populations. Thus, immediately upon secondary contact, there might be a strong barrier to gene flow, but over time this barrier will dissolve.

Note, again, that it is only relevant whether the loci contributing directly to the barrier have reached equilibrium. It is irrelevant whether neutral loci are at equilibrium, as m_e is independent of the frequency of the neutral alleles; for example, m_e for an unlinked neutral locus is the same independent of whether that locus has very recently obtained a new mutation in deme 1 or shows a large allele frequency difference between deme 1 and deme 2.

In non-equilibrium cases, it is important to define the time point/time interval at which we want to determine RI. One can then (at least in principle) calculate or simulate the migration matrix as it changes through time, and thus also calculate how RI changes over time. The analytical approach proposed by Barton and Bengtsson (1986) described above can be used to calculate m_e between demes for any given time point. However, importantly, in a non-equilibrium situation, the frequency of different genetic backgrounds (for selected loci) changes from generation to generation, and therefore this matrix changes over time as well, making the calculation more complicated.

Heterosis and adaptive introgression

Frequently, F_1 hybrids are fitter than either parent. In the short term, such heterosis will increase the effective migration rate, by e.g. increasing $\overline{W}_1$ in Eq. 2.2 (which still applies). However, if the heterosis is due to the presence of different deleterious recessives in the two populations, then it will be transient, since gene flow and selection will together tend to equalise allele frequencies. This is a special case of a more general phenomenon: when an allele that is favoured in both populations sweeps across, it will take with it a surrounding block of genome, thereby increasing neutral gene flow rather than reducing it, as discussed in most of this article. RI can in these cases be negative. If there is recurrent selection at the same locus, there may be a long-term increase in gene flow, and a consistent reduction in RI. One example of this phenomenon may be the frequent observation of introgression of mitochondrial DNA in animals, which inflates the flow of all the alleles carried on these maternally inherited organelles (Toews and Brelsford, 2012; Sloan et al., 2017).

Effect of a physical barrier

As mentioned above, it is important to note that a physical barrier to gene flow (e.g. a river, wall or mountain, or a local reduction in population density due to e.g. less available habitat) does not directly contribute to RI, as it is not genetically based. However, it does restrict gene flow. In a two-deme setting, a physical barrier between demes just reduces m. In continuous space, the situation is more complicated, because the dispersal rate is only reduced in a specific area. In that case it is possible to calculate a B_{phys} for a physical barrier analogous to that for a genetic barrier. For example, a local reduction in population density, ρ, or dispersal, σ^2, will generate a physical barrier:

$$B_{phys} = \int (\frac{\rho_0^2 \sigma_0^2}{\rho^2 \sigma^2} - 1) dx \tag{2.7}$$

(Barton, 1986) (Appendix A.2: Effect of a physical barrier). ρ_0, σ_0^2 are the density and dispersal outside the barrier, and ρ, σ are reduced below ρ_0, σ_0^2 within some local region.

Importantly, even though a physical barrier itself is not part of RI, co-located physical and genetic barriers interact to together form a stronger barrier than would be higher than expected from just adding up the two barriers (i.e. two values of B) (Fig. 2.6).

2.6 Estimating RI from empirical data

Above, we have provided definitions of RI in different spatial settings. In the empirical literature, there are various approaches for estimating RI. They differ particularly in whether they focus on the genetic or organismal level, and on whether they generate a single estimate of RI (genome-wide RI) vs. multiple estimates along the genome (local RI). In the following, we summarise these approaches, discuss how they relate to the above definitions and to each other, and discuss under what conditions they measure RI, as we have defined it.

As discussed above, in non-equilibrium situations (for selected loci) the value of RI will depend on the timescale considered. Different measures of RI implicitly focus on different timescales, and will thus often provide different estimates. Organismal methods, based on immigrant fitness and the fitness of the first few hybrid generations, measure RI on very short timescales. Hybrid zone analysis applies in a limited spatial context, where selected and neutral loci typically equilibrate quickly, and thus reflects processes on a timescale of hundreds or thousands of generations. Methods based on sequence divergence can reflect processes over much longer timescales, of order the effective population size of the whole species.

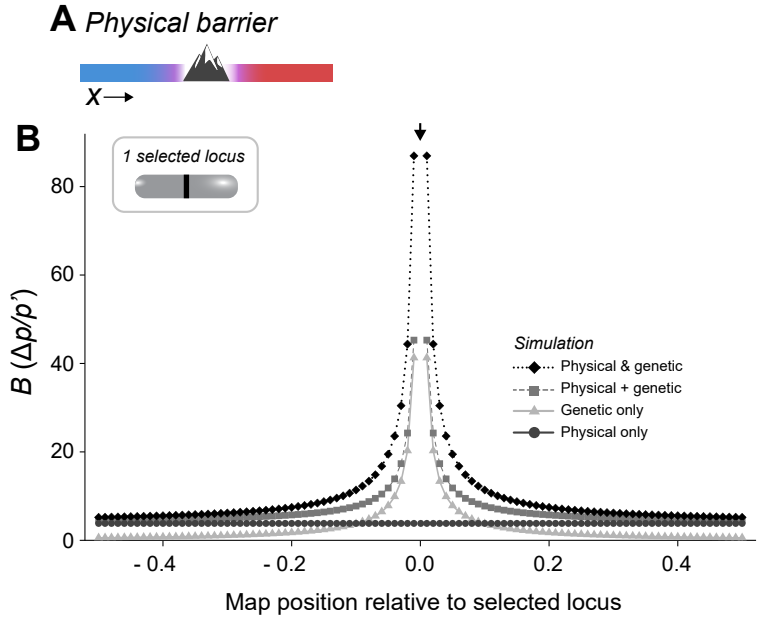


Figure 2.6: **Reproductive isolation in continuous space, with a physical barrier.** A) Two divergent populations meet in a hybrid zone in continuous space; the genetic barrier coincides with a physical dispersal barrier. B) The barrier B along the genome under this model, comparing different situations with and without physical and genetic barrier. Circles: Only a physical barrier, with no selection: Triangles: Only a genetic barrier; Diamonds: the genetic and the physical barrier acting together; Squares: For comparison, the (hypothetical) barrier that would appear if the genetic and the physical barrier just added up. This plot shows the synergy between the physical and genetic barrier. The x-axis gives the recombination rate of the neutral locus relative to position 0.0 (which corresponds to the position of the selected locus with $s = 0.1$, if present). See A.1 for details.

Organismal measures of RI

Several measures of RI focus on the organismal level, not using genetic data to estimate gene flow. The most direct way to measure RI at this level is to estimate the fitness of migrants and hybrids experimentally or in the field. We discuss above how these fitnesses $(\overline{W}_0, \overline{W}_1, \dots)$ relate to genome-wide RI (Eq. 2.2, RI for an unlinked neutral locus in a two-deme model under low gene flow). In a two-deme situation, one would ideally determine the fitness of immigrants, F_1 hybrids and backcrosses in each deme. It might be difficult to do this in laboratory experiments, as it will usually be hard or impossible to include all components of fitness. It might be more promising to use field data: using genetic markers, individuals in a deme can be assigned to the migrant, F_1 or successive backcross generations (in practice, this works only for the first few generations, but the genome-wide barrier depends mostly on these). This categorization can be performed in unmanipulated natural populations or after transplanting a set of "migrants" into the recipient population and tracking their offspring. The offspring numbers for the migrant,

F_1 and backcross individuals can be determined (e.g. by counting eggs or using pedigreed populations) and compared to that for the recipient population, thus producing a measure of relative fitness for each generation. If sample sizes are large, one could alternatively use the migrant and hybrid counts directly as a measure of fitness of the previous generation (However, this requires obtaining migrant numbers before selection). These approaches reflect RI over short timescales (only over the studied set of generations) and should provide more direct estimates than those from sequence divergence discussed below. Note that this approach does use genetic data to assign individuals to hybrid categories, but it uses a direct estimate of fitness (offspring number) to calculate RI. A shortcoming of this approach is that large numbers of genotyped individuals are necessary if hybridization is rare.

Previous work on organismal measures of RI has used definitions of RI deviating from what we propose here, and has focused mostly on measuring different barriers (e.g. pre- and postzygotic barriers) separately. Coyne and Orr (1989) introduced a measure of RI in their classic meta-analysis of Drosophila crosses, based on mating preferences and the viability and sterility of F_1 hybrids. For example, their measure of prezygotic isolation, based on lab mating experiments, was $RI_{pre,CO} = 1-$(frequency of heterospecific matings)/(frequency of homospecific matings). If (frequency of heterospecific matings)/(frequency of homospecific matings) was measuring the fitness of immigrants, this would be equivalent to 1- $\overline{W}_0$ in Eq. 2.2, above. However, this will usually not be the case. For example, this quantity (obtained from mating experiments) may often underestimate the mating component of immigrant fitness. For example, even if the probability of heterospecific matings in lab crosses is low, in the field there might be so many mating opportunities that almost all immigrants will eventually find a mate; then, at least female fitness might not be much reduced. The probability of mating in a single (lab) encounter thus may not reflect the long-term reproductive success of an individual. It might be possible to obtain more realistic measures of the mating component of fitness by designing mating experiments closely mimicking field conditions. However, Coyne and Orr (1989)'s measure also differs from the measure of RI proposed here in that it focuses on specific barriers. Combining multiple barriers mathematically is straightforward in principle, but large numbers of experiments may be required to include all barriers. In addition, this measure does not include the fitnesses of the following hybrid generations. Both would be required for the measure of RI to reflect gene flow.

A main problem with the measure proposed by Coyne and Orr (1989), and other similar measures, is that it uses mating experiments/crosses between individuals that might be sampled distant from the area of hybridization and would never encounter each other in nature. This is problematic e.g. for taxa forming hybrid zones because barriers to gene flow can be resolved by displacing the clines of the contributing loci. For example, this

is expected if divergence involves Dobzhansky-Muller incompatibilities (Orr, 1996). If a derived allele A is favoured at locus A in population 1, and a derived allele B is favoured at locus B in population 2, and alleles A and B are incompatible, then the ancestral alleles a and b will become common in the hybrid zone centre, alleviating incompatibility. There are several such examples in natural hybrid zones (e.g. Hatfield et al. (1992); Virdee and Hewitt (1994)). Clines at loci A and B will be pushed apart in space, so that there is no barrier to gene flow because the incompatible alleles (A and B) do not meet. Measures of m_e based on genomic data (see below) will thus not indicate a genetic barrier to gene flow. However, when individuals sampled distant from the hybrid zone are crossed, they will contain incompatible alleles and give the misleading impression that there is a barrier to gene flow. While it is correct that an incompatibility exists, it is not relevant in the natural context and should thus not be incorporated in a measure of RI.

Sobel and Chen (2014) discuss alternative measures of RI, which are mostly identical or similar to Coyne and Orr (1989) measure (see Sobel and Streisfeld (2015)). They aim for a linear relationship between their measure and $H/(C+H)$, which they consider the "probability of gene flow". This probability is not the rate of gene flow into a deme but the probability of mating in a symmetrical contest, e.g. for the parental taxa when they fully coexist in sympatry. Sobel and Chen (2014) thus propose $RI_{SC} = (1 - H/C)/(1 + H/C)$, where H and C are the numbers of heterospecific and conspecific matings, respectively. They also suggest methods to calculate the total barrier by combining the effects of different barriers. Postzygotic isolation is included by setting H to the fitness of the offspring of heterospecific crosses and C to the fitness of the offspring of conspecific crosses. Similar to Coyne and Orr (1989)'s measure, Sobel and Chen (2014)'s measure of RI is usually calculated from lab mating experiments under artificial settings, and thus may not directly reflect fitness components or the probability of gene flow under field conditions, except under very specific assumptions. It also mostly focuses on mating between the parental taxa, while RI must depend also on the fitness of the following generations. For example, even if individuals of two taxa readily mate and produce fertile F_1 offspring, many F_2 or backcrosses might be sterile.

In summary, as the organismal methods estimate migrant and hybrid fitnesses, genome-wide RI can in principle be estimated directly from organismal measures by multiplying across independent fitness components and across generations (Eq. 2.2). This requires taking multiple generations and barriers into account; it is also important that lab mating probabilities are not equated to fitness components in the field unless the experiment is specifically designed towards this goal or an estimate of the fitness components can be calculated from the mating data (see Perini et al. (2020) for an example in this direction). It is also important to note that hybrid and backcross fitness measured in the lab might differ from the field because deleterious alleles might be purged more effectively in the

field (i.e. increasing $\overline{W}_1, \overline{W}_2, \ldots$ in the field); however, this error is small unless selection is very strong (see Example scenario 1 and Appendix A.2: Flow into a single deme).

Another general limitation of organismal approaches is that they do not explicitly consider the genomic locations of selected and neutral loci; i.e. they cannot measure variation of RI along the genome. Moreover, under high gene flow or in continuous space (for example, across hybrid zones) there are a multitude of fitness classes (each possible genotype at selected loci in each environment might have a different fitness), making calculation of their effect on the net rate of gene flow very complicated; in such cases, the fitnesses of all hybrid genotypes cannot in practice be measured. We expect organismal methods to have the potential to provide reliable RI estimates only in situations with populations connected by low levels of gene flow, either because of a large distance or physical barrier to gene flow, or because strong RI has already evolved and limits exchange.

In summary, as the organismal methods estimate migrant and hybrid fitnesses, genome-wide RI can in principle be estimated directly from organismal measures by multiplying across independent fitness components and across generations (Eq. 2.2). This is only appropriate in cases where gene flow is low. It requires taking multiple generations and barriers into account; it is also important that lab mating probabilities are not equated to fitness components in the field unless the experiment is specifically designed towards this goal or an estimate of the fitness components can be calculated from the mating data (see Perini et al. (2020) for an example in this direction). In addition, hybrid and backcross fitness measured in the lab might differ from the field because deleterious alleles might be purged more effectively in the field (i.e. increasing $\overline{W_2}, \overline{W_3}, \ldots$ in the field); however, this error is small unless selection is very strong (see Section 2.3 and Appendix A.2: Flow into a single deme, Unlinked loci).

A great advantage of organismal approaches over genetic ones is the opportunity to explore the actual barriers contributing to RI and make the distinction between pre- and postzygotic or intrinsic and extrinsic barriers. Genomic data alone are often difficult to interpret, and approaches like F_{ST} scans can frequently generate signals of "barriers" that in fact have nothing to do with RI.

Hybrid zone analysis (Geographic cline analysis)

Unlike the above methods, which often drastically simplify or ignore the geographic setting, hybrid zone analysis uses spatial patterns of genetic variation to quantify the strength of barriers. In example scenario 2, we showed how a genetic barrier impedes the flow of neutral alleles through continuous space. This leads to the formation of a cline with a sharp central step of allele frequencies flanked by relatively shallow tails of introgression (examples in Fig. 2.5). This pattern is seen in many natural hybrid zones,

and has been used to estimate the barrier strength, B, for neutral loci linked and unlinked to selected loci, as described above ($B = \Delta p/p'$).

To calculate B in practice, one needs large sample sizes and dense spatial coverage to resolve the allele frequency gradients at the centre and edges of the zone. The size of the central step (Δp) and the slope of the flanking gradients (p') must then be estimated. This is difficult to do directly, so a three-part 'stepped' cline model (left gradient – step – right gradient) is usually fit to the data (Szymura and Barton, 1986; Porter et al., 1997). Because the left and right tails can be approximated by separate functions, B can be estimated separately for each side of the hybrid zone to quantify asymmetry in gene flow.

As cline shapes for selected loci quickly equilibrate when selection changes (on a timescale of $\sim 1/s$), and on a local scale neutral alleles quickly respond to changes, B reflects gene flow typically over a timescale of hundreds to thousands of generations.

The interpretation of B may be difficult in practice because physical barriers have the same effect on patterns of gene flow as genetic ones (see above), so it is necessary to rule out or correct for their effect. Physical barriers are not always readily apparent (e.g., a mountain or river) and may include more subtle environmental features that make the landscape less permeable to dispersal or harder to inhabit (e.g., differences in soil chemistry, or water salinity). Mapping of the population density and habitat variables may help reveal physical barriers (Hewitt, 1988), but density may also be reduced at a genetic barrier if selection against hybrids is strong (i.e., 'hybrid sink' effect; Barton (1980)). This makes it difficult to disentangle the effects of physical and genetic barriers, but conclusions may be strengthened by other lines of evidence, including inferences from multiple independent transects (e.g. Szymura and Barton (1991)) or direct measurements of dispersal (Barton and Gale (1993)).

Despite its obvious strengths, geographic cline analysis also has some innate limitations. The most obvious is that it cannot be applied to organisms that do not hybridize in nature, or those that form complex mosaic hybrid zones that do not show a smooth one-dimensional cline (e.g. Bierne et al. (2003)). Another practical challenge is that cline fitting is statistically delicate, and application to genome-wide datasets is challenging. Moreover, cline shapes are not only affected by dispersal and selection, but also by genetic drift (Polechová and Barton, 2011); there is little understanding of how this affects estimates of B especially in small populations. Finally, B, calculated in the absence of other methods does not tell us anything about the types of barriers that cause RI.

Inferences about the strength of RI have also been made from the distributions of hybrid index (HI) scores from individuals sampled at the centre of hybrid zones (Jiggins and Mallet, 2000; Irwin, 2020). Because HI scores are calculated from numerous unlinked loci that diagnose different taxa, the distribution of scores in areas of overlap, which summarise

multilocus patterns of LD, must reflect the historical local rate of production and fitness of hybrids. For example, complete isolation in sympatry will maintain maximum LD among loci, so that HI scores remain perfectly bimodal. In contrast, the absence of any barriers will cause LD to decay, resulting in a unimodal distribution. Partial isolation is expected to result in a distribution somewhere in between. However, hybrid zones are usually just classified as either "unimodal" or "bimodal", thus not producing a quantitative measure of RI, and the genome-wide barrier to gene flow has to be very strong to maintain a bimodal distribution.

Using sequence divergence to estimate RI

There is a plethora of methods for estimating rates of gene flow from genetic data, either applied to regions of genome or to the genome as a whole. However, RI is defined as the reduction in gene flow due to genetic differences. Therefore, to estimate RI, it would be necessary to obtain measures of both m_e (either along the genome or for an unlinked neutral locus) and m.

Wright introduced the classic statistic, F_{ST}, which measures allele frequency differences between populations. Assuming an infinite island model, Wright suggested using F_{ST} to estimate the number of migrants between demes, $N_e m$, at an equilibrium between drift and gene flow, $F_{ST} = 1/(1 + 4N_e m)$. When applied to empirical data, F_{ST} reflects the migration actually experienced by the loci analysed (rather than the raw migration rate), and thus in most cases would more accurately be described as $F_{ST} = 1/(1 + 4N_e m_e)$.

Partly for that reason, F_{ST} varies along the genome, and a main premise of speciation genomics has been that genomic regions containing barrier loci ("genomic islands of divergence") can be discovered by "scanning" genomes for high-F_{ST} windows (Ravinet et al., 2017).

However, interpreting F_{ST} is difficult. F_{ST} is a relative measure, which can be inflated by reductions in genetic diversity due to selective sweeps or background selection, independent of gene flow. Indeed, in some systems “islands” appear to have been shaped by these processes (Cruickshank and Hahn, 2014; Burri et al., 2015; Chase et al., 2021). There is also wide variation in F_{ST} even in the absence of such causes, due to the fundamental randomness of the evolutionary process, making it difficult to reliably detect local barriers in the genome. Despite these difficulties, F_{ST} can be a useful indicator of recent divergence, but can only be taken as evidence for reduced effective gene flow across the genome after correction for confounding factors and when combined with other evidence, e.g. from experiments.

An alternative genome scan, complementary to F_{ST}, measures the admixture proportion, f_d. Like other D-statistics, f_d measures introgression from the excess of shared derived sites

in a four-taxon framework, but is modified for application to genomic windows (Martin et al., 2015). The underlying assumption is that 'ABBA' and 'BABA' site patterns should be equally frequent in the genome when sharing of the derived allele (B) results from random sorting or recurrent mutations. Gene flow, on the other hand, creates an excess of one site pattern, which can be used to identify and quantify introgression.

Estimates of f_d are roughly proportional to admixture for small simulated genomic windows (Martin et al. (2015) p. 201), so one would expect scans of f_d to be correlated with variation in m_e across the genome (Martin et al., 2015, 2019). Also, unlike F_{ST}, f_d is robust to variation in nucleotide diversity across the genome, so should largely be unaffected by sweeps or background selection. However, f_d is slightly biased toward regions with low between population divergence (low d_{xy} or shallow between-species coalescence), so scans with multiple statistics may give a clearer picture (i.e., F_{ST} and f_d).

Alternatives to genome scans circumvent some of the problems listed above. For example, Aeschbacher et al. (2017) use a genome-wide pattern rather than focusing on small windows, potentially increasing power. They use the genome-wide negative correlation between recombination rate and population divergence found under highly polygenic RI, which appears because higher recombination rates allow neutral loci to decouple from divergently selected loci and hence introgress. Using coalescent theory, Aeschbacher et al. (2017) fit a model of divergence with gene flow to empirical divergence data and generate estimates for the selection density (the product of the mean selection coefficient and the density of selected sites) and the baseline migration rate (m). Because they specify how m_e is determined by the selection density per map length, r and m, it is possible to calculate m_e, and therefore RI, as a function of the recombination rate. This method thus estimates how m_e generally depends on the recombination rate in the focal system, but it does not actually directly give measures of local RI along the genome. This approach aggregates information across the whole genome and does include a correction for background selection, but nevertheless depends strongly on the demographic model and on the assumed spatial structure.

Various methods for modelling demographic history (including gene flow) have become popular, e.g. those based on the site frequency spectrum (dadi; Gutenkunst et al. (2010)) or summary statistics in an ABC framework (Beaumont et al., 2002). These approaches do not rely on equilibrium or infinite islands assumptions. Focusing on putatively neutral markers and neutral demographic processes, they are often treated as fundamentally separate from genome scan methods (focused on finding selected loci). Demographic modelling has been used to obtain estimates of m, but some recent approaches have also explicitly considered the fact that not all loci are neutral, and fit a distribution of m_e rather than a single value while also taking other potentially confounding processes into account (e.g. pervasive background selection) (Rougemont et al., 2017). Emerging

methods aim to explicitly characterise variation in m_e across the genome by fitting separate demographic models to defined blocks of sequence (Laetsch et al., 2023). However, like genome scans, these approaches try to estimate gene flow from genomic data affected by various processes, and failing to include those in the model might lead to serious errors when estimating gene flow (e.g. Momigliano et al. (2021)).

Importantly, all approaches listed in this section suffer from the same fundamental problem: They do not generate the estimate of m needed to calculate RI. Even though some methods aim to estimate m, they instead estimate m_e, the gene flow actually experienced by the loci analysed. We have seen above that most RI may be due to the aggregate genome-wide effect of divergent selection, i.e. a general barrier to gene flow that affects the whole genome. This means that any estimate of gene flow obtained from genomic data, even from a genomic region distant from any strongly selected locus, includes the effect of this genome-wide barrier, and is thus an estimate of m_e that is lower than m. Genome scans alone, for example, can therefore potentially identify genomic regions where m_e is reduced on top of the effect of the genome-wide barrier to gene flow, but they cannot be used to find m and so generate an RI estimate that is comparable among different systems. Similarly, demographic modelling that aims to estimate m cannot distinguish to what extent a limitation in gene flow is due to physical versus a genetic barrier.

Here, it becomes important to combine genomic data with data closer to those obtained by the "organismal" methods described above. For example, migration rates can be estimated directly with mark-recapture experiments or using observations of dispersal in pedigreed populations. However, it needs to be noted that these organismal methods reflect m on the timescale of a few generations, while genomic estimates of m_e from e.g. IM model fitting reflect long-term gene flow over thousands of generations. Therefore, in non-equilibrium scenarios the combination of these estimates for m and m_e cannot lead to reliable estimates of RI.

A final point to note for the methods covered in this section is that these approaches (except, maybe, for demographic analyses explicitly taking into account hybrid zone settings) should be used with caution when the study system forms a hybrid zone rather than more or less discrete populations. As we have discussed above, reproductive barriers play out differently in continuous space, m_e has no clear definition in continuous space, and assumptions of the models underlying two-deme approaches may be violated in continuous space.

Conclusions about estimating RI

Estimating RI from empirical data is challenging. Methods determining short-term RI based on the fitnesses of migrants, hybrids and backcrosses in the field might be most

promising for spatially discrete populations or sympatric taxa with low levels of gene flow. These approaches are limited to estimating genome-wide RI. However, they could be extended to understand local RI by observing how blocks of introgression are distributed across the genome (e.g.Petr et al. (2019)): In genomic regions with higher RI, introgression is expected to be reduced. However, testing whether there is significant variation in introgression is challenging, since the process is highly random.

In continuous space, hybrid zone analysis is promising if there is detailed sampling of spatial clines. Genomic data used in e.g. genome scans certainly often reflect RI to some extent, but a major challenge here is to disentangle m and m_e. While non-genetic data on migration rates (e.g. mark-recapture experiments) might be useful to some extent, they reflect m over much shorter timescales than the genomic data.

RI estimators based on mating experiments or crosses are unlikely to reflect RI as defined here. For example, experimental estimates of reproductive barriers made in one or a few generations may be high, but these barriers may not be very isolating over timescales that are more relevant to gene flow. However, experiments with organisms are necessary to determine the barriers that reduce gene flow, and careful observations of morphology and behaviour are often necessary to define the groups between which to measure RI in the first place.

2.7 Why should we care about RI?

Reproductive isolation has received much attention because of its central importance to the biological species concept. We define RI in terms of the effect of genetic differences on gene flow. We define RI only for neutral loci in order to separate it from the idiosyncratic effects of selection on particular alleles. However, in most studies of adaptation and speciation, we are interested primarily in traits and loci under selection. Why should we care about a quantity that is only defined for neutral alleles?

RI compares the actual migration rate, m, and the effective migration rate, m_e. Knowing the m_e for an unlinked neutral locus is useful in itself, since it estimates the realised background rate of gene flow, which is relevant not only for neutral, but also for selected loci. Local adaptation can be maintained if divergent selection is stronger than the effective migration rate into a deme, or if the area under divergent selection is sufficiently large, and the barrier sufficiently strong (Piálek and Barton, 1997; Turelli and Barton, 2017). The role of a reduced effective migration rate (due to a genetically based genome-wide barrier) for the accumulation of divergence at further small-effect loci has been highlighted in the discussion about "genomic hitchhiking" (e.g. Flaxman et al. (2013)). Knowing m_e may also be important for predicting the spread of universally favoured alleles, e.g. herbicide resistance between different populations or species of weeds. However, favourable alleles

will relatively quickly penetrate all but the very strongest barriers (See Appendix A.2: Consequences of barriers, for a detailed summary).

Thus, m_e is useful and sufficient if we want to predict how neutral, locally or universally adaptive alleles flow between different demes on the short term. However, in speciation research we are not only interested in how and to what extent gene flow is limited between groups of individuals, but in the extent to which this limitation is caused by inherent differences between these groups, and in how such differences allow them to coexist in sympatry without collapsing. To measure this intrinsic component specifically, m_e is not adequate, as it reflects both limits to the baseline migration rate and the effects of genetic barriers to gene flow. We need both m and m_e, as only the difference between the two is caused by genetic differences. This is why we need to measure RI. We can then relate RI to other features of the system to better understand the processes and barriers that contribute to speciation.

For example, some organismal approaches measure the extent of assortative mating. To understand whether and how mate choice contributes to speciation, we then need to ask: Does assortative mating substantially reduce gene flow? To answer that question, we need to determine RI with and without assortative mating (which could be done in experimental populations or using simulations). On the other hand, if we observe strong RI for some taxon pair, we can ask: which barriers to gene flow do we find on the organismal level, and are these barriers sufficient to explain such a high level of RI?

This idea of using measures of RI to understand the speciation process is represented in the concept of a "speciation continuum". This concept relies on using contemporary populations, varying in their level of RI, to reconstruct the speciation process (Stankowski and Ravinet, 2021). As Stankowski and Ravinet (2021) point out, this approach may be flawed because different contemporary taxon pairs may have followed very different evolutionary trajectories, not representing the same single speciation process. However, comparative analyses can allow us to identify different factors that vary with, and potentially cause, variation in the strength of RI. In the present article, it also becomes clear that the comparative measures of RI necessary for this approach may be difficult to obtain. Most genomic measures of RI will be influenced by differences in the history and spatial situation of the individual taxon pairs. Again, field studies across multiple hybrid generations and hybrid zone analysis might be the most promising ways to infer RI.

RI, even if reliably measured, is not sufficient to predict coexistence in sympatry. While some RI is necessary for maintaining sets of adaptive alleles without being broken up by recombination, long-term full sympatry requires ecological divergence. Even if hybrids are completely inviable, cross-mating will be more costly to the rarer population and may prevent coexistence. Here, it again becomes apparent that a focus solely on genetic patterns and processes is not sufficient, and that ecological processes need to be considered

to comprehensively understand speciation.

2.8 Conclusions

This article is our attempt to clarify several key issues surrounding reproductive isolation, including what RI is, how it can be quantified in principle, and how can it be measured in practice. We define RI based on the reduction in gene flow between populations that is due to genetic differences. We have shown that RI depends strongly on circumstances, including the spatial, temporal and genomic context. This makes it difficult to quantify RI in a way that will be directly comparable across systems. After reviewing methods for estimating it from empirical data, we conclude that it is difficult to measure RI in practice. All existing methods have shortcomings and assumptions that will limit their applicability and accuracy.

A main issue is that existing definitions and measures of RI, including those we prefer, only apply in situations where most of the individuals can be assigned to one of two (or more) distinct populations, without the occurrence of complex hybrids across large spatial areas. In some taxa, hybridisation is pervasive and it is impossible to identify distinct groups between which RI can be measured. In such cases, it is unclear what we would want to measure in the first place. Future work should develop concepts and measures for such scenarios.

While these messages may seem overly negative, we emphasise that in many systems RI is useful and necessary for quantifying the evolutionary independence of populations. While not perfect, existing methods, especially when combined and interpreted with appropriate caution, can give insight into the extent to which populations evolve independently and the underlying barriers to gene flow. Looking to the future, we encourage researchers to explore new, creative approaches to estimating RI in the field, taking advantage of all of the available data and combining measures of RI with experiments and field data on the different contributing barriers.

BOX 1: Origin and meaning of 'reproductive isolation'

The term 'reproductive isolation' first appeared in the 1930s (Emerson, 1935), but the idea of species as reproductive communities can be traced back to Linnaeus' emphasis on reproductive organs in taxonomic classification. Even during Darwin's time, some biologists argued that species should be distinguished from races by their inability to produce fertile offspring (Huxley, 1860). For example, Wallace (1865) stated that 'species are merely those strongly marked races or local forms which, when in contact, do not intermix, and when inhabiting distinct areas are generally believed... to be incapable of producing fertile hybrid offspring'. Around the turn

of the century, Poulton (1904) laid out a verbal theory for how interspecific sterility might evolve through the cessation of interbreeding between groups (i.e. 'asyngamy') owing to geographic isolation, mechanical incompatibilities, or preferential mating. As highlighted by Mallet (Mallet, 2004b,a), this (largely overlooked) work laid the foundation for modern speciation research, ultimately leading to the widespread focus on RI that emerged in the mid to late 20th century (Fig. 2.7A,B).

Although RI became central to the work of Mayr and Dobzhansky in the 1940s, clear definitions did not emerge until the 1950s. Dobzhansky (1937) stated that RI exists between populations when "the gene exchange between species is restricted or suppressed owing to genotypically conditioned differences between their populations". He also coined the term 'isolating mechanism' to refer to the properties or organisms that may cause RI (1937, 1951). Mayr (1959) stated that reproductive isolation is "...what we might call the protective devices of a well-integrated and harmoniously coadapted gene pool against pollution by other gene pools". While very similar, these definitions seem to differ in regard to the precise meaning of the term. In his writing, Mayr (1942, 1959, 1963) tended to emphasize the organismal traits that restrict reproduction (and thus gene flow) between groups of organisms, whereas Dobzhansky emphasised that RI was the reduction in gene flow itself.

Focusing on other influential papers speciation research, we found that later definitions, discussions and studies of RI varied similarly in whether they emphasise patterns of reproduction between organisms (organismal focus) or levels of gene flow between populations (genetic focus) (Table S1 2.8). However, despite being widely used in the literature, RI is usually not specifically defined, making it difficult to gauge how widespread these different views are. To address this, we turned to a recently published survey to gain additional insight (Stankowski and Ravinet, 2021). Survey question 12 asked: 'In a sentence or two, what is reproductive isolation?'. The answers from 231 speciation researchers (Table S2, 2.8) were variable, but we could classify most based on whether they mentioned (i) patterns of interbreeding, (ii) levels of gene flow, (iii) the ability of populations to remain distinct, or some combination of the three (Fig. 2.7C). Forty-two percent of answers had a purely organismal focus, mentioning only patterns of interbreeding (Fig. 2.7D). Answers focusing only on the levels of gene flow were the second most common, accounting for 30% of answers. Answers focusing exclusively on the distinctness of populations were uncommon, accounting for roughly 6% of the total. Seventeen percent of the answers mentioned both an organismal and genetic perspective, though varied depending on whether RI was a reduction in interbreeding (which causes a reduction in gene flow), or a reduction in gene flow (caused by a reduction in mating). These results suggest that speciation researchers are divided on exactly what RI is, highlighting the need for clarification.

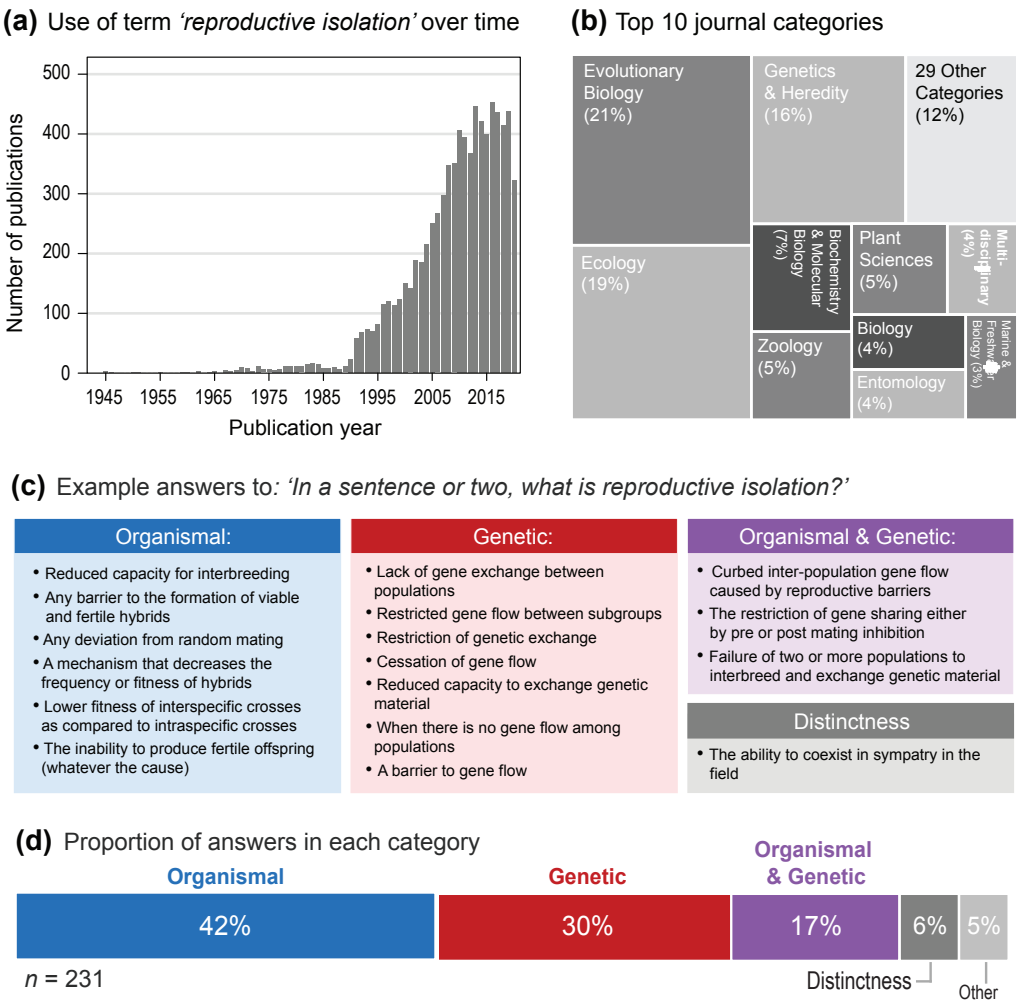


Figure 2.7: **Use of RI in the literature and insights into its meaning from an online survey.** A) Number of papers using the term 'reproductive isolation' in their title abstract or keywords and B) the top 10 journal categories in with the term RI is used, both according to ISI web of Science as of September 23, 2021. C) Example answers to the question: 'in a sentence or two, what is reproductive isolation?', from the speciation survey, classified as described in the text. D) The percentage of answers classified into each category. See Table S2 (2.8) for the full set of answers and methodological details.

BOX 2: Ecogeographic isolation

Sobel and Chen (2014) propose that when the ranges of two taxa are determined by genetic differences between them, the fraction which does not overlap should be included as a component of isolation. Here, we argue that the effect of range overlap on gene flow, and hence, on reproductive isolation, depends on both the geographic context, and on the nature of the genetic differentiation. Simple measures of range overlap are only informative under quite restrictive assumptions.

If the distinct populations have sufficiently different niches, and interbreed sufficiently

rarely, then they may coexist in sympatry over some fraction of their ranges. If their respective habitats each form a mosaic, with some degree of overlap, and if there is substantial gene flow amongst patches of each of the two populations ($Nm \gg 1$), then we can treat them as two well-mixed populations (Fig. 2.8A). The effective migration rate will then be the fraction of range overlap, multiplied by the effective migration rate when they are in sympatry. In this situation, then, the fraction of non-overlap is a sensible component of reproductive isolation, as proposed by Sobel and Chen (2014).

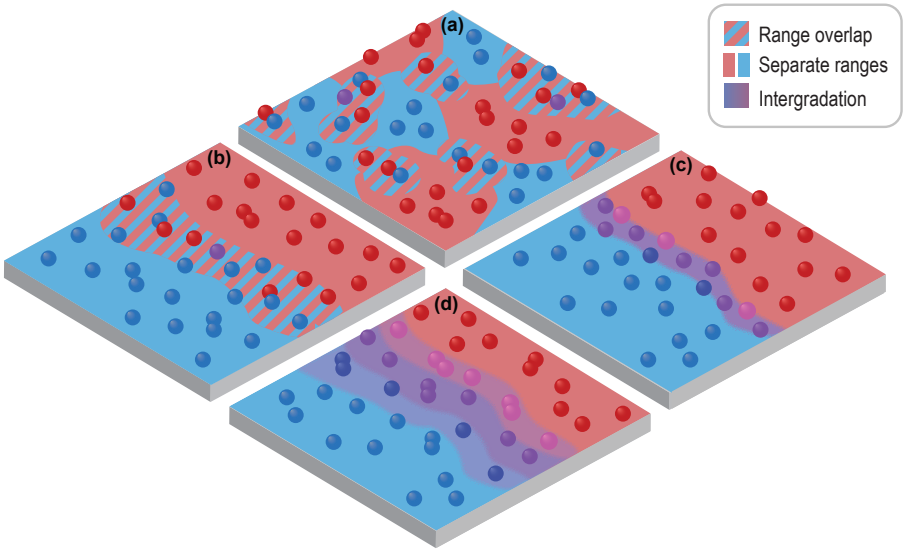


Figure 2.8: **Different patterns of range overlap, in a two-dimensional habitat.** A) Two populations (red, blue) are distributed in a mosaic, and remain distinct where they overlap. B) Two contiguous ranges overlap in an intermediate region, again remaining distinct in an intermediate region of sympatry. C) Two populations are separated by a narrow hybrid zone, in which clines for multiple genetic differences coincide. D) Clines are scattered, so that there is a broader region of intergradation.

Next, suppose that the two populations have contiguous ranges, which overlap in an intermediate region along a one-dimensional continuum (Fig. 2.8B). This region can be treated as a localised barrier to gene flow, which causes a step in the frequency of divergent neutral alleles. If the density of one population is $n_1(x)$, which declines from n_1^* on the left to zero on the right, and the other density is $n_2(x)$, which increases from zero on the left to n_2^* on the right, then one can show that the barrier strength is $B = (\sigma^2 n_1^* n_2^*)/(m \int 2 n_1 n_2 dx)$. Here, m is the proportion of each population that is exchanged between the populations per generation (Fig. 2.8; derivation in Appendix A.2: Ecogeographic isolation). If there are additional incompatibilities, then m should be replaced by the appropriate effective rate (Eq. 2.2). Thus, across a one-dimensional

habitat, the barrier strength B is the appropriate measure of reproductive isolation; if the integral in the denominator is taken as a measure of the distance over which the taxa overlap, then the barrier is inversely proportional to the fraction of range overlap.

It may be, however, that the taxa are separated by a narrow hybrid zone (Fig. 2.8C), or by a broad region of intergradation (Fig. 2.8D), with multiple overlapping clines. In such a case, the barrier may be calculated as described in the previous sections, and may be much weaker than suggested by the net divergence. With more than a few loci, the ranges of the parental genotypes will not overlap at all, even if the clines coincide (Fig. 2.8C), and so the ranges of the parental genotypes do not overlap at all. Yet, as we have seen, a narrow hybrid zone may pose a negligible barrier to gene flow, across most of the genome; if the clines are scattered, the barrier will be still weaker (Fig. 2.8D). In such cases, the fraction of overlap (however defined) is not an appropriate measure of isolation.

Glossary of key terms

Barrier to gene flow A physical or genetic obstacle to gene flow.

Barrier loci Loci that cause a barrier to gene flow.

Effective migration rate (m_e) The migration rate in the absence of a barrier that would have an effect equivalent to the actual migration rate in the presence a genetic barrier.

Geographic isolation A reduction in migration between populations due to geographic barriers (e.g., mountains or rivers) or geographic distance.

Genome-wide RI The RI experienced by a neutral allele that is unlinked to any selected loci.

Local RI The RI experienced by a neutral allele due to selection acting on both linked and unlinked selected loci. Local RI will inevitably vary from locus to locus, depending on the proximity to barrier loci.

Migration rate (m) The fraction of individuals that derive from elsewhere in the previous generation.

Population A group of individuals that are of some interest.

Reproductive isolation (RI) RI is a quantitative measure of the effect of genetic differences on gene flow. RI compares the flow of neutral alleles from one population to another population, given a set of genetic differences that reduce gene flow, with the flow without any such differences. The exact definition depends on the spatial context, among other things.

Response to Commentaries

The response to 5 commentaries on this article can be found at https://doi.org/10.1111/jeb.14082

Supplementary Table

Online supplementary tables S1 and S2 can be found at https://academic.oup.com/jeb/article/35/9/1143/7317908#supplementary-data

References

Aeschbacher, S., Selby, J., Willis, J., and Coop, G. (2017). Population-genomic inference of the strength and timing of selection against gene flow. *PNAS*, 114:7061–7066.

Barton, N. (1980). The hybrid sink effect. *Heredity*, 44:277–278.

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

Barton, N. and Bengtsson, B. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57:357–376.

Barton, N. and Gale, K. (1993). Genetic analysis of hybrid zones. In *Hybrid zones and the evolutionary process, ed. Harrison R*, pages 13–45. Oxford University Press.

Barton, N. and Hewitt, G. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, 16:113–148.

Barton, N. and Shpak, M. (2000). The effect of epistasis on the structure of hybrid zones. *Genetical Research*, 75:179–198.

Bazykin, A. (1969). Hypothetical mechanism of speciaton. *Evolution*, 23:685–687.

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162:2025–2035.

Bengtsson, B. (1985). The flow of genes through a genetic barrier. In *Evolution: Essays in honour of John Maynard Smith*, pages 31–42. Cambridge University Press, Cambridge.

Bierne, N., Borsa, P., Daguin, C., Jollivet, D., Viard, F., Bonhomme, F., et al. (2003). Introgression patterns in the mosaic hybrid zone between *Mytilus edulis* and *M. galloprovincialis*. *Molecular Ecology*, 12:447–461.

Burri, R., Nater, A., Kawakami, T., Mugal, C., Olason, P., Smeds, L., et al. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research*, 25:1656–1665.

Chase, M., Ellegren, H., and Mugal, C. (2021). Positive selection plays a major role in shaping signatures of differentiation across the genomic landscape of two independent *Ficedula* flycatcher species pairs. *Evolution*, 75:2179–2196.

Coyne, J. and Orr, H. (1989). Patterns of speciation in *Drosophila*. *Evolution*, 43:362–381.

Coyne, J. and Orr, H. (2004). *Speciation*. Sinauer Associates, Sunderland, MA.

Cruickshank, T. and Hahn, M. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23:3133–3157.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. Columbia University Press, New York, 1st edition.

Emerson, A. (1935). Termitophile distribution and quantitative characters as indicators of physiological speciation in british guiana termites (isoptera). *Annals of the Entomological Society of America*, 28:369–395.

Fisher, R. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

Fisher, R. (1937). The wave of advance of advantageous genes. *Annals of Eugenics*, 7:355–369.

Flaxman, S., Feder, J., and Nosil, P. (2013). Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. *Evolution*, 67:2577–2591.

Gavrilets, S. (2004). *Fitness Landscapes and the Origin of Species*. Princeton University Press.

Grafen, A. (2006). A theory of Fisher's reproductive value. *Journal of Mathematical Biology*, 53:15–60.

Gutenkunst, R., Hernandez, R., Williamson, S., and Bustamante, C. (2010). Diffusion approximations for demographic inference: DaDi. *Nature Precedings*.

Haldane, J. (1948). The theory of a cline. *Journal of Genetics*, 48:277–284.

Hatfield, T., Barton, N., and Searle, J. (1992). A model of a hybrid zone between two chromosomal races of the common shrew (*Sorex araneus*). *Evolution*, 46:1129–1145.

Hewitt, G. (1988). Hybrid zones-natural laboratories for evolutionary studies. *Trends in Ecology & Evolution*, 3:158–167.

Huxley, T. H. (1860). The origin of species. *Westminster Review*, 17(541570):1862.

Irwin, D. (2020). Assortative mating in hybrid zones is remarkably ineffective in promoting speciation. *The American Naturalist*, 195:E150–E167.

Jiggins, C. and Mallet, J. (2000). Bimodal hybrid zones and speciation. *Trends in Ecology & Evolution*, 15:250–255.

Kruuk, L., Baird, S., Gale, K., and Barton, N. (1999). A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, 153:1959–1971.

Laetsch, D. R., Bisschop, G., Martin, S. H., Aeschbacher, S., Setter, D., et al. (2023). Demographically explicit scans for barriers to gene flow using gIMble. *PLOS Genetics*, 19(10):e1010999.

Lowry, D., Modliszewski, J., Wright, K., Wu, C., and Willis, J. (2008). The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363:3009–3021.

Mallet, J. (2004a). Perspectives Poulton, Wallace and Jordan: How discoveries in *Papilio* butterflies led to a new species concept 100 years ago. *Systematics and Biodiversity*, 1:441–452.

Mallet, J. (2004b). Species problem solved 100 years ago. *Nature*, 430:503–503.

Martin, S., Davey, J., and Jiggins, C. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 32:244–257.

Martin, S., Davey, J., Salazar, C., and Jiggins, C. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLOS Biology*, 17:e2006288.

Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia University Press, New York.

Mayr, E. (1959). Isolation as an evolutionary factor. *Proceedings of the American Philosophical Society*, 103:221–230.

Momigliano, P., Florin, A.-B., and Merilä, J. (2021). Biases in demographic modeling affect our understanding of recent divergence. *Molecular Biology and Evolution*, 38:2967–2985.

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

Orr, H. (1996). Dobzhansky, Bateson, and the genetics of speciation. *Genetics*, 144:1331–1335.

Perini, S., Rafajlović, M., Westram, A., Johannesson, K., and Butlin, R. (2020). Assortative mating, sexual selection, and their consequences for gene flow in *Littorina*. *Evolution*, 74:1482–1497.

Petr, M., Pääbo, S., Kelso, J., and Vernot, B. (2019). Limits of long-term selection against Neandertal introgression. *PNAS*, 116:1639–1644.

Piálek, J. and Barton, N. (1997). The spread of an advantageous allele across a barrier: The effects of random drift and selection against heterozygotes. *Genetics*, 145:493–504.

Polechová, J. and Barton, N. (2011). Genetic drift widens the expected cline but narrows the expected cline width. *Genetics*, 189:227–235.

Porter, A., Wenger, R., Geiger, H., Scholl, A., and Shapiro, A. (1997). The *Pontia Daplidice-Ed Usa* hybrid zone in northwestern italy. *Evolution*, 51:1561–1573.

Poulton, E. (1904). What is a species?(presidential address to the entomological society of london). In *Proceedings of the Entomological Society London*, pages 1889–1907.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Rabosky, D. (2016). Reproductive isolation and the causes of speciation rate variation in nature. *Biological Journal of the Linnean Society*, 118:13–25.

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Westram, A. M., et al. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, 30(8):1450–1477.

Rougemont, Q., Gagnaire, P.-A., Perrier, C., Genthon, C., Besnard, A.-L., Launey, S., et al. (2017). Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. *Molecular Ecology*, 26:142–162.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science*, 323:737–741.

Sloan, D., Havird, J., and Sharbrough, J. (2017). The on-again, off-again relationship between mitochondrial genomes and species boundaries. *Molecular Ecology*, 26:2212–2236.

Sobel, J. and Chen, G. (2014). Unification of methods for estimating the strength of reproductive isolation. *Evolution*, 68:1511–1522.

Sobel, J. and Streisfeld, M. (2015). Strong premating reproductive isolation drives incipient speciation in *Mimulus aurantiacus*: Reproductive isolation in *M. AURANTIACUS*. *Evolution*, 69:447–461.

Stankowski, S. and Ravinet, M. (2021). Defining the speciation continuum. *Evolution*, 75:1256–1273.

Stankowski, S., Shipilina, D., and Westram, A. M. (2021). Hybrid zones. In *eLS*, pages 1–12. Wiley.

Szymura, J. and Barton, N. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina Bombina* and *B. Variegata*, near Cracow in southern Poland. *Evolution*, 40:1141–1159.

Szymura, J. and Barton, N. (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*, 45:237–261.

Toews, D. P. and Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, 21:3907–3930.

Turelli, M. and Barton, N. H. (2017). Deploying dengue-suppressing wolbachia: Robust models predict slow but effective spatial spread in *Aedes aegypti*. *Theoretical Population Biology*, 115:45–60.

Virdee, S. and Hewitt, G. (1994). Clines for hybrid dysfunction in a grasshopper hybrid zone. *Evolution*, 48:392–407.

Wallace, A. (1865). On the phenomena of variation and geographical distribution as illustrated by the Papilionidæ of the Malayan region. *Transactions of the Linnean Society of London*, 25:1–71.

Wright, S. (1943). Isolation by distance. *Genetics*, 28:114–138.

Chapter 3

Effect of assortative mating and sexual selection on polygenic barriers to gene flow[1]

Abstract

Assortative mating and sexual selection are widespread in nature and can play an important role in speciation, through the buildup and maintenance of reproductive isolation (RI). However, their contribution to genome-wide suppression of gene flow during RI is rarely quantified. Here, we consider a polygenic 'magic' trait that is divergently selected across two populations connected by migration, while also serving as the basis of assortative mating, thus generating sexual selection on one or both sexes. We obtain theoretical predictions for divergence at individual trait loci by assuming that the effect of all other loci on any locus can be encapsulated via an effective migration rate, which bears a simple relationship to measurable fitness components of migrants and various early generation hybrids. Our analysis clarifies how 'tipping points' (characterised by an abrupt collapse of adaptive divergence) arise, and when assortative mating can shift the critical level of migration beyond which divergence collapses. We quantify the relative contributions of viability and sexual selection to genome-wide barriers to gene flow and discuss how these depend on existing divergence levels. Our results suggest that effective migration rates provide a useful way of understanding genomic divergence, even in scenarios involving multiple, interacting mechanisms of RI.

Keywords: assortative mating, sexual selection, effective migration rates, polygenic selection, reproductive isolation

[1]This work can be found online at https://biorxiv.org/cgi/content/short/2024.07.30.605898v1

3.1 Introduction

Reproductive isolation (RI) typically results from the interaction of multiple processes that cause loss of hybrid fitness, thus reducing gene flow between populations. Ecological specialisation arising from adaptation to different environmental niches, immigrant inviability, or intrinsic genetic incompatibilities may all generate postzygotic barriers to gene flow and maintain genetic differences between populations (Rundle and Nosil (2005); Coughlan and Matute (2020); Rice and Hostert (1993)). Additionally, processes such as assortative mating and sexual selection can act as both prezygotic and postzygotic barriers— by reducing heterospecific matings (that produce hybrids) as well as the mating success of hybrids. Indeed, it has been argued that the buildup of RI typically involves coupling between both postzygotic and prezygotic barriers (Butlin and Smadja (2018)).

Assortative mating (AM)– which involves positive phenotypic (and genetic) correlation between mating pairs (Lewontin et al. (1968)), is especially common in animals Jiang et al. (2013). Numerous studies have documented its role during RI, e.g., in cichlids (Elmer et al. (2009); Stelkens and Seehausen (2009)), swordtail fish (Schumer et al. (2017)), sticklebacks (Vines and Schluter (2006)), butterflies (Jiggins et al. (2001)), and marine snails (Johannesson et al. (2008)). AM decreases heterozygosity and genic variance within a population, but simultaneously increases total genetic variance (by generating higher levels of linkage disequilibria (LD)) (Lynch et al. (1998)), thus increasing the potential for divergence and speciation.

AM may occur either due to temporal or spatial separation of population groups, via preference-trait mechanisms, or via phenotype matching between potential mates (Kopp et al. (2018); Jiang et al. (2013); Kirkpatrick and Ravigné (2002)). Each mechanism differs in the extent of sexual selection it generates— ranging from strong sexual selection under phenotype matching to almost none when assortment is due to temporal or spatial separation. Sexual selection typically reduces mating success of rare phenotypes, and thus may counter the effects of AM (Servedio and Boughman (2017); Safran et al. (2013)). For example, when AM involves sexual selection on both sexes (e.g. if both males and females suffer reduced mating opportunities by being choosy), then its net effect is to generate stabilizing selection and inhibit divergence (Kirkpatrick and Nuismer (2004)).

Numerous studies have examined how different isolating mechanisms, namely assortative mating (Doebeli (1996); Kondrashov and Shpak (1998); Kopp et al. (2018)), ecological selection (Smith (1966); Maan and Seehausen (2011)), and sexual selection (Servedio (2016); Servedio and Kopp (2012); Schumer et al. (2017)) influence divergence and RI in the presence of ongoing gene flow. These suggest that RI is facilitated by coupling between traits that act as reproductive barriers (Coyne and Allen Orr (2004); Smadja and Butlin (2011); Kirkpatrick and Ravigné (2002)). Such coupling can arise if loci underlying

different traits have pleiotropic effects or are tightly linked and not easily broken apart by recombination (Felsenstein (1981); Butlin and Smadja (2018); Smadja and Butlin (2011); Servedio (2009)). In this regard, speciation is thought to be most effective with magic traits, i.e., if the same trait is under divergent selection and also mediates AM. Magic traits are now known to be more common than previously thought (Servedio et al. (2011a)): well-known examples include wing color pattern in Heliconius (Merrill et al. (2012)) and body size in *Littorina saxatalis* (Perini et al. (2020)).

Magic traits may differ in their mechanism of AM and the extent to which they are under sexual selection, and have been studied extensively to understand the evolution of mate choice during speciation (Servedio et al. (2011b); Servedio and Kopp (2012); Servedio and Boughman (2017)). While previous work largely focuses on one or a few loci that mediate mate choice, here we consider a polygenic trait influenced by a large number of small-effect alleles. Not only is this more realistic in view of the polygenic architecture of most traits but also much less studied (though see Kirkpatrick (2000); Sachdeva and Barton (2017); Muralidhar et al. (2022)). In such cases, the evolution of adaptive divergence and RI depends on the ease with which *multiple* locally advantageous alleles can establish and/or be maintained across the genome despite maladaptive gene flow between populations. This in turn depends on whether selected loci evolve more or less independently or if LD between sets of maladapted alleles is strong enough for these to be eliminated together before they can recombine onto fitter backgrounds. This was first studied in the context of hybrid zones, where it was found that the strength of LD between multiple selected loci depends on the ratio of total selection and total recombination (Barton (1983)).

As we demonstrate here, a powerful way of analysing the effects of LD on divergence is by tracking how introgressing alleles at any one trait locus are transferred between different genetic backgrounds, typically via multiple recombination events over multiple generations. The different genetic backgrounds on which the focal allele finds itself may differ in their hybrid index (i.e., the proportion of genetic material they derive from the parental populations) and thus also have different fitness values, which influences the introgression probabilities of any allele they carry. This is captured by the notion of the *effective* migration rate (m_e), which is the rate at which neutral introgressing alleles are transferred between hybridising populations (Bengtsson (1985)). In essence, strong LD between large numbers of loci reduces m_e across the genome (Barton and Bengtsson (1986)). This increases divergence between hybridising populations, which reduces hybrid fitness and further pushes down effective migration rates. Indeed, it has been argued that this genomewide reduction in m_e between hybridising populations provides a natural way of quantifying RI (Westram et al. (2022)). Subsequent work has generalised theory based on m_e to weakly selected loci with partial and/or transient divergence between populations (Sachdeva (2022)), heterogeneous effect sizes and dominance (Zwaenepoel

et al. (2024)) and sex-linkage (Fraïsse and Sachdeva (2021)). However, all of this work is restricted to randomly mating populations.

In this paper, we extend theory based on effective migration rates to a scenario with assortative mating due to female preference for phenotypically similar males, which leads to sexual selection on one or both sexes. The phenotype underlying assortment is polygenic and under divergent selection across a mainland and island: thus, we have a simple 'magic trait' scenario of speciation. We derive expressions for m_e at an unlinked neutral locus and then use this to predict allele frequency divergence at individual trait loci under migration-selection-assortment balance by assuming that the effect of all other loci on the focal locus is encapsulated by this effective migration rate. This allows us to predict mean trait divergence between mainland and island, and explore how assortment, viability and sexual selection jointly influence the critical threshold at which migration swamps adaptive divergence. We then use this to disentangle the relative contributions of viability and sexual selection to the genomewide barrier to gene flow in parameter regimes where partial RI persists despite limited hybridisation.

3.2 Model and Methods

3.2.1 Model

Consider an island subject to one-way migration from a large mainland, such that a fraction m of individuals on the island are replaced by migrants in every generation. Individuals are haploid and hermaphroditic (i.e. produce both male and female gametes), and can thus play the male or female role during reproduction. Each individual expresses an additive polygenic trait $z = \sum_{i=1}^{L} \alpha_i X_i$ influenced by a large number L of unlinked, biallelic loci, where $X_i = 0, 1$ denote alternative alleles and α_i is the effect size at locus i. The trait is under directional selection on the island– the relative fitness due to viability selection of an individual with trait value z is $W(z) = e^{-\beta z}$, where β denotes the strength of selection. We will not explicitly model the mainland population but assume that trait values follow a normal distribution, or alternatively, that migrant genotypes are maximally deleterious on the island (see section 3.2.2).

Individuals on the island mate assortatively, with trait z also serving as the basis for AM via female choice. We assume a Gaussian choice function, such that the probability of mating between a male and female with trait values z_M and z_F is proportional to $e^{-\gamma(z_M - z_F)^2}$, where γ is the strength of assortment. In other words, females preferentially mate with males with trait values within a few $1/\sqrt{\gamma}$ of their own. Thus, the assortment trait is a 'magic trait', which simultaneously influences mate choice and is under both natural and sexual selection, thereby generating both postzygotic and prezygotic barriers

to gene flow.

We consider two different models of AM– allowing for sexual selection on both sexes (Model I) or only on males (Model II)– sometimes referred to as the 'plant model' and 'animal model' of assortment (Kirkpatrick and Nuismer (2004)). More concretely, the probability $M(z_M, z_F)$ of mating between a male and female with trait values z_M and z_F under the two models is:

$$M(z_M, z_F) = \frac{e^{-\gamma(z_M - z_F)^2}}{\int dz \int dy\, P(z)P(y)e^{-\gamma(z-y)^2}} \qquad \text{Model I}$$

$$M(z_M, z_F) = \frac{e^{-\gamma(z_M - z_F)^2}}{\int dz\, P(z)\, e^{-\gamma(z-z_F)^2}} \qquad \text{Model II}$$

Here, $P(z)$ is the trait value distribution just before mating. Note that under Model II, mating probabilities are normalized such that any female, regardless of trait value, mates with probability 1, while under Model I, females with rare phenotypes have lower mating probability. Thus, sexual selection acts only on males in the former case but on both sexes in the latter.

3.2.2 Effective migration rate at a neutral locus

Our analysis is based on effective migration rates (m_e)– defined as the rate at which neutral alleles entering a population via migration are incorporated into the resident gene pool (Bengtsson (1985); Barton and Bengtsson (1986)), where 'residents' are defined more precisely later. In our model, migrants and their descendants experience both viability and sexual selection and are less fit than residents. Consequently, in order to establish, neutral alleles must recombine away from migrant genetic backgrounds before these are eliminated by selection. Thus, m_e will vary along the genome, being lower at sites that are tightly linked to a selected locus or in genomic regions with a high density of selected loci. However, we will focus on m_e at a neutral locus that is *unlinked* to any selected locus, since we are interested in quantifying the average or genomewide reduction in gene flow.

Under rare migration ($m \ll 1$), the ratio m_e/m for an unlinked neutral allele, also called the gene flow factor g, is equal to the reproductive value (RV) of migrants relative to residents (Kobayashi et al. (2008)). Here, RV refers to the long-term genetic contribution of the migrant in the recipient population (Fisher (1999); Barton and Etheridge (2011)), and is given by:

$$g = \frac{m_e}{m} = \prod_{i=0}^{\infty} \overline{W}_i \tag{3.1}$$

where $\overline{W}_0$ denotes the average fitness of migrants relative to residents, and $\overline{W}_i$ the average relative fitness of i^{th} generation descendants of migrants (see, e.g., Westram et al. (2022)). For $m \ll 1$, individuals with recent immigrant ancestry are rare, so that migrants and their descendants mate primarily with residents. As a result, first-generation descendants of the migrants are F_1 hybrids, second-generation descendants are typically first-generation backcrosses (with residents), i^{th} generation descendants are $(i-1)^{th}$ generation backcrosses (denoted by BC_{i-1}), and so on. Thus, computing m_e boils down to calculating the relative fitness of successive back-crosses.

In the following, we classify individuals on the island by the number of generations leading back to their most recent immigrant ancestor, i.e., into F_1, BC_1, BC_2, up to BC_n (where n is arbitrary). All other individuals, i.e., those with no migrant ancestor in the previous $n+1$ generations, are designated as 'residents'. We further assume trait values to be normally distributed *within* any group of individuals with the same level of recent migrant ancestry. In other words, the trait distributions $P_r(z)$, $P_1(z)$, $\ldots$ $P_i(z)$, $\ldots$ amongst residents, F_1 hybrids, i^{th} generation descendants (who are BC_{i-1}) and so on, are assumed to be normal with means $\overline{z}_r, \overline{z}_1, \ldots \overline{z}_i, \ldots$ and variances $V_r, V_1, \ldots V_i, \ldots$ respectively (see figure 3.1A). For generality, we also take the migrant trait value distribution to be normal with mean $\overline{z}_0$ and variance V_0. However, in the Results, we will only consider a scenario where the mainland is fixed for alleles locally deleterious on the island (so that $\overline{z}_0 = \sum_{i=1}^{L} \alpha_i$ and $V_0 = 0$).

In general, assuming normally distributed trait values is justified for traits influenced by a large number of loci of small effect. The inheritance of such traits is described by the infinitesimal model (Fisher (1918)), which states that the offspring of any two individuals have trait values that are normally distributed about the mean of the parents, with a 'within-family' variance V_*, which depends only on the genetic relatedness between parents, regardless of selection, non-random mating etc. (Barton et al. (2017)). We further assume that V_* depends only on the extent of migrant ancestry of the parental individuals, and does not vary significantly across (for instance) different resident $\times$ F_1 parental pairs, all of which thus have approximately the same V_*. We denote the within-family variance by $V_{r,r}$ when both parents are residents, $V_{r,0}$ when one parent is a resident and the other migrant, and $V_{r,i}$ when one parent is a resident and the other an i^{th} generation descendant of a migrant.

Then, we can express the distribution $P_i(z)$ in terms of the parental trait value distributions $P_r(z_1)$ and $P_{i-1}(z_2)$ (eq. 3.2a). Further, we can write down the mean fitness $\overline{W_i}$ of i^{th} generation descendants (relative to residents) as an integral over the parental distributions

(eq. 3.2b).

$$P_i(z) \propto \int dz_1 \int dz_2 \, P_r(z_1) P_{i-1}(z_2) \, e^{-\beta(z_1+z_2)} \left[\frac{M(z_1,z_2)+M(z_2,z_1)}{2}\right] \frac{e^{-\frac{(z-(z_1+z_2)/2)^2}{2V_{r,i-1}}}}{\sqrt{2\pi V_{r,i-1}}} \tag{3.2a}$$

$$\overline{W}_i = \int dz_1 \int dz_2 \, \frac{e^{-\beta z_1}}{\int dz \, e^{-\beta z} P_r(z)} P_r(z_1) \, \frac{e^{-\beta z_2}}{\int dz \, e^{-\beta z} P_r(z)} P_i(z_2) \left[\frac{M(z_1,z_2)+M(z_2,z_1)}{2}\right] \tag{3.2b}$$

In eq. 3.2a, z_1 and z_2 denote respectively the trait values of the resident parent and the parent who is an $(i-1)^{th}$ generation descendant of a migrant, with corresponding trait value distributions $P_r(z_1)$ and $P_{i-1}(z_2)$. The term $e^{-\beta(z_1+z_2)}$ captures the effect of viability selection on the parents, while $M(z_1, z_2)$ (which is the probability of mating between a male and female with trait values z_1 and z_2 respectively) captures the effect of AM and sexual selection. Note that the non-resident parent may have a different probability of being chosen as a mate depending on whether it takes the female or male role during reproduction: this is captured by the two terms $M(z_1, z_2)$ and $M(z_2, z_1)$ (that are unequal under Model II). Finally, the last term gives the distribution of trait values of the offspring under the infinitesimal model; it captures the effect of recombination between parental genotypes and random segregation. Appendix B.1 provides a detailed derivation of eq. (3.2).

Integrating over eq. 3.2a gives an expression for the mean and variance of trait values among the i^{th} generation descendants in terms of the corresponding means and variances among $(i-1)^{th}$ generation descendants and among residents. This specifies a set of recursions for $\overline{z}_i$, V_i and $V_{r,i}$ which can be solved numerically (see Appendix B.1). In turn, these predict the relative fitnesses $\overline{W}_i$ (using eq. 3.2b), from which one can calculate the gene flow factor g (using eq. (3.1)) and m_e.

However, in the main paper, we will derive a more approximate and intuitive expression for g by simply tracking the difference between trait means of residents and other groups (i.e., F_1s, BC_1s and so on), while neglecting phenotypic variance *within* any group. More concretely, we assume that the mean trait value of hybrid offspring is the average of the trait means of the parental subgroups. Thus, for example, F_1 hybrids are assumed to be midway between residents and migrants, giving $\overline{z}_1 \approx (\overline{z}_r + \overline{z}_m)/2$, which gives: $\overline{z}_1 - \overline{z}_r \approx (\overline{z}_m - \overline{z}_r)/2 = \Delta/2$. Here $\Delta = \overline{z}_m - \overline{z}_r$ denotes the difference in the trait means of migrants and residents, which we refer to as mean trait divergence. It follows similarly that $\overline{z}_i - \overline{z}_r \approx (\overline{z}_{i-1} - \overline{z}_r)/2 = \Delta/2^i$. Thus, in effect, every successive generation of backcrossing halves the average phenotypic distance of the (next-generation) backcross from the resident phenotype (Sachdeva (2022); Zwaenepoel et al. (2024)).

As we argue below and in Appendix B.1, this simple argument holds as long as $\beta V_i \ll \Delta$

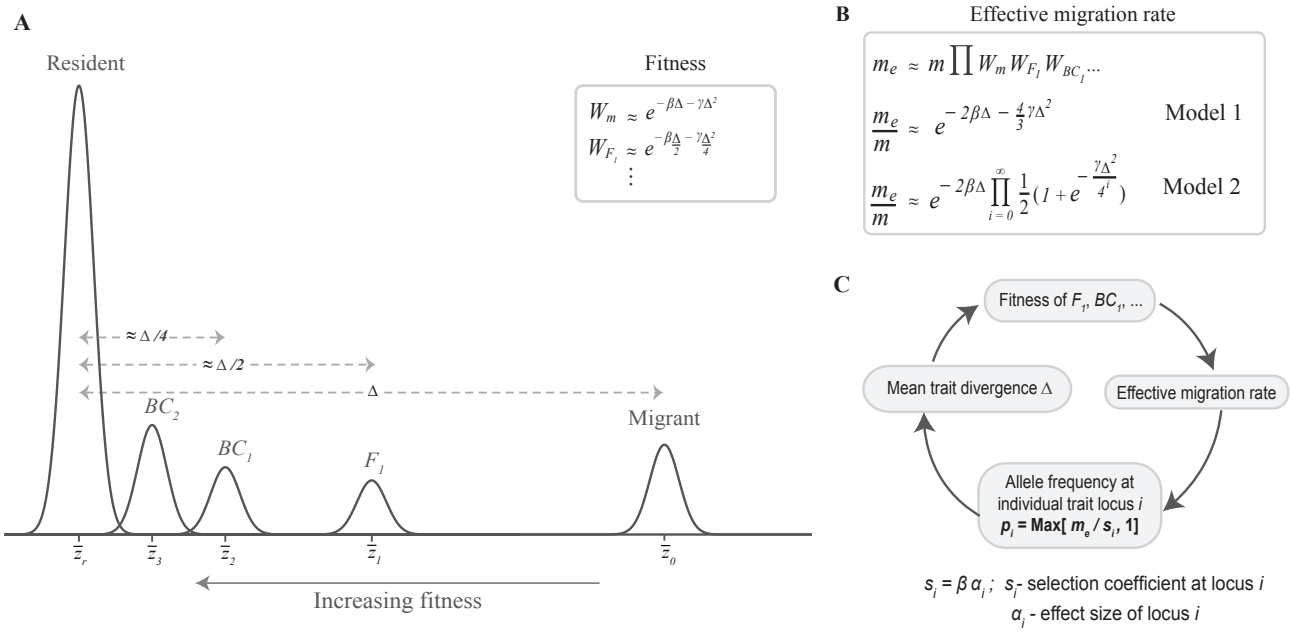


Figure 3.1: Schematic summarising our theoretical approach for predicting mean trait divergence Δ (or allele frequency divergence per locus Y) using effective migration rates (m_e). **A**: Distribution of trait values on the island is the sum of trait value distributions associated with residents, F_1s, BC_1s, and so on, where each such distribution is assumed to be approximately normal. Denoting the difference between trait means of migrants and residents by $\Delta = \overline{z}_0 - \overline{z}_r$, the difference between trait means of F_1s and residents is $\approx \Delta/2$, and more generally, between that of i^{th} generation descendants (of migrants) and residents is $\Delta/2^i$. This allows us to calculate *average relative fitness* of F_1s, BC_1 etc. (inset). **B**: The *gene flow factor* g (which describes the reduction in neutral gene flow at a site unlinked to any selected site) can be calculated from the average relative fitness of migrants and of individuals with recent migrant ancestry (F_1s, BC_1s), as in eq. (3.1). This gives approximate expressions for g under Model I (sexual selection on both sexes) and Model II (sexual selection only on males) **C**: The *mean trait divergence* Δ can be calculated by first using m_e to predict the allele frequency p_i at trait locus $i = 1, \ldots, L$, then summing over loci to obtain the mean trait divergence Δ, then using this to predict relative fitnesses of F_1s, BC_1s and so on (as in A.), and finally using these to predict the effective migration rate m_e (as in B.). This kind of circular dependence, where m_e depends on Δ which depends on m_e (via allele frequencies), allows for a self-consistent solution for the mean trait divergence Δ.

and $\gamma V_i \ll 1$ (for all i). Moreover, under these conditions, the relative fitness of any group of descendants is (from eq. 3.2b): $\overline{W_i} \approx e^{-\beta(\overline{z}_i-\overline{z}_r)-\gamma(\overline{z}_i-\overline{z}_r)^2}$ for Model I (sexual selection on both sexes) and $\approx e^{-\beta(\overline{z}_i-\overline{z}_r)}\frac{1}{2}(1+e^{-\gamma(\overline{z}_i-\overline{z}_r)^2})$ for Model II (sexual selection only on males). Now, using $\overline{z}_i - \overline{z}_r = \Delta/2^i$, where Δ is the mean trait divergence between mainland and island, we obtain the following approximate expressions for the gene flow factor $g = \prod_{i=0}^{\infty}\overline{W_i}$:

$$g(\Delta) \approx e^{-2\beta\Delta-\frac{4}{3}\gamma\Delta^2} \qquad \text{Model I} \qquad (3.3a)$$

$$g(\Delta) \approx e^{-2\beta\Delta} \prod_{i=0}^{\infty} \left(\frac{1 + e^{-\gamma\Delta^2/4^i}}{2} \right) \quad \text{Model II} \tag{3.3b}$$

A more detailed derivation of eq. 3.3 is provided in Appendix B.1. A striking feature of eq. 3.3 is that under both models of sexual selection, the gene flow factor can be expressed as a product of two terms– the first describing the reduction in reproductive value of migrants due to viability selection, and the second due to sexual selection. The first term above, $e^{-2\beta\Delta}$, is the square of the ecological fitness of migrants relative to residents. The second term is related to the relative mating success of migrants and differs between the two models. For instance, under Model I, the sexual selection term is simply the relative mating success of migrants raised to the power 4/3. These powers (of 2 or 4/3) emerge from the specific fitness and mating functions assumed in our model. However, more generally, the fact that the viability and sexual selection components of g bear a simple relationship to the corresponding fitness components (of say migrants or F_1s) is due to there being a simple relationship between the immediate fitness and long-term reproductive value of individuals when fitness is polygenic.

What assumptions underlie our simple approximation for trait means, which is the basis of eq. (3.3)? First, within any group of descendants, say F_1s, individuals closer to the resident phenotype will have higher viability, causing the trait mean of BC_1s (the hybrid offspring of F_1s and residents) to be slightly shifted towards residents rather than exactly midway between the two parental groups. The magnitude of this shift is proportional to βV_1; thus, in ignoring the effects of viability selection *within* subgroups, we assume that $\beta V_i \ll \Delta$, where Δ is the mean trait divergence between mainland and island. As we argue in Appendix B.1, under migration-selection balance, the various βV_i must be comparable to the migration rate m, and can thus be neglected if $m \ll 1$.

Second, we ignore the effects of sexual selection within subgroups, which may lead to small differences in mating success among (say) F_1s, again slightly shifting the BC_1 trait mean towards one of the parental subgroups. Moreover, AM can also increase the phenotypic variance within each group (Fisher (1918); Wright (1921); Crow and Felsenstein (1968)), at least when sexual selection is relatively weak, as under Model II (Kirkpatrick and Nuismer (2004)), thus also influencing the magnitude of these shifts. As we argue in Appendix B.1, both these effects can be ignored if $\gamma V_i \ll 1$, i.e., if the preference function is much wider than the typical phenotypic spread within any subgroup. Thus, in summary, our simple approximation for trait means– $\overline{z}_i - \overline{z}_r \approx \Delta/2^i$ holds as long as $\beta V_i \ll \Delta$ (which is true if $m \ll 1$) and if $\gamma V_i \ll 1$. In effect, these two conditions boil down to assuming that viability and sexual selection are strong enough to affect the survival and mating success of migrants and their F_1 and BC descendants relative to residents, but not strong enough to cause differential mating success *within* any such group of descendants.

In summary, eq. 3.3 relates the gene flow factor $g = m_e/m$ at an unlinked locus to

the mean trait divergence Δ between mainland and island at equilibrium, where Δ now depends on the allele frequency differences across all trait loci. In the following, we relate the allele frequency difference at any locus back to the effective migration rate m_e, thus allowing us to obtain the mean trait divergence in a self-consistent way.

3.2.3 Predicting allele frequency divergence at trait loci using m_e

We now consider allele frequency dynamics at individual trait loci, assuming the mainland to be fixed for alleles that are deleterious on the island. First, consider the equilibrium allele frequency at a locus with effect size α_i under *linkage equilibrium*, i.e., neglecting the effects of LD between the focal locus and other trait loci. In Appendix B.2, we show that sexual selection does not contribute to *direct* selection at a locus (at least under our assumed quadratic preference function), provided α_i is sufficiently small. More concretely, at migration-selection balance, the deleterious allele frequency p_i is simply m/s_i, where $s_i = \beta\alpha_i$, provided $\beta\alpha_i \ll 1$ and $\alpha_i \ll \gamma/\beta$ (Appendix B.2). The latter condition can be re-written as $\alpha_i/\Delta \ll (\gamma\Delta^2)/(\beta\Delta)$— this simply means that the relative contribution of any locus to trait divergence should be much smaller than the ratio of the sexual selection component of fitness to the viability selection component.

Following Sachdeva (2022), we now assume that the main effect of LD between immigrant alleles is to reduce m_e at any locus. This can be justified as long as trait loci are unlinked or weakly linked so that LD between these breaks down much faster than allele frequencies change. Further, if individual loci are weakly selected, then the effective migration rate at a trait locus is approximately equal to that at a neutral locus, so that in a large population, we have: $p_i \approx m_e/s_i \approx \frac{m}{s_i}\, g(\Delta)$. Here, the gene flow factor g is given by eq. 3.3, and captures the effect of both viability and sexual selection on migrants and their descendants.

Note that g depends on allele frequencies $\{p_i,\ i = 1, .. L\}$ at all L loci via the mean trait divergence $\Delta = \sum_{i=1}^{L} \alpha_i(1 - p_i)$. We can thus determine allele frequencies by numerically solving the L equations $p_i = (m/s_i)g(\Delta)$, all coupled via Δ (Zwaenepoel et al. (2024)). However, here we will focus on the simpler case of equal-effect loci, such that $\alpha_i = \alpha$ (and $s_i = s$) for all i. Then, in the absence of drift, equilibrium allele frequencies are equal across all loci (i.e., $p_i = p$), at least in parameter regimes where there is only one stable equilibrium Thus, the mean trait divergence can be written as $\Delta = \alpha L(1 - p)$. In the following, we will express our results in terms of $Y = \Delta/\Delta_0$, the trait divergence between the mainland and the island relative to the maximum possible divergence $\Delta_0 = \alpha L$. Note that for the case of equal-effect loci, we have: $Y = 1 - p$, so that Y is also the mean allele frequency divergence per locus. It follows from eq. 3.3 that Y satisfies:

$$Y \approx 1 - \frac{m}{s} e^{-2\beta\Delta_0 Y - \frac{4}{3}\gamma\Delta_0^2 Y^2} \qquad \text{Model I} \tag{3.4a}$$

$$Y \approx 1 - \frac{m}{s} e^{-2\beta\Delta_0 Y} \prod_{i=0}^{\infty} \left(\frac{1 + e^{-\gamma\Delta_0^2 \frac{Y^2}{4^i}}}{2} \right) \qquad \text{Model II} \tag{3.4b}$$

as long as a solution with $0 \leq Y \leq 1$ exists, and is zero otherwise. Eq. (3.4) can be solved numerically to obtain allele frequency divergence per locus Y or the trait divergence Δ. In the case of Model II, the product in eq. 3.4b converges rapidly and can be computed by taking ~ 20 terms.

3.2.4 Deterministic simulations

We test our approximations against deterministic simulations based on the hypergeometric model, which describes the inheritance of a quantitative trait influenced by equal-effect loci (Barton (1992); Doebeli (1996)). In each generation, migration is followed by viability selection, which is followed by assortative mating (involving sexual selection on one or both sexes), and production of offspring via free recombination between parental genotypes. The mainland is assumed to be fixed for alleles that are locally deleterious on the island. Additionally, unless stated otherwise, simulations are initiated with the island perfectly adapted, i.e., fixed for the locally adaptive allele at each trait locus.

Note that with equal-effect loci, the trait value z of any individual is just αj, where j is the number of locally deleterious alleles it carries. Appendix B.3 specifies the recursions that describe how the distribution $P(j)$ (or alternatively, $P(z)$) changes in any generation under the combined effects of migration, (viability and sexual) selection, AM and random segregation. Recursions for $P(j)$ are iterated until equilibrium, and allele frequency divergence per locus Y obtained using: $Y = \frac{1}{L}\sum_{j=0}^{L}(L-j)P(j)$. This can then be compared with theoretical predictions (eq. (3.4)).

3.3 Results

We now use predictions based on effective migration rates to investigate the combined effects of migration, viability and sexual selection, and assortative mating on the equilibrium allele frequency divergence between mainland and island as well as on the genome-wide reduction in neutral gene flow (as measured by the gene flow factor). We first quantify how mean trait divergence declines with increasing migration (which is parameterized by m/s, migration rate relative to selective effect per locus), for different strengths of viability and sexual selection against migrants. Next, we ask: how strong would viability and sexual selection need to be to *shift* the critical migration threshold beyond which

divergence is swamped by gene flow? In other words, under what conditions does LD between trait loci become sufficiently strong to increase the swamping threshold beyond that for a single divergently selected locus? Finally, we examine the relative contributions of viability and sexual selection to RI under different parameter regimes, by using the increase in F_{ST} at an arbitrary site (that is unlinked to any trait locus) as a proxy for the genome-wide barrier effect. Throughout, we quantify the strength of viability and sexual selection in terms of $\beta\Delta_0$ and $\gamma\Delta_0^2$, which denote respectively the reduction in log fitness of migrants (relative to residents) due to viability and sexual selection under conditions of maximum divergence between the mainland and island, i.e., when $\Delta = \Delta_0$.

If traits are highly polygenic (large L and small effect sizes α), then to a good approximation, divergence (as measured by $Y = \Delta/\Delta_0$) depends only on the composite parameters m/s, $\beta\Delta_0$ and $\gamma\Delta_0^2$, where $\Delta_0 = \alpha L$ (see eq. 3.4). However, for smaller L (say tens of loci), divergence levels may be sensitive to the exact genetic basis of the trait, i.e., depend on α and L separately (Appendix B.5). In the main text, we will focus on the large L limit (where eq. (3.4) applies).

3.3.1 Effect of viability and sexual selection on adaptive divergence.

Figure 3.2 shows Y, the trait divergence between mainland and island relative to the maximum possible divergence (or equivalently, allele frequency divergence per locus) as a function of m/s, under the two models of AM. For each model, the plots depict divergence levels for $\beta\Delta_0 = 0.2, 0.5, 1$ (i.e., stronger viability selection from left to right). Further, for each $\beta\Delta_0$, we consider random mating (grey) as well as populations with increasing assortment ($\gamma\Delta_0^2$ equal to 0.2, 0.5 and 1), which generate increasing levels of sexual selection on both sexes (Model I; top) or only on males (Model II; bottom). The analytical predictions obtained by solving eq. 3.4 (lines) match well with the results of deterministic simulations of a trait influenced by $L = 80$ loci (points) unless net selection (viability and/or sexual selection) is very strong; see, e.g., red curves corresponding to $\gamma\Delta_0^2 = 1.5$ in rightmost panel (with $\beta\Delta_0 = 1$). This discrepancy between simulations and theory for strong selection becomes weaker as the number of trait loci becomes larger (Appendix B.5).

Thus, there is a significant range of parameter space where simple approximations based on m_e (eq. (3.4)) predict equilibrium divergence accurately. Note that these approximations only account for the lower reproductive value of migrants due to sexual (and viability) selection, but ignore other effects of AM such as increased variance among residents, F_1s, etc, and the increased probability of mating between individuals with recent migrant ancestry. That there is, nonetheless, close agreement between theory and simulations

(which make no assumptions about AM), suggests that AM strengthens divergence in our model primarily by generating additional (sexual) selection against migrants and their descendants. This reduces m_e, thus increasing the genomewide barrier effect and facilitating divergence (see also *Discussion*).

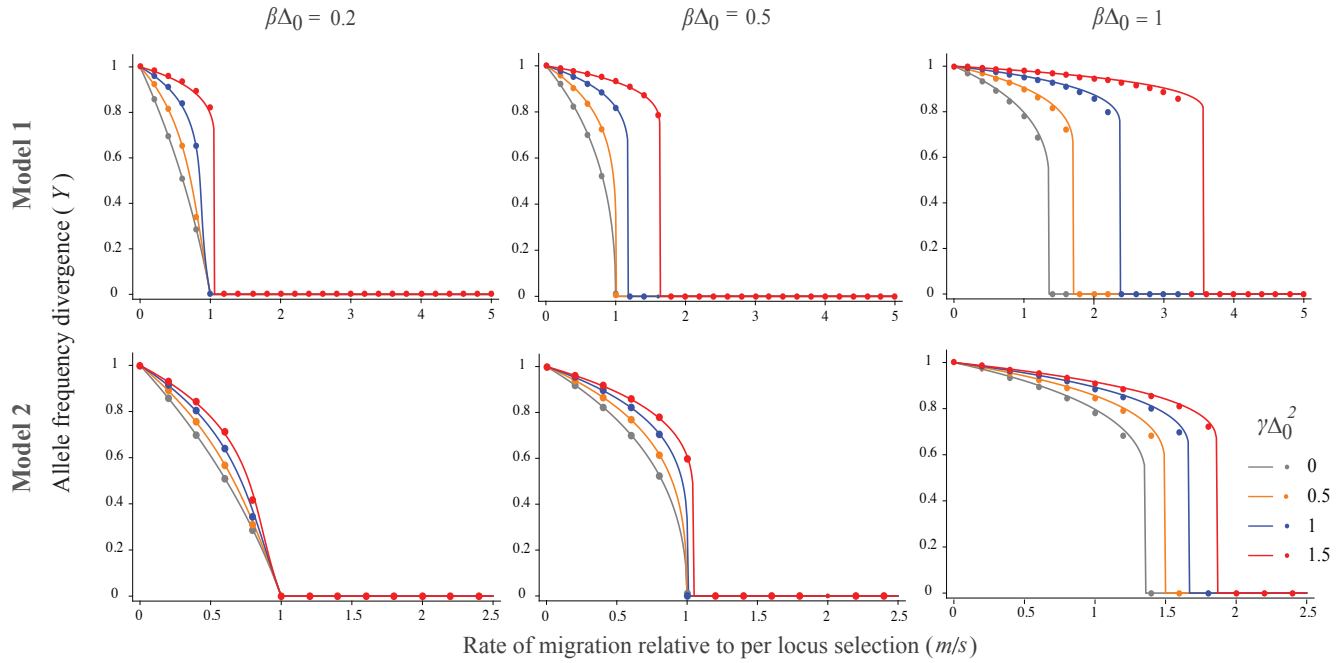


Figure 3.2: Allele frequency divergence per locus Y between mainland and island as a function of m/s, the migration rate relative to selection per locus. The upper and lower panels depict respective results for Model I (sexual selection on both sexes) and Model II (sexual selection only on males). Different subfigures within each panel correspond to different strengths of viability selection ($\beta\Delta_0 = 0.2, 0.5, 1$ from left to right); different colors within each plot correspond to different strengths of AM and sexual selection ($\gamma\Delta_0^2 = 0, 0.2, 0.5, 1$ shown in grey, orange, blue and red respectively). Solid lines show theoretical predictions (obtained by numerically solving eq. (3.4)) while dots represent the results of deterministic simulations of the hypergeometric model with $L = 80$ loci underlying the polygenic trait. Increasing viability and sexual selection cause sharper thresholds ('tipping points') for loss of divergence as well as an increase in the critical migration rate at which divergence is lost.

As evident from eq. (3.3), this reduction in m_e is stronger under Model I (where sexual selection acts on both sexes) as compared to Model II (sexual selection only on males), and also depends on existing levels of divergence (which in turn depend on the strength of migration and viability selection). For instance, consider the effects of moderately strong assortment ($\gamma\Delta_0^2 = 1$), which corresponds to a 63% vs. 32% reduction in sex-averaged mating success of migrants relative to residents under Model I vs. II under conditions of maximum divergence (i.e., if $\Delta = \Delta_0$). Under weak migration, say $m/s = 0.2$ and for $\beta\Delta_0 = 0.2$ (corresponding to rather weak viability selection), the effects of assortment are rather modest, increasing allele frequency divergence Y from 0.86 (under random mating)

to 0.96 and 0.91 under Models I and II respectively. However, with higher migration, e.g., for $m/s = 0.8$, the effect of assortment is more significant, increasing Y from 0.28 under random mating to 0.58 and 0.37 under Models I and II.

Figure 3.2 also highlights how migration affects divergence differently in parameter regimes characterised by strong vs. weak selection against migrants. When sexual selection (due to assortment) and viability selection are weak, divergence declines *smoothly* with increasing m/s, going to zero at a critical threshold $m_c/s = 1$, regardless of the actual value of $\beta\Delta_0$ and $\gamma\Delta_0^2$ (e.g., in leftmost panel in figure 3.2). Note that $m_c/s = 1$ is also the critical migration threshold for loss of adaptation at a single divergently selected locus. Thus, this threshold remains unchanged even if divergence is polygenic, as long as viability and sexual selection on the trait are weak.

By contrast, with strong viability and/or sexual selection, divergence at first decreases mildly with increasing migration, but then exhibits an *abrupt* collapse once it crosses a threshold Y_c (at a migration level $m/s \gtrsim 1$). Such 'tipping points' or discontinuous changes in divergence can be observed in the middle panel of figure 3.2 at higher levels of assortment, and in the rightmost panel regardless of AM. We can use eq. (3.3) to derive expressions for Y_c, the allele frequency divergence per locus associated with tipping points under the two models of AM (details in Appendix B.4):

$$Y_c = \frac{(4\gamma\Delta_0^2 - 3\beta\Delta_0) + \sqrt{(3\beta\Delta_0 + 4\gamma\Delta_0^2)^2 - 24\gamma\Delta_0^2}}{8\gamma\Delta_0^2} \quad \text{Model I} \tag{3.5a}$$

$$Y_c = 1 - \frac{1}{2\beta\Delta_0 + 2\gamma\Delta_0^2 Y_c \sum_{i=0}^{\infty} \frac{4^{-i}}{1+e^{\gamma\Delta_0^2 (Y_c^2/4^i)}}} \quad \text{Model II} \tag{3.5b}$$

For Model II, eq. (3.5b) specifies a transcendental equation for Y_c, which can be solved numerically (e.g., by approximating the sum by the first $\sim$ 20 terms). Further, in Appendix B.4, we show that the migration threshold associated with such a tipping point is: $m/s = (1 - Y_c)/g(\Delta_0 Y_c)$, where g, the gene flow factor, is given by eq. (3.3), and $\Delta_0 Y_c$ is the mean trait divergence at the tipping point.

The threshold Y_c can also be thought of as the minimum allele frequency difference per locus required to 'lock in' divergently selected alleles across the genome into distinct combinations that can be stably maintained by selection despite gene flow. Divergence levels lower than Y_c would result in an insufficient genomewide barrier to gene flow, causing a further decrease in divergence, further increasing m_e and so on, thus generating a tipping point. Interestingly, the threshold Y_c associated with the tipping point is independent of migration rate and only depends on the strengths of viability and sexual selection (eq. (3.5); see also figures 3.3A and 3.3B).

In addition to generating tipping points, strong viability or sexual selection (or both) can

also increase the critical migration threshold for loss of divergence *above* the single-locus threshold $m_c/s = 1$ (see for example, rightmost panel in figure 3.2). In other words, if selection against migrants and their descendants is sufficiently strong, then polygenic divergence can be maintained at migration levels at which a single divergently selected locus (in the absence of LD) would have been swamped.

Both tipping points and increased swamping thresholds may be explained as follows: increasing levels of migration reduce the adaptive divergence between the mainland and island, thereby increasing m_e (see eq. (3.3)), rendering multi-locus selection against locally deleterious alleles less effective, which increases their frequency, further reducing adaptive divergence and increasing m_e, which finally results in the swamping of locally adaptive alleles beyond a certain migration threshold. Crucially, the positive feedback between loss of divergence at individual loci and increase in m_e across the genome is stronger when net selection against divergent phenotypes (either the viability or sexual selection component) is higher. We now ask: how strong do viability and sexual selection need to be to produce these effects?

3.3.2 When does assortative mating and sexual selection lead to tipping points and higher swamping thresholds?

We first illustrate the various thresholds described above using a concrete example assuming weak viability selection ($\beta\Delta_0 = 0.2$) and sexual selection on both sexes (Model I). Under *weak* AM and sexual selection (in this example, $\gamma\Delta_0^2 < 1.2$), divergence decreases smoothly with increasing migration and is completely swamped at $m_c/s = 1$: this behaviour is qualitatively similar to that of a single divergently selected locus. For *intermediate* assortment (here, $1.2 < \gamma\Delta_0^2 < 1.4$), increasing migration leads to a tipping point or discontinuous change in divergence: this occurs once allele frequency divergence per locus approaches a threshold Y_c (given by eq. (3.5)) and involves a strong reduction in the genomewide barrier effect. However, in this intermediate assortment regime, the tipping point still occurs at a migration rate lower than the single-locus swamping threshold $m_c/s = 1$. For instance, for $\gamma\Delta_0^2 = 1.3$, the tipping point occurs at $m/s \approx 0.95$ and is associated with $Y_c \sim 0.6$: thus, a tiny increase in m/s (say from 0.949 to 0.95) causes allele frequency divergence per locus to fall from 0.6 to 0.1. A further increase in migration further reduces divergence, and complete swamping occurs (as before) at $m_c/s = 1$. However, with even *stronger* assortment (in this example, $\gamma\Delta_0^2 > 1.4$), the genomewide barrier effect becomes strong enough to maintain divergence beyond $m/s = 1$, so that the tipping point now occurs at $m_c/s > 1$. Further, divergence is completely swamped (Y goes down to zero) at the tipping point. Thus, in this strong assortment regime, tipping points and swamping thresholds coincide and occur at a critical migration threshold that exceeds the single-locus threshold.

We now ask: in what parameter regimes do we observe these qualitatively different behaviours? Figures 3.3A and 3.3B show the divergence threshold Y_c associated with tipping points (i.e., sharp drops in divergence), while figures 3.3C and 3.3D show the critical migration threshold m_c/s beyond which divergence is completely swamped, as a function of the strength of viability selection (measured as $\beta\Delta_0$). As discussed above, the swamping threshold m_c/s is also the migration level associated with the tipping point if assortment is strong. Figures 3.3A and 3.3C show Y_c and m_c/s for Model I (sexual selection on both sexes), while figures 3.3B and 3.3D show the corresponding thresholds for Model II (sexual selection on males). The various colors in each plot correspond to different values of $\gamma\Delta_0^2$, which quantifies the strength of sexual selection (due to AM). The predictions shown in figures 3.3A-3.3D all follow from eqs. (3.3)-(3.5) (see Appendix B.4 for details).

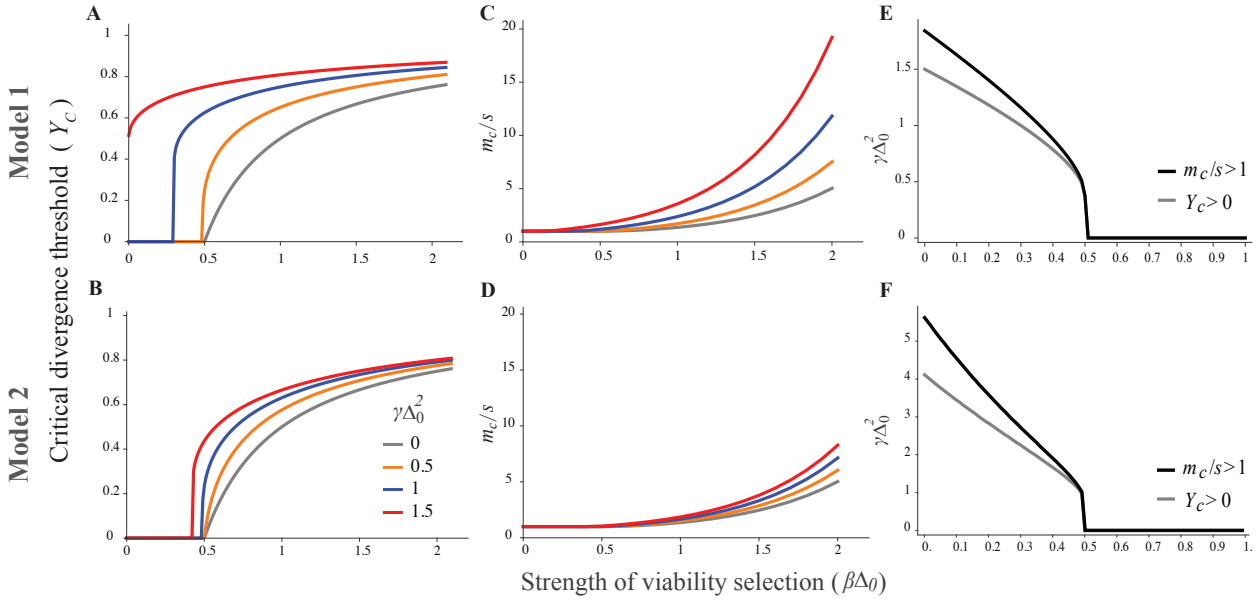


Figure 3.3: (A)-(B) Critical allele frequency divergence Y_c below which a tipping point occurs (i.e, divergence decreases sharply) vs. $\beta\Delta_0$, strength of viability selection. (C)-(D) Critical migration rate m_c/s at which divergence is completely lost vs. $\beta\Delta_0$. The different colors show results for different levels of AM and sexual selection ($\gamma\Delta_0^2$ =0, 0.5, 1 and 1.5). Upper panel (A, C, E) shows results for Model I (sexual selection on both sexes) and lower panel for Model II (sexual selection only on males). In panels A and B, parameter combinations with $Y_c = 0$ are those for which divergence decreases smoothly with increasing migration, i.e., there is no tipping point. In panels C and D, parameter combinations with $m_c/s = 1$ are those for which viability and sexual selection are too weak to shift the swamping threshold above the single-locus threshold ($m/s = 1$). (E)-(F) Grey curves show the minimum level of sexual selection (as measured by $\gamma\Delta_0^2$) required for $Y_c > 0$, i.e., to generate tipping points in divergence, as a function of $\beta\Delta_0$, the strength of viability selection. Black curves show the minimum level of sexual selection ($\gamma\Delta_0^2$) required for $m_c/s > 1$, i.e., to shift the swamping threshold above the single-locus threshold, as a function of $\beta\Delta_0$. Panels E and F correspond to Model I and II respectively. All results in this figure show (large L) theoretical predictions obtained from eqs. (3.3)-(3.5).

Under sufficiently strong viability selection ($\beta\Delta_0 > 0.5$), increased swamping thresholds ($m_c/s > 1$) emerge even in the absence of AM, as shown by the grey lines in figures 3.3C-3.3D (see also Sachdeva (2022)). Thus, in this regime, AM and sexual selection further increase the swamping threshold m_c/s (beyond 1) as well as the critical level of divergence Y_c that can be stably maintained. This can be explained, as before, in terms of the effects of sexual selection on m_e and the overall genomewide barrier to gene flow. As expected, AM has a stronger effect in Model I (where sexual selection is stronger) than in Model II. For example, for $\beta\Delta_0 = 1$, AM with strength $\gamma\Delta_0^2 = 1$ increases the critical migration threshold (m_c/s) from 1.36 (with random mating) to 2.37 under Model I and 1.66 under Model II. Note also that AM makes it harder to maintain intermediate levels of divergence, with divergence becoming unstable and collapsing once mean allele frequency difference per locus approaches 0.75 under Model I and 0.66 under Model II. By contrast, under random mating (and with $\beta\Delta_0 = 1$), we have $Y_c = 0.5$, so that somewhat lower levels of divergence can be stably maintained between the mainland and island, albeit at lower migration rates.

The situation is somewhat different if $\beta\Delta_0 < 0.5$; now, viability selection by itself (i.e., in the absence of AM) is too weak to generate tipping points or shift the swamping threshold m_c/s. However, sufficiently strong sexual selection due to AM will lead to one or both. More specifically, given viability selection $\beta\Delta_0 < 0.5$, we can determine an assortment threshold, i.e., a minimum value of $\gamma\Delta_0^2$ beyond which tipping points occur ($Y_c > 0$ in figures 3.3B and 3.3D), and a second threshold beyond which critical migration rates become higher than the single locus threshold (m_c/s exceeds 1 in figures 3.3A and 3.3C); see Appendix B.4 for details. These assortment thresholds are depicted by grey and black curves respectively in figures 3.3E and 3.3F. As expected, much higher levels of AM are required to generate tipping points and increase swamping thresholds when viability selection is weak, and when sexual selection acts only on males (instead of both sexes). For instance, for $\beta\Delta_0 = 0.2$, we require $\gamma\Delta_0^2$ to exceed 1.41 under Model I and 3.58 under Model II for $m_c/s > 1$. Note, however, that these thresholds are associated with much stronger sexual selection on migrants in the case of Model I (sexual selection on both sexes) than in Model II, despite weaker AM in the former case. For instance, the sex-averaged mating success of migrants relative to residents is $e^{-\gamma\Delta_0^2} \sim 0.24$ (for $\gamma\Delta_0^2 = 1.41$) under Model I and and $(1 + e^{-\gamma\Delta_0^2})/2 \sim 0.512$ (for $\gamma\Delta_0^2 = 3.58$) under Model II, both calculated by assuming maximum divergence between mainland and island.

3.3.3 Effect of viability and sexual selection on neutral gene flow

In the previous section, we have shown that approximations based on m_e accurately predict long-term adaptive divergence between the mainland and island under AM and sexual

selection. In essence, divergent selection at very many loci (whether due to viability or sexual selection) results in a genome-wide reduction in gene flow, which further strengthens divergence across the genome. As discussed above, this genome-wide reduction can be quantified via $g = m_e/m$, the gene flow factor at an unlinked locus, which may be viewed as a measure of RI between hybridising populations (see Westram et al. (2022)). Moreover, g can also be related to commonly used measures of neutral differentiation such as F_{ST}, which depends on both the number of migrants, Nm, and their reproductive value (which is equal to g). More concretely, at equilibrium, neutral F_{ST} between the mainland and island is: $F_{ST} \sim 1/(1 + 2Nm_e) = 1/(1 + 2Nm\,g)$ on average.

We now quantify the extent to which viability and sexual selection reduce gene flow across the genome in a scenario with high migration ($2Nm = 49$). Note that in the absence of divergent selection, such high levels of gene flow would erase most neutral differences between mainland and island, resulting in $F_{ST} = 0.02$. However, strong enough selection can maintain adaptive divergence despite high migration, thus suppressing gene flow and increasing genome-wide F_{ST}. Figure 3.4 illustrates this by plotting F_{ST} as a function of $\beta\Delta_0$ and $\gamma\Delta_0^2$ for $m/s = 1$, under the two models of AM. These figures are generated by solving for the allele frequency divergence per locus Y (using eq. (3.4)) for each combination of $\beta\Delta_0$ and $\gamma\Delta_0^2$, then using this to compute the gene flow factor g (via eq. 3.3), and finally computing $F_{ST} = 1/(1 + 2Nm\,g)$.

In each plot, the region to the left of the solid curve shows parameter combinations for which adaptive divergence is lost ($Y = 0$) at $m/s = 1$, so that $m_e = m$ and $F_{ST} = \frac{1}{1+2Nm} = 0.02$. As before, increasing $\beta\Delta_0$ and $\gamma\Delta_0^2$ (which correspond to stronger viability and sexual selection respectively) allow non-zero adaptive divergence to be maintained, thereby reducing m_e and increasing F_{ST}. The dashed curve in each plot depicts parameter combinations for which mean allele frequency divergence is $Y = 0.95$, which corresponds to a gene flow factor of $g \approx \frac{1-Y}{m/s} = 0.05$ and an F_{ST} value of $\frac{1}{1+2Nm\,g} \approx 0.29$ under either model of AM. To the right of this curve, i.e, with even stronger viability and/or sexual selection, there is a further increase in F_{ST}, which is now *not* due to an increase in adaptive divergence (which is close to its maximum possible value). Instead higher levels of F_{ST} in this parameter regime reflect lower viability and mating success of (nearly maximally diverged) migrants, which further reduces neutral gene exchange between mainland and island.

We further ask: do changes in the strength of viability and sexual selection multiplicatively (i.e., independently) affect gene flow (as measured by g), or are the combined effects of the two stronger (or weaker) than expected from their individual effects? More concretely, consider a scenario with moderately strong viability and sexual selection ($\beta\Delta_0 = \gamma\Delta_0^2 = 0.5$), $m/s = 1$, so that mean allele frequency divergence is $Y = 0.36$ (from eq. 3.4) and $g = 0.64$ (from eq. 3.3) under Model I. A 20% increase in *either* viability or

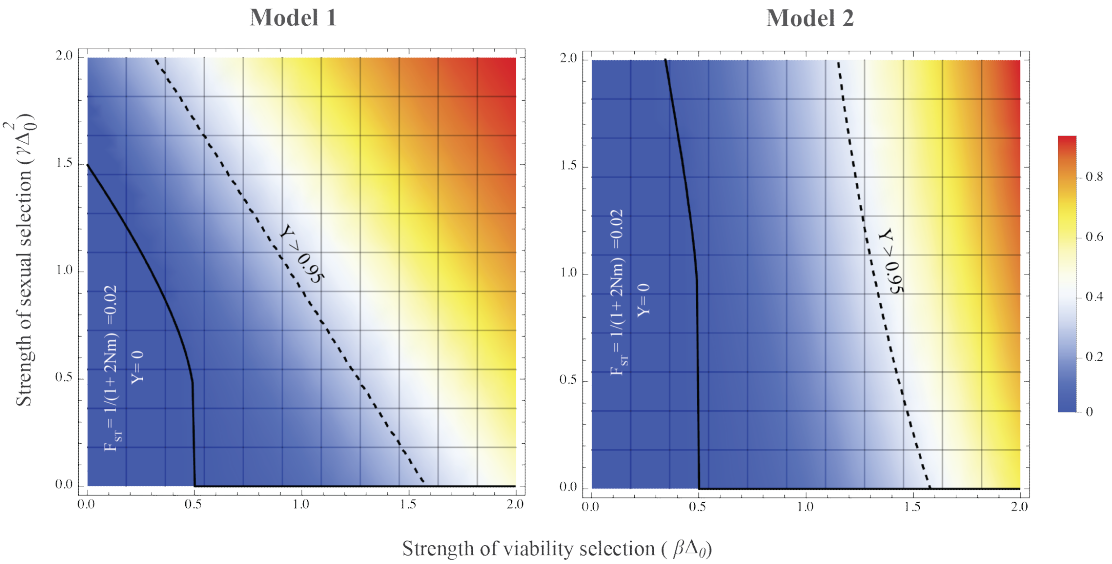


Figure 3.4: Heatmaps depicting F_{ST} at an unlinked neutral locus for different values of $\beta\Delta_0$ and $\gamma\Delta_0^2$ (which parameterize viability and sexual selection respectively) for Model I (sexual selection on both sexes) and Model II (sexual selection only on males) of assortative mating, assuming $m/s = 1$ and $2Nm = 49$. The color bar on the right shows F_{ST} values ranging from $F_{ST} = 0$ to 1. Parameter combinations to the left of the solid line are those for which there is no allele frequency divergence (i.e., $Y = 0$) so that $F_{ST} = 1/(1 + 2Nm) = 0.02$; parameters to the right of this line allow for divergence, so that $F_{ST} = 1/(1+2Nm\, g\,[\Delta_0 Y])$, where Y is obtained by solving eq. (3.4). Parameters to the right of the dashed line are those for which $Y > 0.95$: in this regime, stronger viability or sexual selection increases the genomewide barrier effect and F_{ST} without significantly affecting divergence at trait loci (which is close to its maximum possible value).

sexual selection, i.e., increasing $\beta\Delta_0$ to 0.6 while keeping $\gamma\Delta_0^2$ fixed at 0.5 (or vice versa) causes g to fall to 55% (or 74%) of its original value. Now, if viability and sexual selection were to independently reduce gene flow, then we would expect g to decrease by a factor of $0.55 \times 0.74 = 0.39$, i.e., to 39% of its original value in response to a 20% increase in *both* components of selection ($\beta\Delta_0 = \gamma\Delta_0^2 = 0.6$). Instead, g decreases to 43% of its original value. Thus, in this example, the two kinds of reproductive barriers (lower survival and lower mating success of immigrants and their descendants) cause a weaker reduction in gene flow than if they were to act independently.

Such non-independent effects arise because the gene flow factor is a product of two terms– one due to viability and the other due to sexual selection (see eq. (3.3)), both of which depend on allele frequency divergence Y, which in turn is influenced by both components of selection. Thus, the strongest non-multiplicative effects emerge in parameter regimes where viability and sexual selection are strong enough to significantly influence divergence levels but not so strong as to drive divergence to its maximum possible value. In the latter scenario, i.e., as Y approaches 1 (to the right of the dashed curves in figure 3.4), the effect of viability and sexual selection again becomes multiplicative.

3.4 Discussion

While it has long been recognised that sexual selection can maintain inter-specific differences, questions remain as to its importance relative to other processes during speciation (Safran et al. (2013)). Sexual selection can vary significantly across species and even across environments within a species or over time (Schumer et al. (2017); Magurran and Ramnarine (2004); Rosenthal (2013)), making it difficult to identify broad patterns. Moreover, sexual selection typically acts together with other kinds of selection against heterospecifics and hybrids (e.g., due to ecological mismatch or expression of genetic incompatibilities), making it difficult to disentangle the contributions of different processes to reproductive isolation. This only underscores the broader challenges in quantifying RI– an issue that has attracted considerable attention recently (Westram et al. (2022)).

Here, we examine some of these issues in the context of a polygenic 'magic' trait that is under divergent ecological selection across populations, while also mediating AM and sexual selection. In such a scenario, long-term polygenic divergence depends crucially on LD between trait loci. Such LD is, in turn, a consequence of both (natural and sexual) selection and assortment, making it challenging to model. Previous analyses have relied on one of two approaches- either focusing on how trait variances are affected by selection and assortment while largely ignoring genetic details (e.g., Doebeli (1996); Kondrashov and Shpak (1998)), or explicitly modelling LD between sets of loci (Kirkpatrick (2000); Barton and De Cara (2009)), but assuming LD to be weak (to ensure tractability). Here, we employ a novel approach based on effective migration rates which bridges phenotypic descriptions (involving trait means and variances) and genic descriptions (in terms of divergence at individual selected sites). As we demonstrate, theory based on m_e can accurately predict divergence across a range of parameters, including when LD is strong enough to generate so-called tipping points (Nosil et al. (2017)). An advantage of this approach is that it provides an economical description of tipping points in terms of just allele frequency differences between populations rather than a plethora of LD measures.

The key idea behind our approach is that LD between introgressing alleles breaks down rapidly if these are unlinked or loosely linked (a standard assumption in quantitative genetics), allowing us to represent its effects via an effective migration rate m_e (or alternatively, the gene flow factor $g = m_e/m$) over the longer timescales over which allele frequencies change. The gene flow factor, being the average reproductive value of migrants, is a composite measure of various processes- prezygotic and postzygotic- that affect the fitness of migrants and their descendants. In particular, it captures both the prezygotic effects of AM– which causes fewer F_1 hybrids to be produced (due to sexual selection against migrants) as well as its postzygotic effects– as reflected in the lower mating success of F_1s and later generation hybrids (again, due to sexual selection).

We show that to good approximation, g is a product of two terms- one reflecting viability selection and the other sexual selection on migrants and their descendants (see eq. 3.3). This is in line with previous work that quantifies sexual and natural selection via probabilities of conspecific vs. heterospecific mating, or fitness of offspring from different types of mating, and then assumes that the two components of selection multiplicatively determine the total barrier to gene flow (Ramsey et al. (2003); Sobel and Chen (2014)). Note however, that in our model, the two components (due to viability and sexual selection) both depend on the existing level of divergence Δ, and thus are not independent. For instance, stronger AM and sexual selection will typically increase divergence, which will also reduce the (relative) viability of migrants and their descendants, thus further strengthening the barrier effect of viability selection, even when there is no change in the strength of viability selection. This suggests that measurements of RI are highly context and state dependent, making it difficult to extrapolate from lab measurements or between replicate hybrid zones.

Nevertheless, eq. 3.3 provides an unambiguous way of quantifying the relative contributions of natural and sexual selection to RI, at least in theory. In practice, this requires estimates of various fitness components of individuals with different levels of migrant ancestry, which are only available for a handful of well-studied populations with multi-generational pedigrees (Bérénos et al. (2014); Chen et al. (2019)). However, these populations are necessarily spatially limited and have very little divergence or RI. More typically, we might have indirect fitness estimates of hybrids (and less commonly of backcrosses and recombinants), e.g., from genotype frequencies in hybrid zones between divergent ecotypes or subspecies. However, in such cases, it is usually difficult to estimate individual fitness components. Some progress may be made by combining indirect estimates of total fitness with reciprocal transplant experiments (to measure viability) or mate choice experiments (to measure mating success). It is an open question whether such composite approaches can provide reliable estimates of fitness components in natural populations, thus allowing us to disentangle the contributions of natural and sexual selection to divergence and RI.

Our work also bears upon the interplay between assortative mating and sexual selection during divergence with gene flow. The distinction between these two processes has been highlighted in earlier work (Kirkpatrick and Nuismer (2004); Polechová and Barton (2005); Servedio and Bürger (2014)), which points out that sexual selection can reduce the mating success of outlier individuals, thus constraining genetic variance and reducing the potential for divergence (Kirkpatrick and Nuismer (2004)). By contrast, pure assortment without sexual selection always increases variance, thus facilitating divergence.

The situation is somewhat different following secondary contact between diverged populations: now, sexual selection against migrants reduces introgression of deleterious alleles. This effect is only strengthened by assortative mating (based on ancestry) which increases

LD between introgressing alleles, causing them to be eliminated more effectively. In recent work, Muralidhar et al. (2022) provide a rough estimate of the effect of AM following a pulse of admixture, where an individual's viability declines in proportion to the amount of introgressed material it carries. Their argument is as follows: the change in introgressed allele frequency per generation (due to viability and/or sexual selection) is proportional to the variance of introgressed ancestry across individuals. In an assortatively mating population with correlation coefficient ρ between mates (where ρ is a measure of the strength of AM), the variance in introgressed ancestry declines by a factor $\sim (1+\rho)/2$ per generation, *slower* than it would under random mating, allowing for more efficient selective purging of introgressing alleles. It then follows that AM *by itself* (as a process that is distinct from sexual selection) increases net purging by a factor of $1/(1-\rho)$ (Muralidhar et al. (2022)). This argument implicitly assumes that sexual selection is weak and does not cancel out the variance-increasing effect of AM; this assumption would, for instance, break down under our Model I (see also Kirkpatrick and Nuismer (2004))

The conclusions of Muralidhar et al. (2022) differ from ours– in particular, we show that divergence is shaped essentially by sexual selection on various subgroups (i.e., F_1s, BC_1s etc.), with little to no effect of AM (e.g., on within-subgroup variances). What might explain these differing conclusions? Muralidhar et al. (2022) analyse the consequences of an admixture event that replaces a sizable fraction of the native population by migrants from a diverged population. In this case, individuals with recent migrant ancestry are sufficiently common to mate preferentially with one another under AM: this causes LD between introgressing alleles to persist longer (and variance of introgressed ancestry to decline slower), thus also allowing selection to eliminate them efficiently. By contrast, we consider ongoing migration between two populations at rate m, and focus on the long-term equilibrium between selection and migration. In this case, long-term adaptive divergence requires m_e to be lower than the typical per locus selective effect s, so that m is at most a few times larger than s. Thus, migration is sufficiently low that individuals with recent migrant ancestry mate predominantly with residents rather than with each other. Then, introgression depends essentially on the mating success of (or alternatively, sexual selection on) various hybrids (F_1, BC_1 etc.) relative to residents. We emphasise that assuming low migration rates ($m \ll 1$) does not imply that neutral divergence between populations is high, since the latter depends on the number of migrants Nm per generation, which can be large. Nor does it imply high levels of adaptive divergence (see, e.g., figure 3.2); indeed, this can be arbitrarily low, depending on m/s, and the total strength of viability and sexual selection.

The preceding discussion highlights how the role of AM (as distinct from sexual selection) depends on the context of divergence. Broadly speaking, assortment is likely to be important if phenotypically diverged subgroups are present at comparable frequencies,

as may be the case in the centre of a hybrid zone. In such a scenario, our theoretical approximations would no longer apply, since F_1s would be sufficiently common as to mate amongst themselves and generate F_2 recombinants. Moreover, many hybrid zones are characterised by a bimodal distribution of parental ancestries or equivalently, of hybrid indices (Jiggins and Mallet (2000); Bridle and Butlin (2002); Schumer et al. (2017)). The effects of AM in such a scenario are not amenable to simple descriptions which assume trait variance to increase by a constant factor $(1+\rho)/2$ (Muralidhar et al. (2022)), since the correlation ρ between mating pairs in an assortatively mating population also depends on the distribution of phenotypic values and can be very different for unimodal vs. bimodal populations. The consequences of non-random mating in hybrid zones thus remain an important direction for future work (see also Irwin (2020)), which will require us to expand existing theoretical descriptions.

What are some limitations of our study? We have considered a model of AM via self-referencing, where individuals prefer mates with similar phenotypes. However, sexual selection may be less effective in suppressing gene flow if female preference and the male traits on which it acts are encoded by different genes, requiring these to be in LD. More generally, little is known about the prevalence of different mechanisms of AM in nature and the extent to which these generate sexual selection in different ecological contexts (Safran et al. (2013); Kopp et al. (2018)), making it important to examine the robustness of theoretical findings to alternative assumptions about assortment mechanisms.

We also consider an extreme form of divergent ecological selection, where alternative alleles are favoured in the two populations at every locus. In this case, sexual selection is predominantly directional, reducing the mating success of individuals in proportion to their migrant ancestry. Alternatively, the magic trait could be under stabilizing selection towards different trait optima in the two populations (Sachdeva and Barton (2017)). While selection on migrants would still be directional if trait optima are far apart, adaptive divergence can now be maintained at higher migration rates, allowing for a more complicated interplay between assortment and sexual selection– as mating between individuals with recent migrant ancestry becomes more common. Importantly, in this case, we need to go beyond theory that assumes introgressing alleles as cascading down successively fitter genetic backgrounds and account for F2s etc.

Finally, we neglect linkage between trait loci as well as genetic drift. In general, linkage between introgressing alleles strengthens barriers to gene flow, allowing divergence to be maintained despite stronger migration (Barton and Bengtsson (1986)). Conversely, drift reduces the efficiency of selection at individual loci and also increases the frequency of genotypes containing alleles with opposing effects on fitness: this reduces overall selection against introgressing alleles, making smaller populations more prone to swamping by gene flow. Approximations based on effective migration rates can be extended to include these

complexities (Sachdeva (2022); Zwaenepoel et al. (2024)), making it straightforward to generalise our analysis to more realistic architectures of magic traits.

We focus here on the interplay between natural and sexual selection given a *fixed* level of assortment and following secondary contact between diverged populations. Other questions remain as to how and when mating preferences evolve– e.g., under what conditions mating with conspecifics might be favoured by selection, thus reinforcing divergence (Servedio and Noor (2003)). While our study does not address these questions, it does suggest that analysing the evolution of assortment in terms of differences in reproductive value (or effective rates of exchange) of modifiers of mate choice could be a fruitful direction for future work. More generally, effective migration rates link a 'phenotypic' view of RI (based on fitness components of hybrids) and a 'genic' view (that conceptualises RI as a process that reduces gene flow and increases divergence across the genome), making them a powerful way of analysing various aspects of RI.

Acknowledgements: We thank Nick Barton for useful comments on the manuscript. This research was supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

Data availability: Mathematica notebooks for numerical analysis and simulations are available at 10.15479/AT:ISTA:17344.

Conflict of Interest: The authors declare that there is no conflict of interest.

References

Barton, N. (1992). On the spread of new gene combinations in the third phase of Wright's shifting-balance. *Evolution*, pages 551–557.

Barton, N. and Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57(3):357–376.

Barton, N. H. (1983). Multilocus clines. *Evolution*, 37(3):454–471.

Barton, N. H. and De Cara, M. A. R. (2009). The evolution of strong reproductive isolation. *Evolution*, 63(5):1171–1190.

Barton, N. H. and Etheridge, A. M. (2011). The relation between reproductive value and genetic contribution. *Genetics*, 188(4):953–973.

Barton, N. H., Etheridge, A. M., and Véber, A. (2017). The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:50–73.

Bengtsson, B. (1985). The flow of genes through a genetic barrier. In *Evolution: Essays in honour of John Maynard Smith*, pages 31–42. Cambridge University Press, Cambridge.

Bridle, J. R. and Butlin, R. K. (2002). Mating signal variation and bimodality in a mosaic hybrid zone between *Chorthippus* grasshopper species. *Evolution*, 56(6):1184–1198.

Butlin, R. K. and Smadja, C. M. (2018). Coupling, reinforcement, and speciation. *American Naturalist*, 191(2):155–172.

Bérénos, C., Ellis, P. A., Pilkington, J. G., and Pemberton, J. M. (2014). Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. *Molecular Ecology*, 23(14):3434–3451.

Chen, N., Juric, I., Cosgrove, E. J., Bowman, R., Fitzpatrick, J. W., Schoech, S. J., Clark, A. G., and Coop, G. (2019). Allele frequency dynamics in a pedigreed natural population. *Proceedings of the National Academy of Sciences*, 116(6):2158–2164.

Coughlan, J. M. and Matute, D. R. (2020). The importance of intrinsic postzygotic barriers throughout the speciation process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1806):20190533.

Coyne, J. A. and Allen Orr, H. (2004). *Speciation*. Sunderland, Mass.: Sinauer Associates.

Crow, J. F. and Felsenstein, J. (1968). The effect of assortative mating on the genetic composition of a population. *Eugenics Quarterly*, 15(2):85–97.

Doebeli, M. (1996). A quantitative genetic competition model for sympatric speciation. *Journal of Evolutionary Biology*, 9(6):893–909.

Elmer, K. R., Lehtonen, T. K., and Meyer, A. (2009). Color assortative mating contributes to sympatric divergence of neotropical cichlid fish. *Evolution*, 63(10):2750–2757.

Felsenstein, J. (1981). Skepticism Towards Santa Rosalia, or Why are There so Few Kinds of Animals? *Evolution*, 35(1):124–138.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Proc. Roy. Soc. Edinburgh*, 52:399–433.

Fisher, R. A. (1999). *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.

Fraïsse, C. and Sachdeva, H. (2021). The rates of introgression and barriers to genetic exchange between hybridizing species: sex chromosomes *vs* autosomes. *Genetics*, 217(2):iyaa025.

Irwin, D. E. (2020). Assortative mating in hybrid zones is remarkably ineffective in promoting speciation. *The American Naturalist*, 195(6):E150–E167.

Jiang, Y., Bolnick, D. I., and Kirkpatrick, M. (2013). Assortative mating in animals. *The American Naturalist*, 181(6):E125–138.

Jiggins, C. D. and Mallet, J. (2000). Bimodal hybrid zones and speciation. *Trends in Ecology & Evolution*, 15(6):250–255.

Jiggins, C. D., Naisbit, R. E., Coe, R. L., and Mallet, J. (2001). Reproductive isolation caused by colour pattern mimicry. *Nature*, 411(6835):302–305.

Johannesson, K., Havenhand, J. N., Jonsson, P. R., Lindegarth, M., Sundin, A., and Hollander, J. (2008). Male discrimination of female mucous trails permits assortative mating in a marine snail species. *Evolution*, 62(12):3178–3184.

Kirkpatrick, M. (2000). Reinforcement and divergence under assortative mating. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1453):1649–1655.

Kirkpatrick, M. and Nuismer, S. L. (2004). Sexual selection can constrain sympatric speciation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1540):687–693.

Kirkpatrick, M. and Ravigné, V. (2002). Speciation by Natural and Sexual Selection: Models and Experiments. *The American Naturalist*, 159(S3):S22–S35.

Kobayashi, Y., Hammerstein, P., and Telschow, A. (2008). The neutral effective migration rate in a mainland-island context. *Theoretical Population Biology*, 74(1):84–92.

Kondrashov, A. S. and Shpak, M. (1998). On the origin of species by means of assortative mating. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1412):2273–2278.

Kopp, M., Servedio, M. R., Mendelson, T. C., Safran, R. J., Rodríguez, R. L., Hauber, M. E., Scordato, E. C., Symes, L. B., Balakrishnan, C. N., Zonana, D. M., and van Doorn, G. S. (2018). Mechanisms of assortative mating in speciation with gene flow: Connecting theory and empirical research. *The American Naturalist*, 191(1):1–20.

Lewontin, R., Kirk, D., and Crow, J. (1968). Selective mating, assortative mating, and inbreeding: definitions and implications. *Eugenics Quarterly*, 15(2):141–143.

Lynch, M., Walsh, B., et al. (1998). *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Maan, M. E. and Seehausen, O. (2011). Ecology, sexual selection and speciation. *Ecology Letters*, 14(6):591–602.

Magurran, A. E. and Ramnarine, I. W. (2004). Learned mate recognition and reproductive isolation in guppies. *Animal Behaviour*, 67(6):1077–1082.

Merrill, R. M., Wallbank, R. W. R., Bull, V., Salazar, P. C. A., Mallet, J., Stevens, M., and Jiggins, C. D. (2012). Disruptive ecological selection on a mating cue. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749):4907–4913.

Muralidhar, P., Coop, G., and Veller, C. (2022). Assortative mating enhances postzygotic barriers to gene flow via ancestry bundling. *Proceedings of the National Academy of Sciences*, 119(30):e2122179119.

Nosil, P., Feder, J. L., Flaxman, S. M., and Gompert, Z. (2017). Tipping points in the dynamics of speciation. *Nature Ecology & Evolution*, 1(2):0001.

Perini, S., Rafajlović, M., Westram, A. M., Johannesson, K., and Butlin, R. K. (2020). Assortative mating, sexual selection, and their consequences for gene flow in *Littorina*. *Evolution*, 74(7):1482–1497.

Polechová, J. and Barton, N. H. (2005). Speciation through competition: A critical review. *Evolution*, 59(6):1194–1210.

Ramsey, J., Bradshaw Jr, H., and Schemske, D. W. (2003). Components of reproductive isolation between the monkeyflowers *Mimulus lewisii* and *M. cardinalis* (phrymaceae). *Evolution*, 57(7):1520–1534.

Rice, W. R. and Hostert, E. E. (1993). Laboratory experiments on speciation: what have we learned in 40 years? *Evolution*, 47(6):1637–1653.

Rosenthal, G. G. (2013). Individual mating decisions and hybridization. *Journal of Evolutionary Biology*, 26(2):252–255.

Rundle, H. D. and Nosil, P. (2005). Ecological speciation. *Ecology Letters*, 8(3):336–352.

Sachdeva, H. (2022). Reproductive isolation via polygenic local adaptation in sub-divided populations: Effect of linkage disequilibria and drift. *PLOS Genetics*, 18(9):e1010297.

Sachdeva, H. and Barton, N. H. (2017). Divergence and evolution of assortative mating in a polygenic trait model of speciation with gene flow. *Evolution*, 71(6):1478–1493.

Safran, R. J., Scordato, E. S., Symes, L. B., Rodríguez, R. L., and Mendelson, T. C. (2013). Contributions of natural and sexual selection to the evolution of premating reproductive isolation: a research agenda. *Trends in Ecology & Evolution*, 28(11):643–650.

Schumer, M., Powell, D. L., Delclós, P. J., Squire, M., Cui, R., Andolfatto, P., and Rosenthal, G. G. (2017). Assortative mating and persistent reproductive isolation in hybrids. *Proceedings of the National Academy of Sciences*, 114(41):10936–10941.

Servedio, M. R. (2009). The role of linkage disequilibrium in the evolution of premating isolation. *Heredity*, 102(1):51–56.

Servedio, M. R. (2016). Geography, assortative mating, and the effects of sexual selection on speciation with gene flow. *Evolutionary Applications*, 9(1):91–102.

Servedio, M. R. and Boughman, J. W. (2017). The role of sexual selection in local adaptation and speciation. *Annual Review of Ecology, Evolution, and Systematics*, 48(1):85–109.

Servedio, M. R. and Bürger, R. (2014). The counterintuitive role of sexual selection in species maintenance and speciation. *Proceedings of the National Academy of Sciences*, 111(22):8113–8118.

Servedio, M. R., Doorn, G. S. V., Kopp, M., Frame, A. M., and Nosil, P. (2011a). Magic traits in speciation: 'magic' but not rare? *Trends in Ecology & Evolution*, 26(8):389–397.

Servedio, M. R., Doorn, G. S. V., Kopp, M., Frame, A. M., and Nosil, P. (2011b). Magic traits in speciation: 'magic' but not rare? *Trends in Ecology & Evolution*, 26(8):389–397.

Servedio, M. R. and Kopp, M. (2012). Sexual selection and magic traits in speciation with gene flow. *Current Zoology*, 58(3):510–516.

Servedio, M. R. and Noor, M. A. (2003). The Role of Reinforcement in Speciation: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics*, 34(1):339–364.

Smadja, C. M. and Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, 20(24):5123–5140.

Smith, J. M. (1966). Sympatric Speciation. *The American Naturalist*, 100(916):637–650.

Sobel, J. M. and Chen, G. F. (2014). Unification of Methods for Estimating the Strength of Reproductive Isolation. *Evolution*, 68(5):1511–1522.

Stelkens, R. B. and Seehausen, O. (2009). Phenotypic divergence but not genetic distance predicts assortative mating among species of a cichlid fish radiation. *Journal of Evolutionary Biology*, 22(8):1679–1694.

Vines, T. H. and Schluter, D. (2006). Strong assortative mating between allopatric sticklebacks as a by-product of adaptation to different environments. *Proceedings of the Royal Society B: Biological Sciences*, 273(1589):911–916.

Westram, A. M., Stankowski, S., Surendranadh, P., and Barton, N. (2022). What is reproductive isolation? *Journal of Evolutionary Biology*, 35(9):1143–1164.

Wright, S. (1921). Systems of Mating. III. Assortative mating based on somatic resemblance. *Genetics*, 6(2):144–161.

Zwaenepoel, A., Sachdeva, H., and Fraïsse, C. (2024). The genetic architecture of polygenic local adaptation and its role in shaping barriers to gene flow. *Genetics*, page iyae140.

Chapter 4

Effects of fine-scale population structure on the distribution of heterozygosity in a long-term study of *Antirrhinum majus*[1]

Abstract

Many studies have quantified the distribution of heterozygosity and relatedness in natural populations, but few have examined the demographic processes driving these patterns. In this study, we take a novel approach by studying how population structure affects both pairwise identity and the distribution of heterozygosity in a natural population of the self-incompatible plant *Antirrhinum majus*. Excess variance in heterozygosity between individuals is due to identity disequilibrium (ID), which reflects the variance in inbreeding between individuals; it is measured by the statistic g_2. We calculated g_2 together with F_{ST} and pairwise relatedness (F_{ij}) using 91 SNPs in 22,353 individuals collected over 11 years. We find that pairwise F_{ij} declines rapidly over short spatial scales, and the excess variance in heterozygosity between individuals reflects significant variation in inbreeding. Additionally, we detect an excess of individuals with around half the average heterozygosity, indicating either selfing or matings between close relatives. We use two types of simulation to ask whether variation in heterozygosity is consistent with fine-scale spatial population structure. First, by simulating offspring using parents drawn from a range of spatial scales, we show that the known pollen dispersal kernel explains g_2. Second, we simulate a 1000-generation pedigree using the known dispersal and spatial distribution and find that the resulting g_2 is consistent with that observed from the

[1]This work can be found online at https://doi.org/10.1093/genetics/iyac083

field data. In contrast, a simulated population with uniform density underestimates g_2, indicating that heterogeneous density promotes identity disequilibrium. Our study shows that heterogeneous density and leptokurtic dispersal can together explain the distribution of heterozygosity.

Keywords heterozygosity, identity disequilibrium, population structure, isolation by distance

4.1 Introduction

For most organisms, gene dispersal and therefore relatedness are spatially structured, such that individuals closer in space are more likely to mate, and be more closely related, than individuals further apart (Wright, 1946; Vekemans and Hardy, 2004). Such spatial population structure causes decreasing genetic similarity with geographic distance (isolation-by-distance (Wright, 1943)); this reduces the mean heterozygosity of the whole population relative to a well-mixed population. Despite the ubiquity of these patterns in nature, the role of demography and gene dispersal in determining the spatial pattern of genetic variation has not been thoroughly explored. Commonly used spatial models typically assume discrete demes and/or a uniform population density. However, natural populations are typically patchy, with heterogeneity in both the distribution and density of individuals. Patchy and heterogeneous spatial distributions within natural populations should result in spatial variation in inbreeding and, consequently, excess variance in heterozygosity. Despite this prediction, the effect of spatial heterogeneity on heterozygosity has rarely been examined in the population structure literature. Moreover, it is the interplay of heterogeneous density and dispersal that likely shapes the spatial structuring of genetic relatedness between individuals. This highlights the importance of understanding the factors (e.g., life history, demography, population structure) that contribute to shaping the full distribution of heterozygosity and relatedness in a spatially structured population.

Understanding the drivers of variation in inbreeding within populations is fundamental, given its importance to genetic diversity and to fitness. Quantifying variation in inbreeding and combining this with measures of fitness (or fitness proxies) makes it possible, in principle, to estimate inbreeding depression either through pedigrees Charlesworth and Charlesworth (1987); Lynch and Walsh (1996) or heterozygosity-fitness correlations (HFCs). For HFCs, inbreeding depression is estimated by comparing proxy measures of fitness against heterozygosity, with the expectation that offspring from related individuals will have lower heterozygosity. Variance in inbreeding is therefore essential for HFCs to be detected (Szulkin et al., 2010). In addition, variance in inbreeding is interesting per se because it depends on both demographic history (e.g., (Sin et al., 2021)) and mating

system (selfing, partial selfing or outcrossing) (Winn et al., 2011)). Outcrossing species, with generally low levels of inbreeding, provide an opportunity to examine factors other than mating system variation that may affect inbreeding variation, and thus, variance in heterozygosity.

If there is variation in inbreeding between individuals, heterozygosity at different loci will be correlated. The covariance between loci in heterozygous state is termed identity disequilibrium (ID), by analogy with linkage disequilibrium, which is the covariance in allelic state between loci. ID can be calculated across individuals and divided by the square of the mean heterozygosity to calculate the population statistic g_2, which is a measure of variance in identity by descent (Szulkin et al., 2010) amongst individuals. For an outcrossing organism with fine-scale population structure, spatial patterns of density and mating could have strong effects on the degree of mating with related individuals, and thus affect identity disequilibrium and g_2. Furthermore, as sessile organisms, mating and offspring dispersal in plants are mediated by external vectors (pollinators and seed dispersal mechanisms) (Loveless and Hamrick, 1984). Consequently, the shape of the distribution of dispersal of both pollen and seed will also have an impact on g_2. Additionally, as partial selfing will produce identity disequilibria across loci for selfed individuals, g_2 can be used to estimate the selfing rate of a population, with this estimator being robust to null alleles and biparental inbreeding (David et al., 2007; Hardy, 2016). If the sources of variation in inbreeding are better understood, we may be able to combine g_2 with other statistics of population structure to improve inferences about demographic history (Milligan et al., 2018; Bradburd and Ralph, 2019).

For over a decade, we have sampled a population of the self-incompatible plant *Antirrhinum majus*, the long-term aim being to build a pedigree that will allow us to estimate fitness and dispersal directly. Through that project, we have collected an exceptionally large sample of individuals with SNP genotypes that are spatially mapped. This dataset enables a powerful test of whether the observed density and dispersal in this population can account for both the decay of pairwise relatedness with distance, and for the distribution of heterozygosity across individuals. Here, we first verify that there is excess variance in heterozygosity, which reflects an underlying variance in inbreeding. Second, to understand the role of spatial patterns of dispersal in generating variance in heterozygosity, we compare the empirical distribution of heterozygosity with that of offspring from simulated matings where parents were drawn from different dispersal scales. Third, we ask whether heterogeneous population density promotes variation in inbreeding, by comparing simulated pedigrees conditioned on uniform density versus on the observed locations of plants. Taken together, addressing these questions provides insight into the underlying drivers of the distribution of heterozygosity and relatedness, and provides novel ways to study the effects of mating patterns and demography in nature.

4.2 Methods

Study system

Antirrhinum majus is a self-incompatible, hermaphroditic, short-lived perennial herb native to the Iberian Peninsula. It has a seed bank with most individuals' parents recorded 3-4 years before they are sampled (D. Field, unpublished data). It grows in a variety of microhabitats with relatively bare soil or frequent disturbance, including rail embankments, rocky cliffs, and regularly mowed roadsides. Our study includes two "subspecies" that differ only in flower color: *A. majus pseudomajus* has magenta flowers and occurs in northern Spain and south-western France, including the Pyrenees. *A. majus striatum* has yellow flowers and a smaller range, encircled by *A. m. pseudomajus*. The subspecies are parapatric; narrow clines with intermediate color hybrids form wherever they meet, and there is no evidence for post-zygotic reproductive barriers (Andalo et al., 2010). We focus on such a hybrid zone in the Vall de Ribès, Spain (Whibley et al., 2006), where we have collected demographic data annually since 2009. Across nearly all of the genome, there is little divergence within our study area between plants with different flower color, except for limited regions associated with floral pigmentation, which show steep clines (Tavares et al., 2018). Thus, the study area can be considered as a single population for studying neutral genetic variation.

Figure 4.1: Distribution of *A. majus* individuals (shown as white circles) in Vall de Ribès, Spain from the years 2009 to 2019.

Field sampling

Genetic samples were obtained annually from 2009-2019 from every accessible flowering individual in ∼ 5 km stretches of two parallel roads that cross the Vall de Ribès, dubbed the "lower road" (GIV-4016; ∼ 1150 m elevation) and "upper road" (N-260; ∼1350 m) (4.1). We also sampled along small side roads, railroad embankments, rivers, and hiking trails. The plants grow preferentially along exposed areas such as roads, therefore, density was very low away from these disturbed areas between the main sampling sites of the

lower and upper roads. In some years, we were limited to genotyping only in the core area, ∼1 km along each road. The total genotyped sample summed over the eleven years is 22,353 plants, ranging from ∼750 plants in the smallest year (2018), to ∼5500 plants in the largest year (2014). Eighteen percent of individuals were sampled in more than one year. Sampling was conducted during peak flowering (early June to late-July). Each year there were fewer than 100 visible but inaccessible plants; consequently, we estimate that we found the majority of individuals in the sampled area.

For each plant, we collected leaf samples for genotyping, and recorded spatial locations with GeoXT handheld GPS units (Trimble, Sunnydale, CA, USA). These devices are accurate to within 3.7 m, determined by the mean distance between samples comparing samples that had been inadvertently recorded twice in the field (individuals with similar geographic location and near-identical genotypes, allowing for SNP errors). Leaf samples were refrigerated upon return to the field station, dried in silica gel and stored for several weeks.

SNP panel

Previously, a panel of 248 SNPs spread throughout the genome was designed for the focal population (see methods in (Ringbauer et al., 2018)). We follow these methods but include an additional five years of data (2015-2019) and use a subset of 91 non-clinal SNPs; the mean sample size per SNP was 21,212, or 95% of the total. (see Appendix D.1 for SNP filtering methods).

Identity by descent vs identity in state

Throughout this paper, it will be important to distinguish between identity by descent (IBD) and identity in state. We denote the probability that two genes are identical by descent by F; this is defined relative to an ancestral reference population, and can in principle be calculated from the pedigree that descends from that population, independent of the actual allelic state. What we observe are biallelic SNP genotypes; the two homologous genes in a diploid individual will be identical in state if the genes are identical by descent, or if the ancestral genes carried the same allele. Thus, probabilities of identity by descent (F) can be estimated from observed identities in state. We denote the heterozygosity at locus i in a particular individual by h_i, with $h_i = 0$ if the genes are identical in state, and $h_i = 1$ otherwise. The mean heterozygosity of an individual is the average of h_i over n loci, denoted multilocus heterozygosity $H = \frac{1}{n}\sum_{i=1}^{n} h_i$.

Isolation by distance

The panel of 91 SNPs was used to calculate F_{ST} and isolation by distance, both of which relate to the mean heterozygosity. We imputed the ∼5% missing genotypes for each SNP by randomly assigning genotypes according to the population-wide allele frequencies at each marker. F_{ST} is defined as the average identity by descent among individuals within a subpopulation, F_S, relative to the total population, F_T : $F_{ST} = \frac{F_S - F_T}{1 - F_T}$ (Wright, 1931). These identities are estimated from SNP genotypes since we do not have the full pedigree. Two genes will have a different allelic state only if they are not identical by descent, and if they derive from different alleles in the ancestral population. Given overall ancestral allele frequencies $p + q = 1$, the expected heterozygosity ($\overline{H}$) of offspring from parents whose genes have a probability of identity by descent F is $\overline{H} = (1 - F)\overline{2pq}$, where $\overline{2pq}$ is an average over loci. Thus, there is a direct relation between F_{ST} and the mean heterozygosity: $F_{ST} = \frac{\overline{H_T} - \overline{H_S}}{\overline{H_T}}$. We use this relation to compute F_{ST} for this dataset (Jakobsson et al., 2013). (Note that here, H is the probability of non-identity in state, which depends on the SNP genotype. The subscripts S and T refer to the specified quantity within subpopulation and total population, respectively). Since we have a single continuous population, a subpopulation is defined as the set of pairs of individuals within a geographic separation of 20m and total population denotes all distinct pairs of individuals in the population. Note that 20m is an arbitrary choice of distance class used to define F_{ST}.

Isolation-by-distance – the decay of genetic similarity with geographic distance – can be observed by measuring pairwise relatedness between individuals. If individuals are separated by a distance r, then pairwise relatedness can be calculated as an extension of F_{ST} (which we refer to as pairwise F_{ij}, denoting the relatedness between individuals i and j) by setting F_S to be the probability of identity by descent and, correspondingly, $\overline{H_S}$ to be the probability of non-identity in state between genes which are at a distance r apart, thereby extending the idea of F_S from subpopulation to a set of pairs of individuals separated by any geographic distance class. $\overline{H_S}$ is calculated by finding the average pairwise heterozygosity between every pair of individuals which are within some interval $\{r, r + \delta r\}$ of distance apart. This formulation is used to estimate F_{ij} between every pair of individuals relative to the total population, as a function of their geographic separation. Pairs of individuals are binned into distance classes of 20m each (i.e individuals within 20m, 21-40m, and so on) and the average pairwise F_{ij} and the distance corresponding to each bin is calculated. This was done for every year from 2009 to 2019, and the average was calculated.

Variation in inbreeding

We calculated multilocus heterozygosity for each individual pooling across all years, denoted here by H, defined as the fraction of heterozygous loci in an individual. In this system "generations" cannot be clearly defined because of seed dormancy and perenniality. However, pooling data across years only reduced H by 0.08%.

We observed an excess of individuals with around half the mean heterozygosity (see Results). To check whether the pattern was consistent with rare selfing, we compared the likelihood of a single Gaussian to a mixture of two Gaussian distributions, one with the observed mean and variance and the other with half its mean and variance.

The variance in individual heterozygosity consists of two components. The first is due to the variance in whether an individual locus is heterozygous, and decreases in proportion to the number of SNP, n: it equals $(1-F)^2\overline{2pq(1-2pq)}/n$. The second is due to covariance in heterozygosity between loci, which is termed the identity disequilibrium (ID). For a given pedigree, unlinked genes flow independently. Thus, heterozygosity is independent across unlinked loci, and so this second component is proportional to the variance in inbreeding across individuals, var(F). The first component can be estimated from the allele frequencies, or simply by shuffling the data across individuals within loci, to eliminate ID. The excess variance is then proportional to the variance in F across individuals, and is measured by the statistic g_2:

$$g_2 = \frac{\sum_{i \neq j} cov[h_i, h_j]}{E[h]^2} = \frac{var(F)}{(1-E[F])^2}$$

(from eq. 1 in (Szulkin et al., 2010)). Here, $cov[h_i, h_j]$ is the ID between loci i and j, and the sum over all distinct i, j is the excess variance in H due to ID. Dividing by the square of the mean heterozygosity $E[h]^2$ eliminates dependence on allele frequency, such that g_2 estimates the variance in F across individuals.

To describe the variance of inbreeding across individuals, we first check if the variance in the distribution of individual heterozygosity is significantly greater than the average variance obtained from 100 replicates. This was done by shuffling heterozygous status randomly across individuals within loci, which would eliminate correlations between loci generated by ID. We then computed g_2 using the g2_snps function from the R package InbreedR (in R version 3.6.1 (R Core Team, 2014)), which implements a modified formula for large data sets to estimate g_2, and provides confidence intervals via bootstrapping to account for the finite number of individuals sampled (Stoffel et al., 2016). We decomposed ID into components due to linked and unlinked SNPs by comparing correlations of H for all individuals to those with low H, at several scales: across all pairs, within linkage groups, and between adjacent SNPs (Table D.1).

Additionally, g_2 can be used to estimate selfing rate within a population (David et al., 2007). Using the software SPAGeDi (Hardy and Vekemans, 2002), which implements the g_2-based selfing rate calculation described in (David et al., 2007), the selfing rate was estimated for the full population using the 91 SNP data.

Effects of pollen dispersal on heterozygosity

With isolation by distance, the distribution of heterozygosity is expected to depend on the distance between parents: heterozygosity of offspring from nearby parents will have a lower mean and higher variance compared to offspring from distant parents. To test this prediction, we simulated offspring using all field individuals as mothers and choosing fathers from a given distance away (details in Appendix D.3). Then we compared the distribution of H between the field data and offspring simulated from matings with three models of pollen dispersal: the nearest neighbor to the mother, a Gaussian distribution ($\sigma = 300m$), and a leptokurtic dispersal kernel sampled from 1463 empirical measurements of pollen dispersal, estimated as the distances between assigned parents (electronic supplementary material; D. Field, unpublished data). A CDF of the latter distribution (Fig. D.3) shows that 75% of the matings occur within 60m and has a kurtosis of 16.5 showing that the distribution is indeed leptokurtic. The genotype of the offspring was assigned using Mendelian inheritance, either without linkage between markers, or using the known linkage map (electronic supplementary material; courtesy of Yongbiao Xue, Beijing Institute of Genomics). Including linkage did not substantially change results, so we mainly show results for simulations without linkage. We compared distributions, means, and variance of H using Kolmogorov-Smirnov tests, t-tests, and F-tests, respectively. For the leptokurtic pollen dispersal simulation, we checked for an excess of low-heterozygosity individuals generated by mating between close relatives by asking whether a mixture of two Gaussian distributions is more likely than a single Gaussian distribution.

Heterozygosity in a simulated spatial pedigree

In order to compare the actual distribution of heterozygosity with that expected for a spatially structured population, we simulated a continuous two-dimensional population, conditioned on the known locations of the individuals and the empirically measured seed and pollen dispersal distances (electronic supplementary material; D. Field, unpublished data), using Mathematica 12.0 (Wolfram Research, Inc., 2019). Our simulation differs from commonly used models (e.g., island (Wright, 1931), stepping stone (Kimura and Weiss, 1964) and continuous Wright-Malécot model (Wright, 1943; Malecot, 1948)) in that we include heterogeneity in density by specifying actual locations to determine relationships in the pedigree. Thus, our simulation parameters should be seen as "effective" values,

analogous to the traditional N_e. Additionally, we also validated our simulation by comparing pairwise relatedness directly from the simulated pedigree and from replicate genotypes, and compared the realized and proposed dispersal kernels (Fig. D.6, D.7).

First, we simulated a population with uniform density (the continuous Wright-Malécot model) as a null model, to compare expected heterozygosity with and without heterogeneous spatial structure. We simulate a region of ∼1.1 x 1.8 km that was sampled consistently in the *A. majus* focal population (Fig. D.4). Locations were assigned by randomly sampling N points from a uniform distribution each generation, for 1000 generations. Genetic diversity is shaped over the coalescent timescale ($2N_e, \sim 170,000$, generations in *A. majus* (Tavares et al., 2018)), which is far longer than the 1000 generations that we simulate. However, we are concerned here with the local population structure that determines the variation in inbreeding amongst individuals within an area of a few km^2, which will equilibrate rapidly (Malecot, 1948). The spatial pedigree was generated by choosing parents for each individual according to a backwards dispersal distribution measured empirically. The seed and pollen dispersal distances are estimated respectively as the distance between offspring and nearest parent (assumed to be the mother) and between parents (electronic supplementary material; D. Field, unpublished data). For every offspring, the mother and father are chosen from randomly drawn distances from the seed and pollen dispersal distributions. To choose a parent from a distance r, 6 points are assigned randomly on a circle of radius r centred at the focal individual and the nearest individual to each of them are found. The closest individual to any of these points is then chosen as the parent. The accuracy of our algorithm is verified by comparing the specified and realised seed and pollen dispersal distributions for the simulated pedigrees (Fig.D.7, Table D.4). The same procedure is repeated for the father, taking the mother as the starting point. Since *A. majus* is self-incompatible, the mother and father are not allowed to be the same individual.

Once the spatial pedigree is generated, 10 replicate sets of genotypes are assigned by dropping genes down the pedigree, starting with equal expected frequencies of both alleles at each of 91 loci. In fact, one could start with any initial frequencies, since F_{ST}-like measures are independent of them. Population size was adjusted so that F_{ST} matched the empirical data for the simulated sampling area. This was done by first simulating the population with an initial population size (N) and then repeating the process with higher or lower N until the desired F_{ST} is attained.

Next, we simulated a population with realistic heterogeneous spatial structure by using the individual locations available for the years 2009 to 2019 in the *A. majus* focal population (Fig. D.5). There were fewer individuals from 2017-2018, so these were merged, giving distribution data for 10 time points. We randomly sample from the ten consecutive time points, and repeat for 100 cycles, thus iterating for 1000 generations. We sub-sample

from these locations to maintain a constant population size (N). If N is greater than the number of plants available in a given time point, say k, all k plants are first included and the remaining $N - k$ locations were re-sampled from the same time point, displaced at a random angle on a circle of radius 3m to avoid having plants in the same location. This naïve approach allows us to simulate a spatial structure that is realistic over at least small scales. We then generated a pedigree following the procedure used for the uniform population, again adjusting population size to match the empirically observed F_{ST}. Ten replicate sets of genotypes were run for each of five replicate pedigrees.

Patterns of isolation by distance, heterozygote deficit (F_{IS}) and identity disequilibrium were compared between the two simulation types and the field data (calculated from the simulated sub-area of the field site). As the fitted population sizes were large (see Results), obtaining direct estimates of identity by descent and thus F_{ST} from the pedigrees was not feasible. Instead, F_{ST} was obtained for a pedigree as the average of replicate genotype sets generated from that pedigree. F_{IS} was calculated from the observed and expected heterozygosity. Values of g_2 were calculated for each replicate from each pedigree using InbreedR (in R version 3.6.1 (R Core Team, 2014)).

4.3 Results

Isolation by distance

If we consider pairs of individuals within 20m of each other, the average F_{ST} over the eleven years is 0.0244; however, this is an average over a quantity that depends strongly on distance. The average pairwise F_{ij} was calculated each year for individuals separated by different distance classes and then averaged across years. Pairwise relatedness (pairwise F_{ij}) between individuals decreased rapidly with geographic distance, showing isolation by distance (Fig. 4.2A). The sharp decline in pairwise identity over short spatial scales corresponds precisely to a rapid increase in H with distance between parents (Fig. D.1), since heterozygosity is determined by the probability of identity by descent between the genes from each parent. Note that over large separations (>1Km), pairwise F_{ij} values are necessarily negative, because distant individuals are less closely related than the average for the whole population.

Variation in inbreeding

Excess variance in the distribution of individual heterozygosity (H) in the field data shows that there is variance in inbreeding in the population (Fig. 4.2B). Furthermore, there is an excess of individuals with around half the mean heterozygosity (i.e., with $H \sim 0.22$, rather than 0.44; Fig. 4.2B, blue, lower left). These might be due to a low rate of selfing,

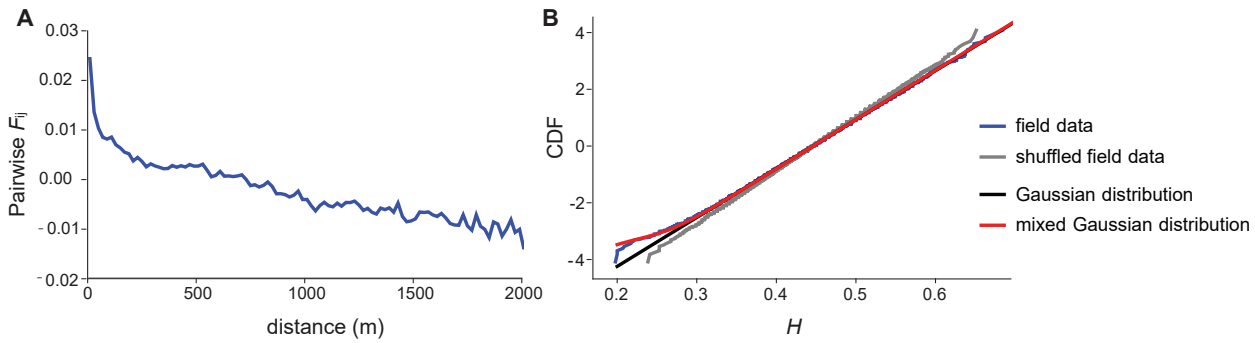


Figure 4.2: A: Pairwise relatedness (pairwise F_{ij}) between individuals decreases rapidly with geographic distance showing isolation- by- distance in the field data. B: Probit transform of the cumulative distribution function (CDF) of the distribution of individual heterozygosity (H). A Gaussian appears as a straight line on a probit scale, and the y-axis is the number of standard deviations of the standard normal distribution.

and using the g_2 estimator calculated with SPAGeDi, the selfing rate for the population is estimated to be 1.2%. Indeed, a mixture between two Gaussian with means ~ 0.22 and 0.44, and variances in the same ratio, fits significantly better than a single Gaussian (Fig. 4.2B, compare red and black to blue) with an increased likelihood of 11.3. However, we shall see in the next section that this excess is also consistent with matings between close relatives, without the need to invoke a breakdown in self-incompatibility.

To examine whether the observed distribution of heterozygosity is significantly different to a distribution taken from a population with zero identity disequilibrium (ID), we compared the field data with heterozygous values shuffled across individuals, which eliminates ID by removing correlations between loci. We found greater variance in heterozygosity in the observed compared to the randomly shuffled field data (Fig. 4.2B, gray). For both data sets, the mean heterozygosity (0.44602) necessarily remains the same, but the observed variance in the field data ($\text{var}(H) = 0.00336$) was significantly higher than the average variance in 100 shuffled replicates (mean $\text{var}(H) = 0.00282$, s.d. 0.000029). This excess variance between the observed and shuffled data implies that the mean standardized ID is $g_2 = 0.0029$ (95% CI: 0.0026-0.0033), representing a significant variance in inbreeding between individuals.

The overall ID, as measured by g_2, is due to correlations in heterozygosity between all pairs of loci, most of which are unlinked. We expect stronger correlations between linked loci, because relatives will share blocks of genome. We found that the mean covariance in heterozygosity between SNP on the same linkage group is substantially stronger than the overall mean (0.00265 vs. 0.00056). If we restrict attention to those individuals with $H < 0.3$, we find that the covariance in heterozygosity between SNP on the same linkage group is still higher (0.00649), as expected if close relatives share long blocks of genome IBD. This higher covariance in heterozygosity translates to higher mean g_2, which is seen within linkage groups compared to the overall value (Table D.1).

Effects of pollen dispersal on heterozygosity

The heterozygosity of simulated offspring depends on distance between their parents, with a rapid increase in mean H with distance (Fig. D.1). We compared the observed distribution of heterozygosity with three alternative scenarios for pollen dispersal. There was no significant difference between the mean and variance of heterozygosity between the field data and offspring simulated from the observed leptokurtic dispersal. However, the mean and variance of heterozygosity differed between the field data and simulated matings with either nearest neighbors, or with Gaussian dispersal (Fig. 4.3A, Tables D.2, D.3). While all three dispersal schemes differed in the distribution tail as assessed by Kolmogorov-Smirnov tests, Gaussian and nearest neighbour matings are very different from the field data compared to the leptokurtic distribution (Table D.3). These comparisons were made for a single replicate, but because each involves 22,353 individuals, there was little variation in the mean and variance between replicates.

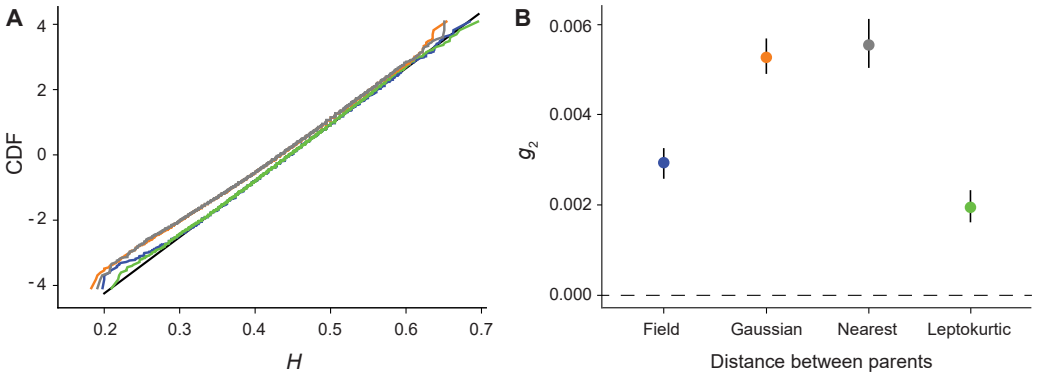


Figure 4.3: A: Probit transform of the CDF of multilocus heterozygosity, H, for the field data (blue) versus a single replicate of offspring simulated from Gaussian pollen dispersal (orange), nearest neighbor matings (gray), and leptokurtic pollen dispersal (green). A normal distribution (black) with the same mean and standard deviation as the field data is included for comparison B: Identity disequilibrium (g_2) for the same data as above indicating mean and 95% CI.

We next examined deviations in the left tail of the distribution, where an excess of low heterozygosity individuals might arise from selfing or from matings between close relatives. We focused on the leptokurtic dispersal curve, which was the distribution closest to the field data. We estimated the increase in likelihood between fitting a single versus mixed Gaussian distribution (see "Variation in inbreeding") for 100 replicate simulations. We found that the mixed Gaussian was a better fit than a single Gaussian, with an increase in log likelihood greater than 2 for 69 of 100 replicates. The estimated fraction of putatively "selfed" individuals was 0.00043, averaged over replicates, which is about half the estimate from the actual data, 0.00086. In comparison, only 4/100 replicates gave higher estimates than that observed (Fig. D.2). This suggests that the excess of individuals with low

heterozygosity can to a large extent be explained by matings between relatives under leptokurtic pollen dispersal. Nevertheless, there is a marginally significant excess of such individuals, with twice as many being seen as expected from our simulations. There is considerable variation in fit between replicates, simply because deviations in the tail involve few individuals.

The coefficient g_2 reflects excess variation due to identity disequilibrium, and showed similar patterns as the variance in H. Here, we found no significant difference between g_2 from field data and offspring from simulated matings with leptokurtic pollen dispersal. However, g_2 from Gaussian and neighbor matings were 80% higher than g_2 from field data and leptokurtic matings. This nominally represents a significant difference given that the 95% confidence intervals between these groups do not overlap (Fig. 4.3B). However, as we discuss below, these confidence intervals only include sampling error, and not the additional variance due to random evolutionary realizations.

Heterozygosity in a simulated spatial pedigree

In the previous section, we simulated offspring across one generation. To examine whether the observed heterozygosity is consistent with a spatially structured model, we simulated pedigrees over 1000 generations with uniform and heterogeneous density, conditioned on the locations of individuals observed over ten years, repeated over 100 cycles for the latter case. The realized seed and pollen dispersal matched the empirical seed and pollen dispersal distribution for both density types (Fig. D.7, Table D.4). We required $N = 15500$ individuals for the heterogeneous density model and 40000 individuals for the uniform density, in order to match the observed $F_{ST} \sim 0.022$ calculated over a 20m scale from the simulated sub-area of the field site (Table D.5). Up to distances of 1km, the decline in pairwise identity with distance matched between the field data and the five replicate pedigrees simulated with heterogeneous density (Fig. 4A, Fig. D.8A). High variation among replicates suggests that many more SNPs would be needed to match the pattern from the pedigree (Fig. D.8B); moreover, linkage would increase this variance to some extent. We also compared the pattern of isolation by distance from the field data to that from the pedigrees generated for both the heterogeneous and uniform density scenarios (Fig 4.4B, Fig. D.9); the heterogeneous density is a much better fit than the uniform density (Table D.5).

Identity disequilibrium (g_2) estimates from the genotypes from pedigrees simulated with heterogeneous density showed substantial variation between the five simulated pedigrees, and between the ten draws of 91 SNPs from each pedigree (Fig. 4.5). The average g_2 estimated from the five pedigrees (each with 10 replicates) is 0.00264, which is consistent with the observed mean annual g_2 from the field of 0.00262. On the other hand, when assuming a uniform density, the average g_2 of 0.00171 is significantly lower than the field

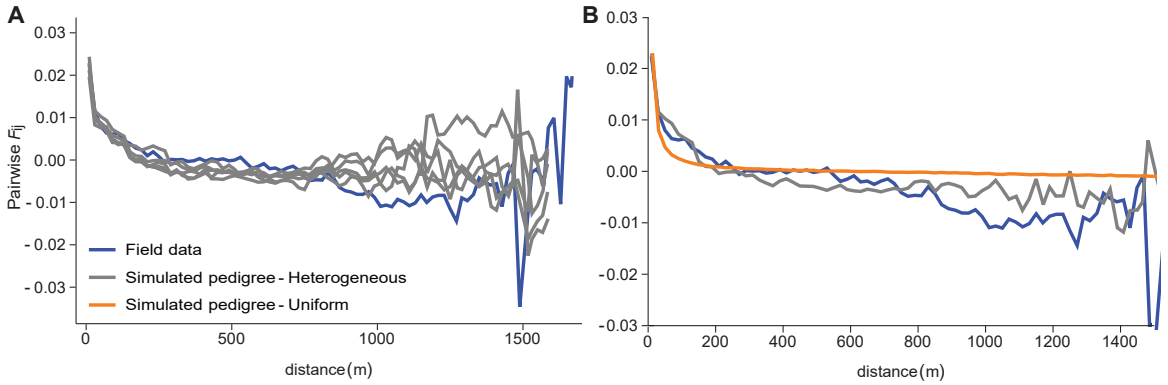


Figure 4.4: A:Isolation by distance compared between the field data (blue) and five replicate simulated pedigrees (gray) based on a heterogeneous population density. B: Isolation by distance from the field data (blue) compared between the simulated pedigree with a heterogeneous (gray) and uniform (orange) population density.

data. Note that the confidence limits for the field data, generated by InbreedR, only include error due to sampling a limited number of individuals. These errors do not account for sampling a limited number of SNPs, or the random variation between evolutionary realizations (see Discussion).

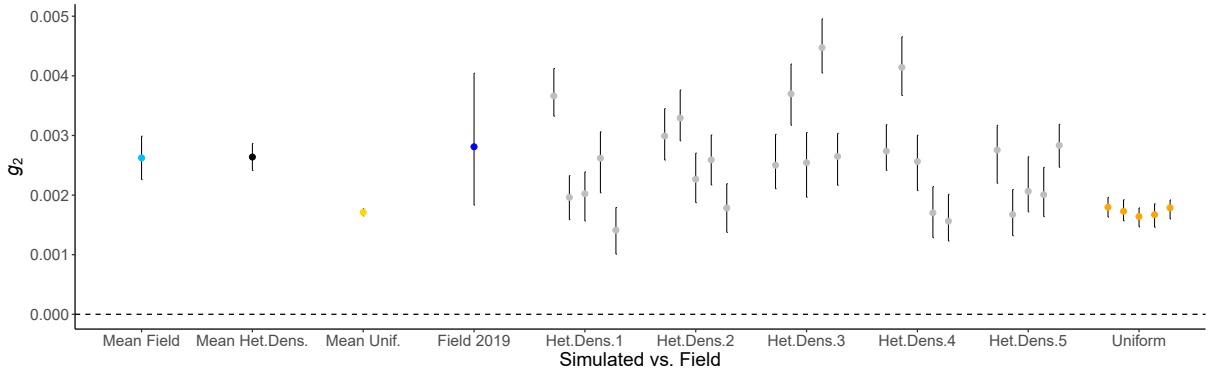


Figure 4.5: Identity disequilibrium (g_2) calculated from field data versus simulated pedigrees. Five of ten replicates per pedigree are shown (gray: heterogeneous density, with five simulated pedigrees; orange: uniform density, with one simulated pedigree). Mean from the field (light blue) is across 2009-2019, while mean from the heterogeneous (black) and uniform (yellow) simulations is across all replicates. The final year of field data (dark blue) is comparable to g_2 calculated from the final year of pedigree replicates (gray and orange).

4.4 Discussion

An enduring problem in evolutionary biology is understanding how demographic processes, such as heterogeneous density and dispersal, interact with spatial structure to determine the distribution of heterozygosity within populations. In this study of a long-term dataset, including more than 20,000 plants sampled over 11 years, we combine field data and simulations to address questions central to understanding how demography can influence

patterns of heterozygosity. Namely, can we predict the distribution of heterozygosity for an outcrossing species from key demographic parameters? To address this question, we first confirmed that there was significant correlation in heterozygosity between markers (g_2, a measure of identity disequilibrium), which implies variation in inbreeding. By simulating offspring from matings between geo-referenced, genotyped individuals, we show that the mean heterozygosity increases, and the variance of heterozygosity decreases, with increasing distance between parents; strikingly, these changes occur over very short scales (∼10m, Fig. D.1). We found that the observed distribution of heterozygosity is consistent with the known leptokurtic distribution of pollen dispersal. We also simulate the population over 1000 generations using the actual seed and pollen dispersal kernels, and the observed heterogeneous density. We found that this model matches the observed identity disequilibrium, whereas a model with uniform density substantially underestimates the observed patterns. Thus, we explain the distribution of heterozygosity (mean, variance and tails) using known features of the population. Moreover, our results also highlight the limitations of making theoretical predictions from simulations that only assume simple demographies. Taken together, our findings highlight the potential for using the observed demography to explain the distribution of genetic diversity, and specifically the variance in inbreeding in spatially continuous populations.

Variation in heterozygosity within populations provides the potential for selection to reduce the frequency of less fit, inbred individuals. The association between inbreeding and fitness is often tested through heterozygosity-fitness correlations (HFC), which quantify inbreeding depression in natural populations by correlating measures of fitness with heterozygosity [6]. Many studies that test for HFCs find that the excess variation in heterozygosity, g_2, which arises from identity disequilibrium, is low and rarely significant (Miller and Coltman, 2014). In our study, we estimate a significant g_2 of 0.0029 (95% CI: 0.0026-0.0033). Although low, this estimate is of the same order as most of the g_2 values found across 105 vertebrate populations in a meta-analysis of 50 HFC studies (average of 0.007) (Miller and Coltman, 2014), and on the same order as ∼60% of the local populations surveyed in a long-lived tree (Rodríguez-Quilón et al., 2015). Our estimate of significant variation in heterozygosity provides the opportunity to examine potential drivers of this variance and examine how density, spatial structure and dispersal contribute to a non-uniform distribution of heterozygosity.

In our study, beyond simply estimating identity disequilibrium, we use two types of simulation to explore how demography shapes variation in inbreeding. The first simulation shows how the spatial pattern of pollen dispersal affects the distribution of heterozygosity. Simulated matings with the empirically measured leptokurtic pollen dispersal curve were consistent with the actual g_2, compared to matings with nearest neighbours or a Gaussian pollen dispersal. This result is somewhat surprising because we did not include the

complexities of the mating system of *A. majus*. *Antirrhinum majus* has a gametophytic self-incompatibility system (GSI (McCubbin et al., 1992)), whereby the pollen detoxifies secretions from the style unless the pollen and style genotypes share alleles at the S-locus (Fujii et al., 2016). This system not only prevents selfing, but also reduces mating among relatives (i.e., biparental inbreeding) because related plants are more likely to share S-alleles (Charlesworth and Charlesworth, 1987; Cartwright, 2009). Thus, we might expect that our simulated matings would have lower mean heterozygosity than the empirical measurement; yet we found no evidence for such an effect. Indeed, we found that the excess of individuals with low heterozygosity, around half the mean, can be explained largely by a small amount of bi-parental inbreeding with leptokurtic pollen dispersal (Fig. 4.3, Fig. D.2). However, we have little statistical power to distinguish this from rare selfing, which can occur in self-incompatible species. In fact, using the g_2 estimator of selfing rate from eq. 9 in (David et al., 2007), our significant g_2 value would imply a selfing rate of 1.2% for this population. However, as shown by (Hardy, 2016), this estimate could be within the bounds of the upward bias of the estimator if strong biparental inbreeding is present, hence, this does not necessarily imply a breakdown of self-incompatibility. We believe that our method, fitting a model of two Gaussians, is a more robust way to estimate selfing than using g_2, since it focuses on the low-H individuals rather than the whole variance. However, it is still challenging to distinguish selfing from close inbreeding.

Our second simulation approach asked whether heterogeneous density promotes variation in inbreeding, given strong fine-scale population structure indicated by a rapid decay in pairwise F_ij (over a few metres, Fig. 4.2A). We only provide a proof-of-principle, by asking whether a plausible model of spatial structure can explain the observed heterozygosity. We do not include all features of the actual population – in particular, we extrapolate by repeatedly sampling ten years of spatial distributions; we ignore linkage; we simplify the self-incompatibility system; and we assume an annual life cycle (no perenniality or seed bank). Indeed, simulated pedigrees with uniformly distributed plants gave less identity disequilibrium than we observed. In contrast, simulated pedigrees conditioned on the actual, heterogeneous density of plants were consistent with identity disequilibrium measured in the field. This indicates that patchiness combined with leptokurtic dispersal shapes the distribution of heterozygosity. Simulations with heterogeneous density also better capture empirical isolation-by-distance patterns than those with a uniform density (Fig 4.4B, Fig. D.9). However, the effective population size of 15,500 individuals in the heterogeneous-density simulations is an order of magnitude larger than the average number of plants observed in a year ($\sim$ 2500). We believe that most plants are sampled each year, so that this discrepancy is more likely to be due to a seed bank, which is expected to substantially increase the effective population size (Heinrich et al., 2018). Nevertheless, despite simplifications such as non-overlapping generations, no seed bank, and a simple SI system, the heterogeneous-density simulation accurately captures patterns

of identity disequilibrium and isolation-by-distance.

Our estimation of identity disequilibrium illustrates a general problem with statistical comparisons in evolutionary biology. There are three sources of error in estimating g_2: firstly, error generated from sampling a limited number of individuals, secondly, from sampling a limited number of SNPs, and thirdly from random variation between evolutionary realizations or trajectories. In our study, the first source (a limited number of individuals) is shown by the confidence intervals in Fig. 4.5, obtained by bootstrapping across individuals (Stoffel et al., 2016). The second source of error (a limited number of SNPs) is shown by the substantial variation in g2 of the ten replicates of each of five pedigrees. Here, variation is generated by random meiosis amongst unlinked markers on a fixed pedigree. This variation could be reduced by increasing the number of SNPs, but the effective number of segregating sites that can be included in the analysis is fundamentally limited by the length of the genetic map. Finally, there is additional variation between pedigrees, due to the random assignment of parents in the simulations, which generates a random pedigree. The wide variation in estimates of g_2 due to random meiosis, and to the random generation of the pedigree (Fig. 4.5) is an important reminder that estimates of parameters are typically limited by the randomness of evolution. The stochasticity of evolution can potentially generate error variance far higher than that due to the limited number of individuals or SNPs sampled.

In addition to analyzing the effect of population structure on the distribution of heterozygosity, our study highlights the potential of utilizing multiple statistics to estimate population structure. We have shown that the variance of heterozygosity due to identity disequilibrium can distinguish alternative dispersal and density distributions, which implies that in combination with pairwise F_{ij} as a function of distance, g_2 can help estimate the demography. Genetic data contain far more information than is described by F_{ST} and g_2; for example, the mean squared disequilibrium can be used to estimate effective population size (Hill, 1981; Vitalis and Couvet, 2001), and this extends naturally to the covariance of pairwise linkage disequilibrium as a function of distance. We could simply use a set of such statistics to inform demographic inference via ABC (Beaumont, 2010). However, our preference would be to first develop a theoretical understanding of how realistic demographies influence statistical measures of spatial covariance in allele frequency, identity disequilibria, and linkage disequilibria.

The distribution of heterozygosity has often been measured to estimate inbreeding depression and examine correlation with fitness. Yet, this type of data has rarely been used to investigate population structure per se and as a complement to the more widely used pairwise identity, F_{ST}. By bringing together local inbreeding and isolation-by-distance, our study provides a novel assessment of how dispersal and population density can explain both pairwise identity and the distribution of heterozygosity in spatially continuous popu-

lations. However, we have only begun to investigate how the distribution of heterozygosity can be shaped by population structure and demographic parameters. Our future work will focus on understanding how other features such as a seed bank influence genetic diversity, with the ultimate goal of deriving information about demographic history from the distribution of heterozygosity in populations that have fewer measured parameters. New models that include these complexities, as well as ecological, mating system and life history factors are required to extend our understanding of the drivers of population structure in natural populations.

Data availability

All data and code used to generate simulated data and carry out analysis is available at: https://doi.org/10.15479/AT:ISTA:11321. Data includes processed field data for 11 years of *Antirrhinum majus* sample collection, including SNP values, GPS locations and trait measurement values for each plant. Also included are dispersal data and a linkage map of 91 SNPs.

Acknowledgments

We thank the many volunteers and friends who have contributed to data collection in the field site over the years, in particular those who have managed field seasons: Barbora Trubenova, Maria Clara Melo, Tom Ellis, Eva Cereghetti, Lenka Matejovicova, Beatriz Pablo Carmona. Frederic Ferrer and Eva Salmerón Mateu have been immensely helpful with logistics at our informal field station, El Serrat de Planoles. We thank Sean Stankowski for technical help in producing Figure 4.1. This research was also supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

Funding

Part of this work was funded by Marie Curie COFUND Doctoral Fellowship and Austrian Science Fund FWF (grant P32166).

Conflict of interest

The authors declare that there is no conflict of interest.

References

Andalo, C., Cruzan, M. B., Cazettes, C., Pujol, B., Burrus, M., and Thébaud, C. (2010). Post-pollination barriers do not explain the persistence of two distinct *Antirrhinum* subspecies with parapatric distribution. *Plant Syst. Evol.*, 286(3):223–234.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol*, 41:379–406.

Bradburd, G. S. and Ralph, P. L. (2019). Spatial population genetics: It's about time. *Annu. Rev. Ecol. Evol*, 50:427–449.

Cartwright, R. A. (2009). Antagonism between local dispersal and self-incompatibility systems in a continuous plant population. *Mol. Ecol.*, 18(11):2327–2336.

Charlesworth, D. and Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. *Annual review of ecology and systematics*, pages 237–268.

David, P., Pujol, B., Viard, F., Castella, V., and Goudet, J. (2007). Reliable selfing rate estimates from imperfect population genetic data. *Mol. Ecol.*, 16(12):2474–2487.

Fujii, S., Kubo, K. I., and Takayama, S. (2016). Non-self- and self-recognition models in plant self-incompatibility. *Nat. Plants,*, 2(9):1–9.

Hardy, O. J. (2016). Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Mol. Ecol. Resour.*, 16(1):103–117.

Hardy, O. J. and Vekemans, X. (2002). spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes*, 2(4):618–620.

Heinrich, L., Müller, J., Tellier, A., and Živković, D. (2018). Effects of population- and seed bank size fluctuations on neutral evolution and efficacy of natural selection. *Theor. Popul*, 123:45–69.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium1. *Genet. Res. (Camb).*, 38(3):209–216.

Jakobsson, M., Edge, M. D., and Rosenberg, N. A. (2013). The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics*, 193(2):515.

Kimura, M. and Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561.

Loveless, M. D. and Hamrick, J. L. (1984). Ecological determinants of genetic structure in plant populations. *Annu. Rev. Ecol. Syst*, 15:65–95.

Lynch, M. and Walsh, B. (1996). *Genetics and Analysis of Quantitative Traits*. Sinauer Sunderland, MA.

Malecot, G. (1948). *The Mathematics of Heredity (English translation)*. San Francisco: WF Freeman.

McCubbin, A., Carpenter, R., Coen, E., and Dickinson, H. (1992). Self-incompatibility in antirrhinum. In *Angiosperm Pollen and Ovules*, pages 104–109. Springer.

Miller, J. M. and Coltman, D. W. (2014). Assessment of identity disequilibrium and its relation to empirical heterozygosity fitness correlations: a meta-analysis. *Mol. Ecol.*, 23(8):1899–1909.

Milligan, B. G., Archer, F. I., Ferchaud, A. L., Hand, B. K., Kierepka, E. M., and Waples, R. S. (2018). Disentangling genetic structure for genetic monitoring of complex populations. *Evol. Appl.*, 11(7):1149–1161.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ringbauer, H., Kolesnikov, A., Field, D. L., and Barton, N. H. (2018). Estimating barriers to gene flow from distorted isolation-by-distance patterns. *Genetics*, 208(3):1231–1245.

Rodríguez-Quilón, I. et al. (2015). Local effects drive heterozygosity-fitness correlations in an outcrossing long-lived tree. *Proceedings. Biol*, 282:2015–2230.

Sin, S. Y. W., Hoover, B. A., Nevitt, G. A., and Edwards, S. V. (2021). Demographic history, not mating system, explains signatures of inbreeding and inbreeding depression in a large outbred population. *Am. Nat*, 197(6):658–676.

Stoffel, M. A. et al. (2016). inbreedr: an R package for the analysis of inbreeding based on genetic markers. *Methods Ecol. Evol.*, 7(11):1331–1339.

Szulkin, M., Bierne, N., and David, P. (2010). Heterozygosity-fitness correlations: A time for reappraisal. *Evolution*, 64(5):1202–1217.

Tavares, H. et al. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proc. Natl. Acad. Sci. U. S. A.*, 115(43):11006–11011.

Vekemans, X. and Hardy, O. J. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol. Ecol.*, 13(4):921–935.

Vitalis, R. and Couvet, D. (2001). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, 157(2):911–925.

Whibley, A. C. et al. (2006). Evolutionary paths underlying flower color variation in *Antirrhinum*. *Science*, 313(5789):963–966.

Winn, A. A. et al. (2011). Analysis of inbreeding depression in mixed-mating plants provides evidence for selective interference and stable mixed mating. *Evolution*, 65(12):3339–3359.

Wolfram Research, Inc. (2019). *Mathematica*. Wolfram Research, Inc. Champaign, Illinois.

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

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114.

Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics*, 31(1):39–59.

Chapter 5

Genetic analysis of flower colour clines in *Antirrhinum majus*

Abstract

Understanding the processes that shape populations and species is a major goal of evolutionary biology. In this respect, hybrid zones have provided excellent natural laboratories for studying local adaptation and speciation in the wild. Most inferences from hybrid zones are based on the properties of geographic clines, which are shaped by the interaction of selection and gene flow. However, hybrid zone theory makes a number of simplifying assumptions that may lead to erroneous conclusions. In this study, we analyzed a flower colour hybrid zone in the common snapdragon (*Antirrhinum majus*) in Val di Ribes, Spain using a long-term dataset of 22494 individuals genotyped at 6 colour loci. We find coincident stepped clines at all loci suggesting that flower colour is under selection. Linkage disequilibrium (LD) between unlinked loci was positive throughout the transect and stronger outside the cline centre, unexpected from traditional cline theory, which assumes diffusive gene flow. In order to explain this, we inferred the full seed and pollen dispersal distribution, which was leptokurtic, with a small fraction of long-distance migrants. A novel simulation framework with the inferred dispersal and multilocus selection showed that long-range dispersal is the primary cause of the stepped cline shape, with a weaker effect from LD. This approach enabled us to obtain realistic estimates of selection and to separate it into components of direct selection due to its effect on fitness and indirect selection due to the effect of LD with other colour loci. This study also underlines how disentangling the causes of stepped shape bears important conclusions for speciation. Overall, our results shed new light into the processes that shape this hybrid zone, and highlight the need to account for the effects of realistic dispersal in cline theory.

Keywords: hybrid zone, long-term study, stepped clines, long-range dispersal, linkage disequilibrium, flower colour

5.1 Introduction

Quantifying the evolutionary forces that shape populations and species is a major goal of evolutionary biology. Considerable effort has focused on understanding how broadly distributed species can adapt to diverse conditions across their geographic range (Endler, 1977; Slatkin, 1987; Pinho and Hey, 2010). Although we know that selection can in principle overcome the homogenizing effects of dispersal to maintain local adaptation (Endler, 1977), measuring the strength of these opposing forces is very difficult. Theory provides some intuition about the strength of selection needed to balance dispersal, but makes simplifying assumptions that are hard to relate to nature.

Some of the most detailed inferences of selection and dispersal to date have come from empirical studies of hybrid zones (Szymura and Barton, 1986; Mallet and Barton, 1989; Phillips et al., 2004; Kawakami et al., 2009). Hybrid zones are narrow areas where genetically differentiated populations come into contact with one another, producing hybrid offspring (Barton and Hewitt, 1985; Hewitt, 1988). They are found in a wide range of organisms and environments, and are considered natural laboratories for studying divergence and speciation in the wild (Hewitt, 1988; Harrison, 1993). Most well-studied hybrid zones are thought to reflect the long-term balance between selection, which keeps them narrow, and dispersal, which acts to widen the zone (Endler, 1977). They have been extensively studied theoretically, providing a framework for inferring evolutionary processes from empirical datasets (Haldane, 1948; Slatkin, 1973; Endler, 1977; Barton, 1979a).

Most inferences from hybrid zones are based on the properties of geographic clines (i.e., gradients in trait means or allele frequencies). When a single locus or trait is selected, and genes diffuse through the habitat, the corresponding cline has a sigmoid shape (Fig. 5.1A), with width, w, proportional to the ratio between dispersal and selection. Thus, narrower clines imply stronger selection (Haldane, 1948; Bazykin, 1969; Slatkin, 1973; Nagylaki, 1976). However, when clines at multiple selected loci coincide in a hybrid zone, they may have a 'stepped' shape in which a steep sigmoid gradient is flanked by shallow tails of introgression (Fig 5.1B, Fig. C.1 Barton (1983); Szymura and Barton (1986)). When multiple loci influence the trait under selection, each locus experiences an increase in effective selection due to positive associations (i.e., linkage disequilibrium; LD) between them at the cline centre (Barton, 1983; Kruuk et al., 1999). Specifically, the total selection experienced by a given locus is determined not only by its own direct effect on fitness, but also by indirect selection, i.e how allele frequency at a locus is influenced

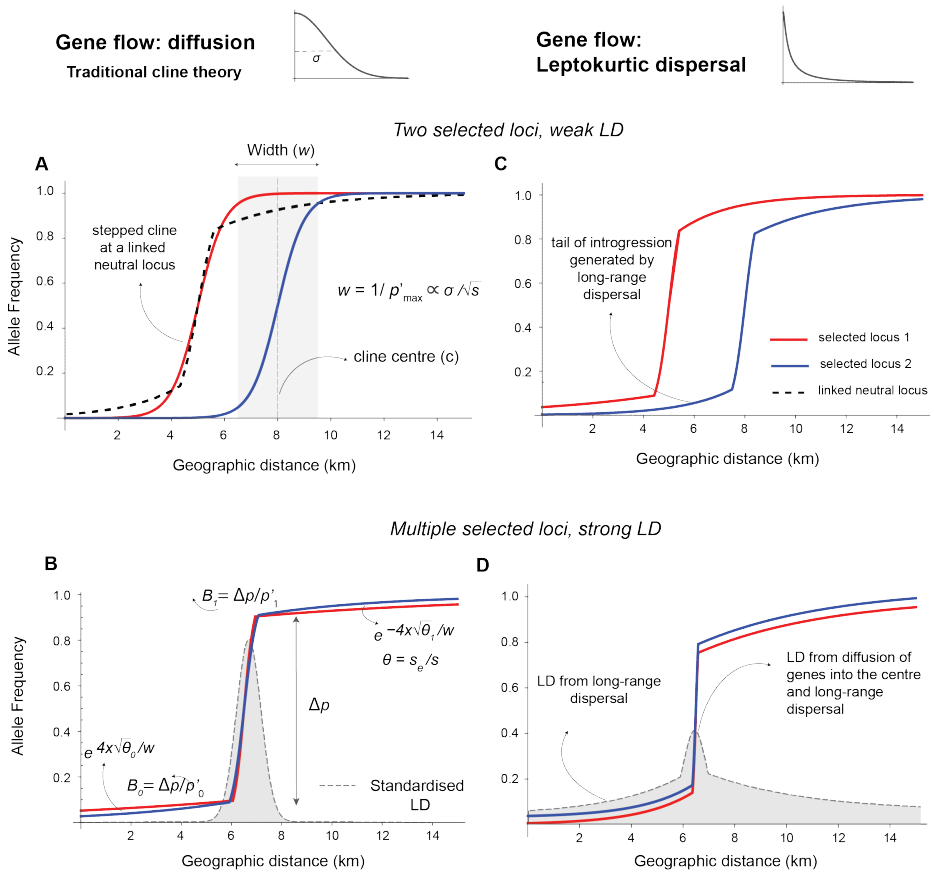


Figure 5.1: **Illustration of cline theory under diffusive (left column) vs leptokurtic dispersal (right column):** The red and blue curves in panels A-D show the cline shapes for 2 selected loci. Panels A and B show expected cline shapes when LD is (A) weak and (B) strong. When LD is weak, clines have a sigmoid shape. The cline width (w) is defined as the inverse of the maximum gradient $w = 1/p'_{max}$ and is proportional to $\sigma/\sqrt{s}$, where σ is the standard deviation of the parent-offspring distance and s is the effective selection at the center (c). When LD among selected loci is strong, it increases the effective selection at the center, distorting the cline shape. These are referred to as 'stepped' clines because they have a steep central step flanked by shallower tails of introgression (also see Fig S1). The tails take the form $\exp\left\{4x\sqrt{\theta}/w\right\}$, where θ is the ratio between effective selection at the centre (s) and selection against introgressing alleles (s_e) away from the centre, and may differ to the left (θ_0) and right (θ_1). Under this scenario, the step reflects a genetic barrier to gene flow (B), described as $B = \Delta p/p'$, where Δp is difference in allele frequency across the central step and p' is gradient of allele frequency at the edge of the step and the tail. (Panel C) When dispersal is leptokurtic, cline shapes may be stepped even if LD is weak. This is because long distance migration transports foreign alleles to the opposite tail in a single dispersal event, in contrast to diffusion where alleles must cross the barrier via many small steps. In this scenario, the step is caused only by long-range dispersal, not LD. (Panel D) Both leptokurtic dispersal and LD contribute to the step. This scenario is distinguished from panel B, by the presence of correlations between loci (i.e., standardized LD) in the tails of the cline which is generated by long-distance migration (gray distributions in B and D). Under diffusive gene flow, these correlations are only expected at the cline centre (as in panel B). In all four scenarios neutral loci linked to selected loci are all expected to show stepped clines (dashed black curve in panel A), indicating a barrier to the flow of neutral alleles.

by its associations with other selected loci (even when they are unlinked). This effect is strongest in the center of the cline, where LD is strongest, generating a barrier to the flow of alleles from one side of the hybrid zone to the other. However, as alleles escape their associations by recombining onto the alternative genetic background, the strength of indirect selection decreases and loci behave independently (due to weak LD), leading to the shallower tails of introgression that lead away from the centre of the zone (Barton and Hewitt, 1985; Barton and Gale, 1993). The stepped cline model has been applied to a variety of empirical systems, enabling researchers to quantify a range of biological parameters, including the selection acting on each locus (s), the strength of the barrier to gene flow in units of geographic distance (B), and rates of introgression into the tails (θ)(e.g., (Szymura and Barton, 1986; Kawakami et al., 2009; Raufaste et al., 2005; Porter et al., 1997). However, other factors can cause clines to have a stepped shape, so observing them in nature need not always indicate the presence of a genetic barrier to gene flow.

The pattern of dispersal may also influence cline shape. A key assumption of 'traditional' cline theory (Fig 5.1A, B) is that the distribution of dispersal distance has finite moments, and can thus be approximated as diffusion, provided that selection is weak (Nagylaki, 1976; Slatkin, 1973). This enables dispersal to be described by a single parameter σ – the root mean square of parent-offspring distance along some axis, which can be estimated from mark-recapture studies (Mallet and Barton, 1989) or from LD at the cline center (Barton and Gale, 1993). In reality, many species show leptokurtic dispersal, meaning that dispersal mostly occurs over short distances, but occasionally includes long-range movements (Nathan et al., 2008; Cayuela et al., 2018). These long-distance events can be hard to detect, yet can have significant effects on the distribution of patterns of genetic variation (Fig 5.1C,D).

In this study, we infer how selection and dispersal have shaped a hybrid zone between two subspecies of the snapdragon *Antirrhinum majus*. *A. majus* has been a model for understanding trait variation dating back to crossing experiments by Mendel and Darwin (Hudson et al., 2008). The natural range of *A. majus* is mainly centered in France and Spain, though its popularity in horticulture has seen populations established throughout Europe and across the globe (Hudson et al., 2008). We focus on a natural hybrid zone between two subspecies, *A.m. pseudomajus* and *A.m.striatum* (Whibley et al., 2006). They have largely distinct geographic ranges, but occupy a similar range of local habitats and are pollinated by the same bee species (Tavares et al., 2018). The only known trait that distinguishes them is their different flower colour patterns, which are thought to be alternate solutions to attracting bees and signposting bee entry into the flower (Whibley et al., 2006; Tavares et al., 2018): *A.m.pseudomajus* has magenta flowers coloured by anthocyanin pigment with a small patch of yellow aurone pigment below the bee entry point (Fig. 5.2a). *A.m.striatum* has primarily yellow flowers due to widespread presence

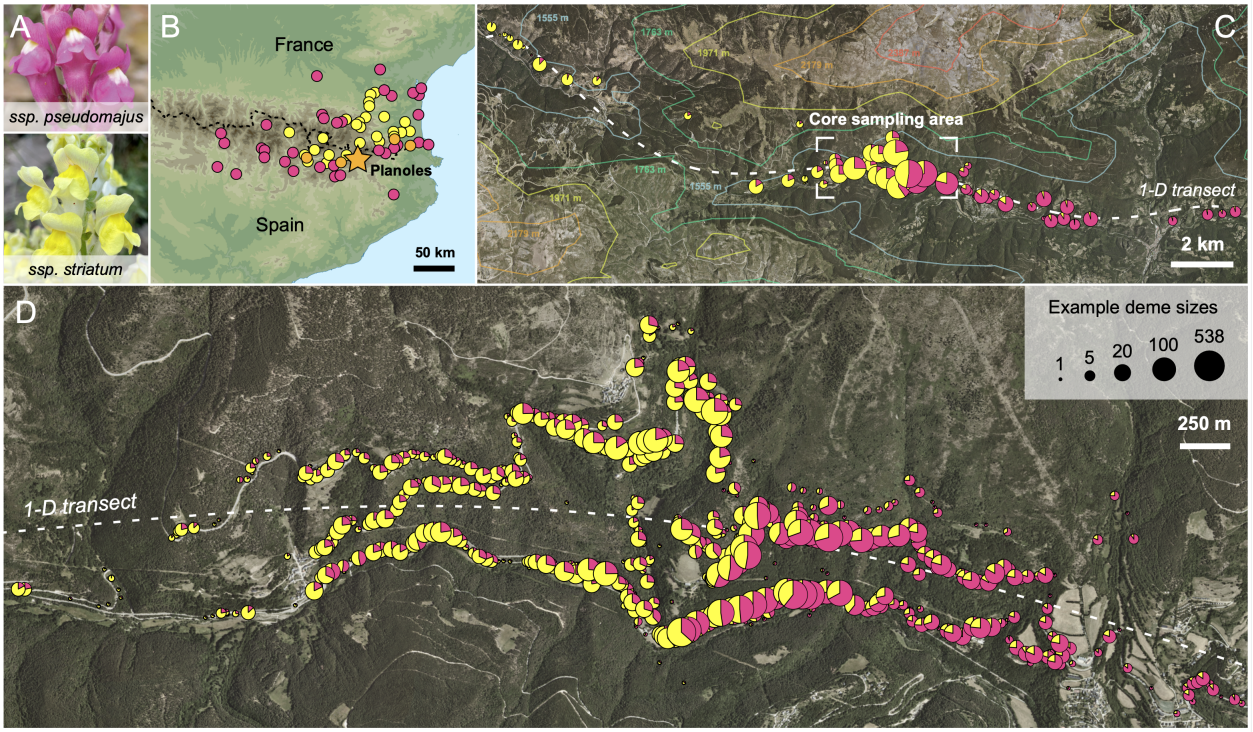


Figure 5.2: **The snapdragon hybrid zone in the Spanish Pyrenees** (a) Typical phenotypes of the subspecies. (b) Partial ranges of *A.m.pseudomajus* (magenta) *A.m.striatum* (yellow) and known hybrid zones (orange) based on sampling in Whibley et al. (2006). The magenta flowered *A.m.pseudomajus* is found in northern Spain and south-western France, its range encircling the yellow-flowered *A.m.striatum*. The dashed black line marks the French-Spanish border. The star indicates our study area. (c) The hybrid zone in the Val di Ribes. For visualisation, the 22,494 individuals have been clustered into 500m demes; pie charts show the mean hybrid index within each deme, defined as the proportion of *A.m. pseudomajus* alleles at five unlinked colour loci (excluding Eluta which is tightly linked to Rosea). The dashed white line shows the one-dimensional transect line and the contour lines show altitude. (d) The core sampling area. Individuals are clustered into 25 m demes, which is the scale used for all analyses. In c and d, pie charts are scaled according to the log of the number of samples that they contain.

of aurone pigment, with restricted veins of magenta anthocyanin pigment above the bee entry point (Fig. 5.2a). This difference in colour pattern is determined by a handful of loci that control the production of anthocyanin and aurone pigment in floral tissue (Table 5.1). Three loci—Rosea (Tavares et al., 2018), Eluta (Tavares et al., 2018) and Sulfurea (Bradley et al., 2017)—have major effects, while three others—Rubia (Field et al. in prep), Cremosa (Bradley et al in prep), Flavia (Richardson et al in prep)—have more subtle effects on colour pattern and/or intensity.

During the last ice age, *A. majus* was restricted to areas of lower elevation, but subsequently expanded into the Pyrenees as the climate warmed. As a result, *A.m.pseudomajus* and *A.m.striatum* have come into contact, forming narrow hybrid zones in multiple valleys separated by mountains that rise above the altitude tolerance of *A. majus* (Fig. 5.2B).

Segregation and recombination of colour alleles in hybrids has generated a wide range of colour phenotypes. These include individuals that are phenotypically intermediate between *A.m.pseudomajus* and *A.m.striatum*, as well as unusual phenotypes that fall outside the range seen in the genus Antirrhinum (Tastard et al., 2008). Selection by pollinators is thought to maintain hybrid zones in two ways. First, experiments have shown that some hybrid colour patterns are less likely to be visited because they do not fit the visual preferences of bee pollinators (Tastard et al., 2008). *A.m.pseudomajus* and *A.m.striatum* colour patterns are also thought to be subject to positive frequency-dependent selection, each having higher fitness on the side of the hybrid zone where they are more common (Tastard et al., 2014).

Several studies at one hybrid zone in the Val di Ribes suggested that selection acts on flower colour. First, past work has shown that the transition from *A.m.pseudomajus* and *A.m.striatum* is abrupt, and is characterised by a steep cline in flower colour only a few km wide (Whibley et al., 2006; Tavares et al., 2018). Studies of genome-wide sequence variation, including preliminary cline analyses, have revealed strong allele frequency differentiation and steep clines around colour loci (Whibley et al. (2006), Field et al. in prep). In contrast, the remaining genomic background shows low differentiation potentially due to the long term effects of dispersal between the two subspecies (Tavares et al. (2018); Ringbauer et al. (2018) Field et al. in prep). Despite this past work, we still lack a robust, quantitative understanding of how selection and dispersal shape this hybrid zone.

In this study, we conduct a genetic analysis of the hybrid zone in the Val di Ribes, based on 22,494 individuals sampled and genotyped over 10 years. Our main aim was to interpret cline shapes and LD among SNPs that are tightly linked to 6 known color loci in order to quantitatively understand how selection and dispersal shape the zone. We do not attempt to distinguish alternate forms of selection or analyze phenotypes in detail. A novel aspect of our approach, made possible by our intensive sampling, is that we are able to infer the full distribution of pollen and seed dispersal, which is highly leptokurtic. These long-range dispersal events, not accounted for in traditional cline theory, substantially distort the cline shape. By incorporating the empirical dispersal into a multilocus simulation, we explain the causes of stepped clines and observed LD and decompose the total effective selection into direct and indirect components. Our results expand our understanding of the processes that shape this hybrid zone, and highlight new approaches to infer selection and dispersal in scenarios when 'traditional' cline analysis fails.

5.2 Results

5.2.1 Sampling and description of datasets

We conducted fieldwork each year from 2009 to 2019 during the peak of the flowering season (late May to early August), attempting to sample every flowering individual within our core study area (See Methods; Fig. 5.2C and D). Although the area is large, snapdragons are restricted mainly to south-facing rocky cliffs and disturbed areas that are accessible from two parallel roads. Repeated surveys of adjacent forested slopes, which are heavily shaded and dominated by conifer species of Pinus or Abies, have found very few individuals, suggesting that we survey most of the suitable habitats in the area. Other suitable habitat includes disturbed deforested areas further from the roads, small side roads, railway embankments, and hiking trails. In some years, we also sampled beyond the core area to improve sampling in the left and right tails of the cline (Fig. 5.2C). After accounting for individuals that were sampled multiple times (identified by genotype) because they lived more than one year, our dataset includes 22,494 unique individuals used in subsequent analysis.

Our analysis is based primarily on three types of data. First, we have the sampling year and spatial position of each plant. Second, each plant has been genotyped at $\sim$ 100 SNP markers. Six of the SNPs (used in the subsequent cline analysis) are in tight linkage disequilibrium with causal alleles at six of the known colour loci. These include three loci that influence the production and spread of magenta anthocyanin pigment (*Rosea, Eluta* and *Rubia*) and three that influence the production of yellow aurone pigmentation (*Sulfurea, Flavia, Cremosa*). Except for *Rosea* and *Eluta*, which are located about 0.5 cM apart on LG6, the loci are on different chromosomes (Table 5.1). The remaining SNPs were chosen based on their power to identify parent-offspring trios across our temporal dataset. Because the spatial locations of the parents and offspring are known, inferred trios provide us with direct estimates of dispersal in our study area. Third, we have qualitative colour scores for each individual, describing the level of magenta and yellow colouration of one sampled flower. Because these scores are subjective (flowers were scored by different observers over the last decade), we limit our analysis of these scores to a qualitative comparison with the SNP clines.

Table 5.1: Information about the colour loci and SNPs studied, with basic descriptions of the effect of each locus. The dominant allele is capitalized (e.g., *ROSEA*), the superscript letters indicate which subspecies the allele is typically associated with: P = *A.m.pseudomajus*, S = *A.m.striatum*. LG is the linkage group that the locus is found on. SNP position (cM) is the position of the genotyped SNP in centimorgans. ΔP is the allele frequency difference between the left (yellow) and right (magenta) flanks, defined as $z < 13$ km and $z > 16$ km on the 1-D transect respectively.

Name (SNP marker)	**Phenotypic Effect**	**LG**	**SNP position (cM)**	**ΔP**
Rosea (ros_assembly _543443)	Promotes expression of magenta anthocyanin pigments. The semidominant $ROSEA^P$ allele causes strong magenta pigmentation across the corolla. Individuals homozygous for the $rosea^S$ allele produce little to no anthocyanin pigment, depending on alleles at other loci (Tavares et al., 2018).	6	49.29	0.855
Eluta (ros_assembly _715001)	Regulates the distribution of anthocyanin. In the presence of $ROSEA^P$, the dominant $ELUTA^S$ allele restricts magenta colouration to the bee entry point. Individuals homozygous for the $eluta^P$ allele have smooth magenta pigmentation across the corolla (Tavares et al., 2018).	6	49.86	0.759
Sulfurea (s91_122561)	Regulates the distribution of yellow aurone pigment. $SULFUREA^P$ restricts pigmentation to the bee entry point; The dominant $sulfurea^S$ causes the expression of aurone across the entire corolla (Bradley et al., 2017)).	4	9.37	0.614
Flavia (s316_93292)	Regulates the expression of yellow aurone pigments. $flavia^P$ reduce pigment intensity across the flower. $flavia^P$ has a smaller effect on reducing yellow colouration than $SULFUREA^P$ (Richardson et al. in prep.).	2	43.18	0.640

Name (SNP marker)	**Phenotypic Effect**	**LG**	**SNP position (cM)**	ΔP
Rubia (s261_720757)	Regulates the production of anthocyanin pigments when. $RUBIA^P$ mainly increases the intensity of magenta pigmentation in the presence of $ROSEA^P$.*RUBIA* has little to no effect in individuals homozygous for $rosea^S$ (Field et al. in prep.).	5	32.32	0.427
Cremosa (s1187_290152)	Encodes an additional regulator that has a subtle effect on yellow patterning. The *pseudomajus* and *striatum* alleles are semi-dominant (Bradley et al. in prep.).	1	2.18	0.654

5.2.2 Clines at flower colour loci have similar positions but different shapes

We first fit clines to each of the 6 colour loci separately (see Methods: SNP Panel, Table 5.1). To facilitate cline fitting, we first clustered plants into 999 demes and calculated their position along a one-dimensional transect (z) (see Methods). A 25m diameter was chosen because it yielded demes with enough individuals to estimate local allele frequencies and LD (mean of 23 individuals per deme; 337 demes with at least 10 individuals, Table C.1), and minimized departures from Hardy-Weinberg equilibrium (Fig C.3, Appendix C.1.2). We then used the Metropolis-Hastings algorithm to fit three cline models to the allele frequency data, while describing variation around the cline using F_{ST} (see Methods). These included (i) a simple sigmoid model, (ii) a sigmoid model with polymorphism in the tails of the cline, and (iii) a stepped model, characterized by a central sigmoidal shape flanked by long exponential tails of introgression which may differ to the left and right (Fig 5.1B). Analysis of one locus (*Rosea*) found no changes in cline position or width over time (Fig. C.5, Table C.3), so we combined individuals across all years. We used likelihood-ratio tests to identify the best-fitting model for each locus, and obtained maximum-likelihood estimates (MLE) for each cline parameter (Table 5.2).

All six of the colour loci show sharp clines which match the transition in flower colour scores (Fig. C.4, Fig 5.3A). The (asymmetric) stepped cline model fits the data better than the simpler sigmoid model for all 6 loci (Table C.2, Fig 5.3A). We examined the MLE (see Table 5.2) and the marginal likelihood to see how parameters differed among

loci. The cline centres were highly similar among loci (Fig 5.3B, Table 5.2), ranging from 14.43 km to 15.3 km (mean c = 14.96 km), and differing from one another by an average of 0.34 km (maximum of 0.9 km). Cline widths, on the other hand, differed markedly among loci (Fig 5.3C), ranging from 0.76 km (*Rosea*) up to 4.5 km (*Rubia*). We also found striking differences in the degree of polymorphism and introgression in the tails of the cline. While most loci showed fixation of alternate alleles on either side, some loci showed appreciable polymorphism in the yellow flank (i.e., *Rubia* and *Sulfurea* where the *A.m.pseudomajus* allele has a frequency around 0.2). Additionally, *Flavia* and *Cremosa* showed strong asymmetries in the cline shape with rates of decay of tails of introgression that are much higher on one side (e.g., $\theta_1 > \theta_0$).

Taken together, and in the context of the low genome-wide differentiation across this hybrid zone (Tavares et al. 2018), our results are consistent with the hypothesis that loci are targets of pollinator-mediated selection. The varying widths indicate that some loci experience stronger selection than others, likely owing to their individual effects on coloration. However, the similar positions of the cline centers is expected, given that the difference in colour pattern between *A.m.pseudomajus* and *A.m.straitum* is determined by the individual contributions and interactions of these loci. This coincidence of clines may be due to the presence of associations among six loci, which we examine next.

SNP	B_0 (km)	B_1 (km)	θ_0 ($*10^{-3}$)	θ_1 ($*10^{-3}$)	Width (km)	Centre (km)	F_{ST}
Rosea	14.4 (8.2 – 21.8)	6.1 (1.05 – 17.7)	2.4 (0.7 – 4.3)	31.2 (9.9 – 82.4)	0.8 (0.6 – 0.9)	14.94 (14.91 – 14.95)	0.07 (0.07 – 0.09)
Eluta	48.5 (27.4 – 125.5)	29.6 (17.9 – 57.1)	0.4 (0.1 – 1.3)	0.4 (0.2 – 1.1)	0.8 (0.7 – 0.9)	14.97 (14.95 – 15.0)	0.06 (0.06 – 0.08)
Rubia	9.6 (8.2 – 10.5)	59.8 (30.5 – 67.7)	8.1 (8.2 – 18.2)	3.6 (3.5 – 19.4)	4.5 (4.5 – 6.4)	14.43 (14.02 – 14.42)	0.04 (0.04 – 0.06)
Sulfurea	28.3 (16.1 – 51.7)	9.1 (3.2 – 19.9)	1.2 (0.5 – 3.0)	30.7 (12.6 – 76.5)	3.5 (3.0 – 3.9)	14.94 (14.87 – 14.98)	0.03 (0.03 – 0.04)
Flavia	29.6 (16.9 – 83.0)	2.4 (2.3 – 3.1)	2.0 (0.4 – 3.6)	10.8 (5.3 – 13.0)	1.3 (0.9 – 1.4)	15.18 (15.11 – 15.22)	0.04 (0.03 – 0.05)
Cremosa	19.0 (10.4 – 62.0)	2.4 (2.0 – 3.2)	3.6 (0.5 – 8.7)	19.0 (11.8 – 29.0)	1.7 (1.5 – 2.0)	15.31 (15.25 – 15.37)	0.04 (0.03 – 0.04)

Table 5.2: Maximum likelihood estimates of the 7 cline parameters for the stepped cline model for each flower colour locus. The cline width is the inverse of the maximum gradient in allele frequency, the cline centre denotes the position on the transect at which both alleles are equally frequent and F_{ST} captures variation in allele frequency around the cline. B_0 and B_1 are the strength of the barrier of gene flow to the left and right of the cline centre, and θ_0 and θ_1 are the rate of decay of the exponential tail to the left and right sides, respectively.

5.2.3 Associations between flower colour loci extend beyond the cline center

Studies of hybrid zones often find strong associations between alleles at unlinked loci (i.e., linkage disequilibrium, LD) in areas where clines coincide (e.g., Bombina, Szymura and Barton (1986, 1991); Heliconius, Mallet et al. (1990); Vandiemenella, Kawakami et al. (2009), Mus musculus, Searle (1991)). The LD is mainly attributed to the influx of distinct multilocus genotypes as parental individuals disperse into the zone from both

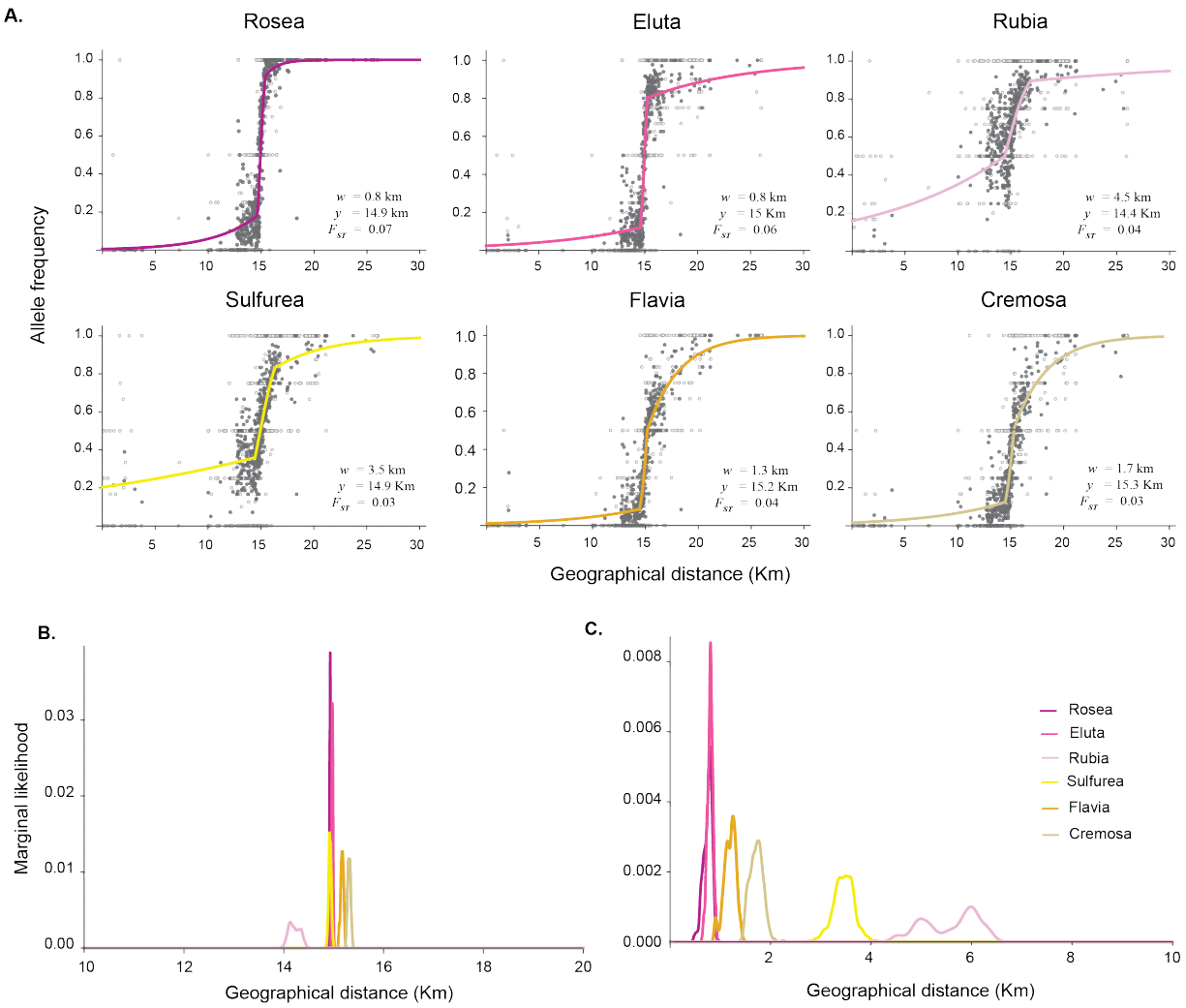


Figure 5.3: **Stepped and coincident clines in snapdragons:** (A) Best fit cline shape and parameters, including the cline width (w), cline centre (c), and F_{ST}, for the 6 flower colour loci. Gray and light gray circles denote demes with greater and less than 5 individuals respectively. The curves show the asymmetric stepped cline shapes fitted to the observed allele frequencies for all loci. (B) The marginal likelihood of cline centres for all 6 loci. Clines coincide, with all cline centres positioned within 1 km. (C) The marginal likelihood of cline width for each locus. Cline widths vary, with Rosea narrowest and Rubia widest. For B and C, the 1D marginal likelihood distribution is obtained from the last 2500 runs of the Metropolis.

sides (Barton, 1983, 1979b; Barton and Shpak, 1990). If individuals disperse only short distances, then since recombination halves LD each generation, it should decay rapidly with increasing distance from the cline centre.

To test for associations in the hybrid zone, we calculated standardized LD within each deme for the five unlinked loci (excluding *Eluta*, noting that choosing *Rosea* or *Eluta* makes little difference due to their tight association). We used an LD estimator, $\hat{R}$, based on the hybrid index (HI), calculated separately for each pair of loci (maximum likelihood gives similar results). HI is defined as the number of *A.m.pseudomajus* alleles carried by each individual, such that with 5 loci, HI ranges from 0 to 10. The mean HI in each deme, calculated from all 5 loci, is plotted along the transect in Fig. 5.4A; it shows a clear stepped pattern and has a shape and position qualitatively similar to the single locus clines. The variance of the hybrid index, $Var(HI) = 2\sum_i p_i q_i + 2\sum_i p_i q_i F_{IS,i} + 2\sum_{i \neq j} D_{ij}$, separates into components due to heterozygosity (V_{PQ}), heterozygote deficit (V_F), and linkage disequilibrium respectively (modified from Barton and Gale (1993)). Both LD and heterozygote deficit inflate the variance in the hybrid index, meaning that the excess variance can be used to estimate the standardised LD or correlations relative to Hardy-Weinberg linkage equilibrium given as $\hat{R} = \frac{Var(HI) - V_{PQ} - V_F}{V_{PQ}}$. $\hat{R}$ ranges from -1 to 1, where 0 implies no correlation, and -1 and 1 imply negative and positive correlations respectively. Because selection against hybrids could lead to a deficit of heterozygotes, we also calculated heterozygote deficit, F_{IS}, for each locus within each deme. To reduce noise among demes, we calculated mean $\hat{R}$ and F_{IS} in 12 non-overlapping bins along the transect, and 3 coarser bins to examine LD inside and outside the cline centre.

Combining individuals from the centre and flanks, we observed significantly positive correlations between all but one pair of loci (permutation test, see Methods). *Rosea* had the strongest associations with *Flavia* (0.089) and weakest with *Rubia* (0.029), while *Rubia* and *Sulfurea* showed weak LD (-0.002) (see Table C.7, Fig C.7). However, the mean correlations ($\hat{R}$) across all pairs of loci were variable and non-zero across the 12 bins along the transect, ranging from 0.006 to 0.12 (Table C.6). $\hat{R}$ was significantly positive but modest in the cline centre, averaging 0.027 across the central 2km region (Fig 5.4B). Additionally, we found no evidence for a marked deficit of heterozygotes at the cline centre, as most bins had a mean F_{IS} near 0, after averaging over the 5 loci (Table C.6, Fig 5.4C). However, *Rosea* showed a deficit of heterozygotes ($F = 0.034$) possibly due to strong selection or assortative mating (Table C.5, Fig C.6). Overall, the low levels of LD and F_{IS} show that there is only a weak genome-wide barrier, consistent with Ringbauer et al. (2018).

Traditional cline analysis allows the inference of dispersal and selection from associations and cline widths (Barton and Gale, 1993). Mean correlations in the centre of the hybrid zone can be used to estimate dispersal as $\sigma = w\sqrt{\hat{R}/8}$, where w is the cline width and

σ is the root mean square of parent-offspring dispersal distance along some axis. The mean correlations between all unlinked loci in the centre, 0.027 and the average cline widths from 5 unlinked loci,$w = 2.1km$ gives $\sigma = 137$m. This estimate of dispersal can in turn be used to infer the strength of selection as $w = \sigma\sqrt{8/s}$, assuming selection against heterozygotes (Bazykin, 1969). Using MLE for cline width (Table 5.2), estimated selection coefficients ranged from 0.23 in *Rosea* to 0.007 in *Rubia* (Table 5.3). This is the effective selection between the populations within and away from the centre of the hybrid zone that is required to maintain the cline of the observed width. However, these inferences are valid only when the assumptions of basic cline theory are met.

Our results highlight two major discrepancies from traditional cline theory and the necessity to adopt a different method, as detailed below. Firstly, the standardized LD in the center of the hybrid zone is weak (∼0.027). In our system, relatively few loci are under selection (Field et al in prep.), and the LD among them is insufficient to cause a sharp step (see Fig C.16). Stepped clines can also reflect a geographical barrier to dispersal (Barton and Hewitt, 1985; Westram and Stankowski, 2022); however, this seems unlikely, as we see no indication of a geographical barrier in this hybrid zone. They may also be caused by introgression of a marker SNP as it recombines away from a tightly linked causal locus (Barton, 1979b); we return to this issue in the *Discussion*.

A second unexpected finding is the presence of positive LD across the entire transect (Fig 5.4B); in fact, the correlation between alleles is higher in the flanks than in the center ($\hat{R} = 0.088$ in the yellow flank and 0.034 in the magenta flank, compared to 0.027 in the center). The simplest explanation is that this is generated by long-distance dispersal from one side of the hybrid zone to the other (Szymura and Barton, 1986, 1991). Indeed, many plant species show substantial rates of long-distance pollen and seed dispersal (Cain et al., 2000; Nathan et al., 2008; Ashley, 2010). Long-distance dispersal directly brings in sets of foreign alleles generating introgression, causing local increases in LD away from the cline center. In such scenarios σ may not suffice to describe dispersal: the diffusion approximation fails, and inferences from traditional cline analysis may be inaccurate.

5.2.4 Dispersal is highly leptokurtic in snapdragons

To understand how dispersal influences the cline shape and LD, we inferred the full pollen and seed dispersal distributions from our long-term dataset. To infer the dispersal kernel, we first used the program SNPPIT to identify 2342 parent-offspring trios, based on our panel of 98 SNP markers (see Methods). For each trio, we first determined the pollen dispersal distance, as simply the distance between the two inferred parents. The vast majority of identified trios (∼99%) come from the densely sampled central area of our study site, meaning that our observed pollen dispersal distances necessarily fall

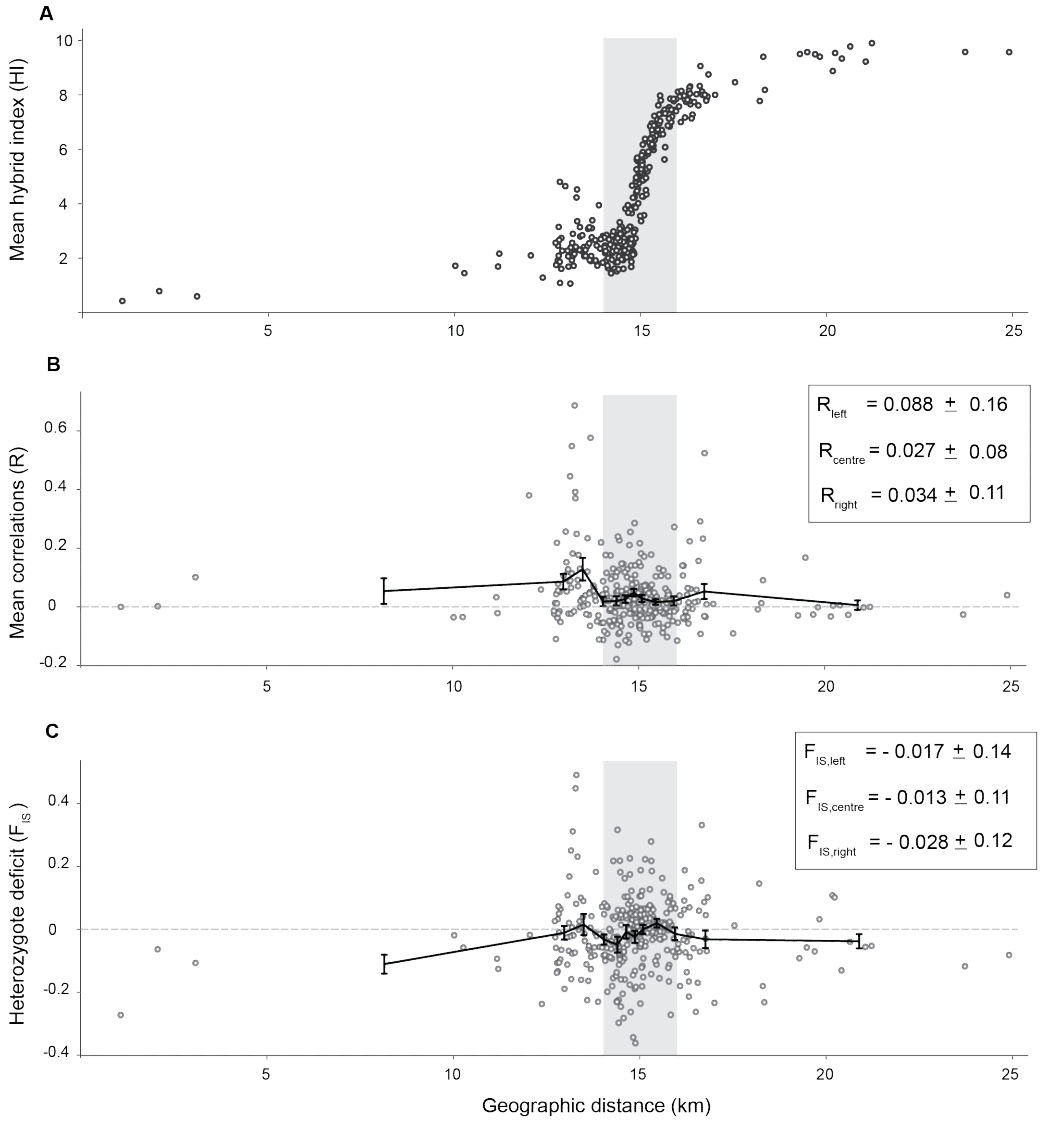


Figure 5.4: **Weak heterozygote deficit and positive correlations across the transect:**(A) Mean hybrid index from 5 unlinked loci shown across geographical distance for 337 demes with at least 10 individuals. (B) Mean correlations ($\hat{R}$) for all 10 pairs of unlinked loci across geographic distance and (C) Mean heterozygote deficit for the 5 loci (F_{IS}). Each point is the estimate for 1 deme, while the black line represents the mean and standard deviation for the 12 bins across the transect. The legends in B and C show the mean F_{IS} and $\hat{R}$ in the centre (denoted by grey bar), magenta flank (right of the grey bar) and yellow flank (left of the grey bar).

within ∼1Km. However, the shape of this short-range distribution closely matched a log-Gaussian distribution, so we extrapolated the kernel beyond this range (Fig 5.5A).

We also used the parent-offspring trios to infer seed dispersal kernel up to ∼1Km. This was less straightforward than for pollen because we lack uni-parentally inherited markers, so could not directly determine which parent was the mother. When parents are close by, both parent-offspring distances are similar, so we can estimate directly; when they are far apart (>100m, say), one parent-offspring distance is typically much lower, suggesting that the closer parent is the mother (see Fig. C.8). We can do better by using a recursive algorithm, based on the assumption that seed and pollen dispersal are independent; this gives a similar estimate to the two simple approaches (see Methods).

To infer dispersal beyond the range of the pedigree, we identified long-range migrants in the flanks as having multilocus genotypes which are unlikely to be produced locally by random mating given the local allele frequencies (see Appendix C.1.3). Using this approach, we found that 3.7% and 1.1% of individuals in the yellow and magenta flanks respectively were more likely to be the result of mating between a yellow and magenta individual (i.e an F1 hybrid) or a long-range seed disperser ($p < 2 * 10^{-4}$; see Table C.17). We attribute these primarily to long-range seed dispersal, because pollen dispersal would generate F1-like individuals, which are rarer than genotypes from the opposite parental population. Because we assume long-range seed dispersal to be more common and due to the lack of F1-like individuals in the improbable individuals, we assume long-range pollen dispersal to follow a log-Gaussian distribution. We inferred long-range seed dispersal separately for individuals in each flank (i.e., yellow flank to magenta flank and vice versa), assuming an exponential function which was spliced onto the short-range distribution at 300m, to give a full seed dispersal kernel (see Methods).

Together, these approaches showed that dispersal is highly leptokurtic. Half of fathers were within 18m of the mother, 75% within 55m, and 93% within 200m (Fig 5.5A). Seed dispersal however had longer tails than pollen, with ∼60% of dispersal occurring within 5m and ∼80% within 10 m (Fig 5.5A), but with very long tails in both directions (Fig 5.5B); the scale of decay in the tails was inferred to be λ_{left}= 4.5km to the left flank and λ_{right}= 2.6 km to the right. Long-range seed dispersal was higher from the magenta to yellow end, consistent with the high rate of introgression seen from the hybrid index score and greater LD observed in the yellow flank (Fig 5.4A,B). These long-range seed dispersers can directly bring in foreign genotypes, thereby generating positive associations between alleles at different causal loci away from the cline center.

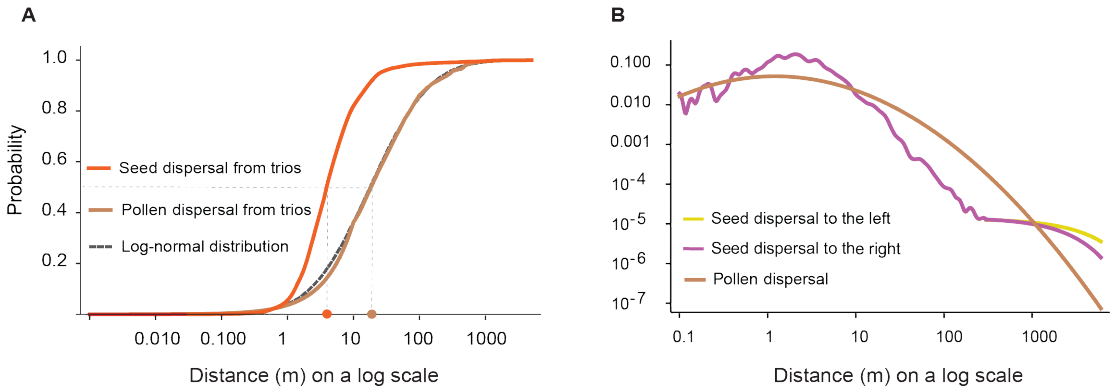


Figure 5.5: **Dispersal is leptokurtic is snapdragons** (A) Cumulative distribution (CDF) of seed and pollen dispersal estimated from the pedigree trios. The brown curve denotes the CDF of the pollen dispersal (i.e. the distance between the parents). The black dashed line denotes the fitted log-Gaussian distribution. The orange curve denotes the CDF of seed dispersal, estimated from the trios. For each curve, the gray dashed lines denote the median distance, where the probability is 50%. (B) The full seed and pollen dispersal distributions, with the probability of movement of a given distance shown in the Y axis. The brown curve denotes the full pollen dispersal kernel. The magenta and yellow curves denote the seed dispersal kernel estimated by splicing the PDF estimated from the trios (as shown in A) with that estimated from individuals in the magenta and yellow flanks respectively.

5.2.5 Using simulations with empirical dispersal to explain cline shape and infer selection

We observe associations (LD) even between unlinked loci, and see strongly leptokurtic dispersal. In order to find the relative contribution of these to distorting clines away from a simple sigmoid, and to estimate the strength of selection that best explains our data, we simulate the five unlinked loci (excluding *Eluta*, but noting that due to tight linkage between *Rosea* and *Eluta* and rarity of recombinants (Tavares et al., 2018), simulating *Rosea* includes the effect of both these loci), using the observed dispersal kernel. There is a plethora of possible models for selection: with 5 loci, we have 243 diploid genotypes, which can each be assigned a fitness that depends on the frequency of all the genotypes. We are skeptical that we have much power to estimate much more than the net selection on each locus, and so choose as simple a model as possible, at least for this first analysis. We assume random mating, so that diploid genotype frequencies can be constructed from the frequencies of the 2^5=32 haplotypes. We assume an infinite population, using a deterministic simulation that neglects drift. We also assume that selection is multiplicative across loci, and acts against heterozygotes; it may be asymmetric, so that contribution to fitness at a locus is $1 + s_{left} : 1 : 1 + s_{right}$. Though we follow haplotype frequencies, we find the best fit to the allele frequency clines. Because the clines are strongly stepped, we use unevenly spaced demes, so that we can follow steep gradients in the center, and shallower tails out at the edges. The migration matrix is constructed from the CDF of

the observed dispersal, to give the fraction in each 'deme' that derives from each of the other demes in every generation. This scheme should be seen as a surrogate for a range of more detailed models; for example, positive frequency-dependent selection without dominance is equivalent to underdominance. Essentially, the selection coefficients give the rate of change of allele frequency on either side, and represent marginal values that absorb effects of dominance, epistasis and frequency-dependence.

Clines that are maintained by selection against hybrids, or by positive frequency-dependence, are not tied to any particular location, and so will tend to move. Such movements are likely to be countered by small density gradients; assuming diffusion, a wave speed v can be balanced by a gradient in log density v/σ (Barton, 1979a). Therefore, we shift the simulation such that it is always centred in the same place, and find its most likely location with respect to the observed allele frequencies. When we simulate a single locus, asymmetric selection is redundant, since it just shifts the cline without altering its shape. However, when we simulate multiple loci, they can be held together by LD, so that differences in asymmetry between loci cause differences in position, as well as differences in the rate of decay of tails of introgression on either side. Simulated allele frequencies were fit to the observed allele frequencies after 500 generations, to infer MLE of the strength of selection, cline centre, and F_{ST}. We first validated the simulation by considering Gaussian dispersal with σ =200m and s= 0.1 and following the allele frequencies at a single locus. As expected, we get a sigmoid cline with width given by $\sqrt{8}\sigma/\sqrt{s} = 1800$m (see Fig C.11).

To understand how long-range dispersal affects cline shape in the absence of LD, we first simulated each locus separately. Simulated clines with only the short-range dispersal (i.e dispersal inferred from the pedigree trios with distances up to 300m) failed to generate a strong step (Fig C.12). However, inclusion of long range dispersal generated stepped clines at all loci (log $L = -9385.3$ with 15 df, Fig C.13, Fig 5.6A). The inferred selection coefficients for the 5 unlinked loci ranged from 0.196 to 0.024 (Table 5.3). The simulated clines matched the observed step and patterns of introgression in the tails for *Rosea* but failed to explain the complete cline shape at other loci (Fig C.13, Fig 5.6A). We next simulated clines at 5 unlinked loci, allowing asymmetric selection, which causes differences in the cline position and rate of introgression. Asymmetric multilocus selection, including long-range dispersal, best explained the observed stepped clines at all loci (log $L = -9257.8$ with 12 df, Fig 5.6A). The inferred selection estimates (see Table 5.3) were roughly symmetric for *Rosea*, *Sulfurea* and *Rubia*, and differed only by ∼0.02 between the two homozygotes for *Flavia* and *Cremosa* which showed strong asymmetric cline shapes, consistent with descriptive cline fits. Similarly, cline centres from cline fitting ($c = 14.96$km) matched closely with that from the simulations ($c = 14.97$ km). However, this simple model does not capture the complete cline shapes, especially at *Flavia* and *Cremosa*.

We next compared the observed standardized LD with that seen in multilocus simulations. We found that LD tended to be overestimated by the simulation. One cause of this discrepancy might be due to dependence of LD on the life stage at which it is measured, which is immediately after populations mix in the simulations. Secondly, long-range seed dispersers might experience greater selective disadvantage since bees tend to prefer the common phenotype (also, confirmed by the lack of F1's in the flanks), causing a reduction in introgression and thus lowering LD. However, there is good agreement between the simulated and observed LD patterns (Fig. 5.6B): mean LD from the simulations was positive across the entire transect, and stronger in the tails of the cline ($\hat{R}_{left} = 0.073$, $\hat{R}_{right} = 0.053$) than at the center ($\hat{R}_{centre} = 0.051$). We also found that the strength of the simulated pairwise LD varied as expected, with stronger LD among pairs of loci under stronger selection (Fig C.14). These results confirm that long-range dispersal can generate widespread LD and stepped clines.

The multilocus simulations also allow us to separate the direct, indirect, and total effective selection acting on each locus. The direct estimates of selection (taken as the average of s_{left} and s_{right} for asymmetric selection) were lower than from the single locus simulations. This is due to the effect of LD (indirect selection) which increases the total effective selection experienced at a locus, thus lowering the inferred direct selection. The total effective selection at any locus i ($s_{eff,i}$) is the sum of direct selection experienced by that locus individually (s_i), and the indirect selection due to associations with other selected loci ($s_{LD,i}$). This can be calculated in the cline centre as $s_{eff,i} = s_i + s_{LD,i} = s_i + \sum_{i \neq j} s_j \hat{R}_{ij}$, where $s_i = (s_{left} + s_{right})/2$ is the average selection coefficient from asymmetric multilocus simulations at locus i and $\hat{R}_{ij}$ is the correlation between locus i and j. This varied from 0.199 to 0.028 among the 5 loci at the center of the hybrid zone (Table 5.3).

Effective selection at each locus was similar to that from single locus simulations, suggesting that multilocus selection has a weak effect on cline shape. This is also seen qualitatively from the cline shape, where LD due to multilocus selection only caused a moderate increase in the step (Fig 5.6A, Fig C.13). This effect of LD is captured by the indirect selection on each locus, s_{LD} and its contribution to effective selection is s_{LD}/s_{eff}. For each locus, the indirect selection contributed roughly 30% (on average) of the total effective selection. It was weakest in *Rosea* (6.7%) and strongest in Rubia and Sulfurea (44%) as expected, because the effect of LD is expected to be strongest for weakly selected loci. Moreover, multilocus simulation in the absence of long-range dispersal (i.e taking dispersal to be Gaussian with $\sigma = 161$m as estimated from cline widths from descriptive cline fitting and inferred s from simulations) was unable to cause stepped clines (Fig C.16). In this scenario, the pattern of correlations among loci was as expected from traditional cline theory: correlations were stronger at the cline center ($\hat{R}$ =0.036) and decayed quickly outside the center (gray curve in Fig 5.6B). Together, these suggest the

multilocus selection alone (without long-range dispersal) is insufficient to explain the observed cline shape and LD. Thus, stepped clines in snapdragons are caused primarily by the effect of long-range dispersal, rather than by the multilocus barrier to gene flow.

In fact, since introgression in the tails is driven by long-range dispersal from the opposite parental type, we can consider the left and right flanks as two demes and calculate the factor by which selection on incoming long-range migrants reduces the rate of influx of unlinked neutral alleles, which is given by the effective migration rate (m_e). Under low rate of migration (m), $g = \prod_i W_i$, where W_i is the fitness of i^{th} generation descendenat of a migrant (see Chap. 2). Taking the average selection coefficient for each locus, the fitness of each successive migrant (F1, BC1..) can be calculated under our simulation model, giving $m_e/m = 0.5$. Thus, the rate at which an incoming neutral migrant allele associated with the opposite phenotype gets established is half the rate at which they enter.

SNP	Selection from 'traditional' cline analysis	Direct selection from single locus simulation	Direct selection from multilocus simulation s_{left}	Direct selection from multilocus simulation s_{right}	Indirect selection due to LD s_{LD}	Effective selection (direct+indirect selection due to LD) $s_{eff,i}$
Rosea	0.228	0.196	0.187	0.185	0.013	0.199
Rubia	0.007	0.024	0.016	0.015	0.012	0.028
Sulfurea	0.012	0.057	0.033	0.028	0.024	0.055
Flavia	0.098	0.091	0.076	0.058	0.023	0.090
Cremosa	0.050	0.097	0.070	0.052	0.022	0.084

Table 5.3: Inferred selection coefficients for single locus cline simulations and multilocus cline simulations with asymmetric selection for the 5 unlinked loci. s_{left} and s_{right} are the selection coefficients for the two homozygotes respectively. The last column is the effective selection at each locus at the cline centre due to the effect of LD between other selected loci.

5.3 Discussion

Our analysis of genotype data from 22,494 individuals has provided new insights about the processes shaping the flower colour clines in this hybrid zone. We found classic hallmarks of hybrid zones, including stepped geographic clines at all six colour loci (Fig 5.3A), and significant LD at the cline center (Fig 5.4B). However, some of our results were not predicted by traditional cline theory, including positive LD across the entire transect, and leptokurtic dispersal of seed and pollen. We used a multilocus simulation that allowed us to estimate selection to show that the stepped cline shapes are caused primarily by long-distance dispersal rather than reflecting a strong genetic barrier to gene flow. Here we elaborate on some of the key conclusions and highlight the general implications of our work for the study of hybrid zones.

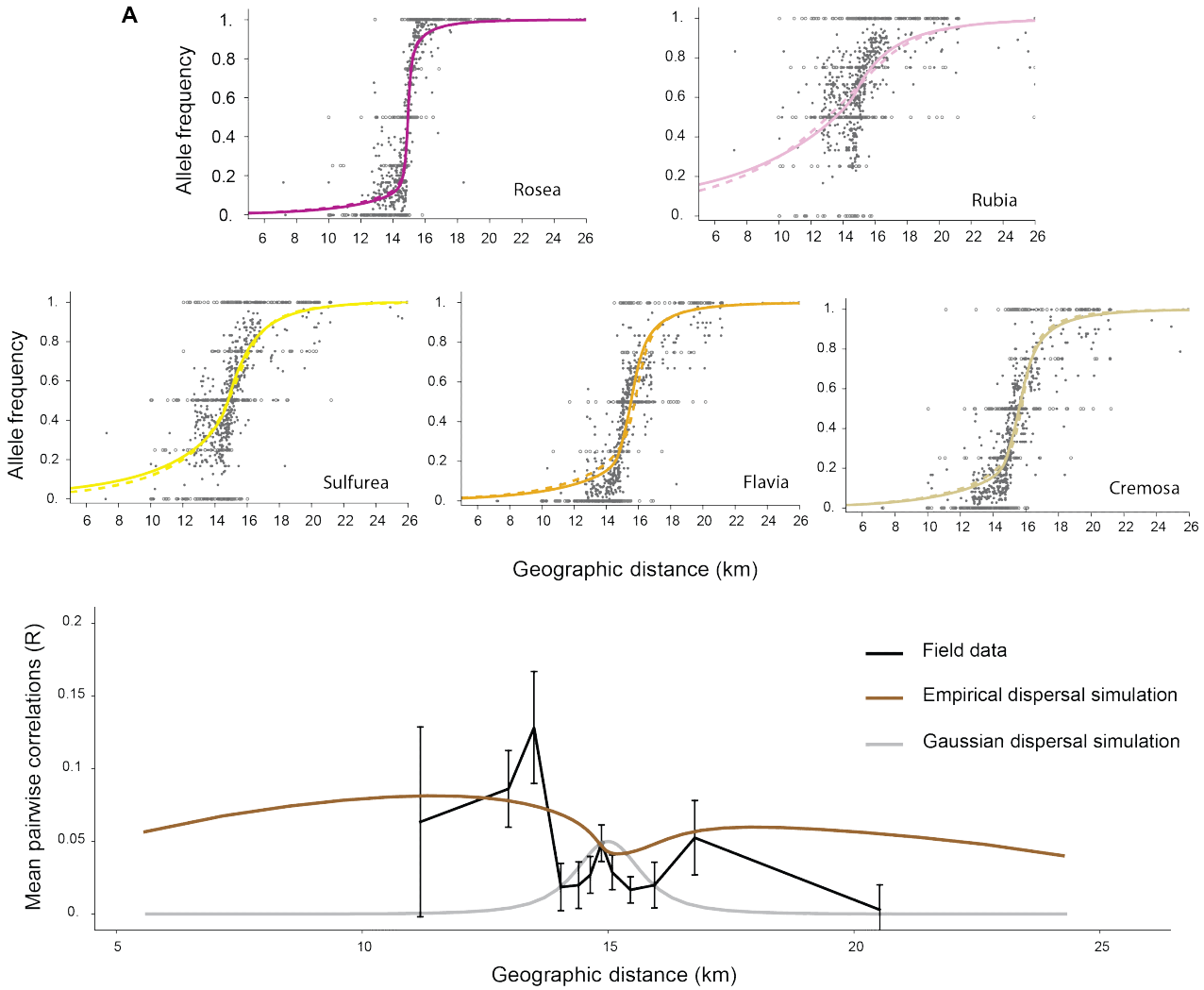


Figure 5.6: **Multilocus cline simulations with the inferred dispersal explain the step and LD** (A) Cline shapes inferred from simulations with asymmetric selection and full dispersal kernel for each locus. The dashed line is the best fit simulated cline from single locus simulations whereas the solid line represents the fit from multilocus simulations with asymmetric selection. The corresponding selection estimates are given in Table 5.3. The light gray and dark gray circles denote the observed allele frequencies of demes with at least 5 and greater than 5 individuals respectively. (B) Mean correlations from 10 pairs of unlinked loci in the observed data (black) are shown for the 12 bins across the whole transect. The brown and gray curves show the mean correlations from multilocus simulations with the inferred dispersal and Gaussian dispersal (with σ =161m) respectively.

Our study provides compelling evidence that the flower colour clines are maintained by a balance between dispersal and selection. Although several studies have characterized phenotypic and genetic variation across this hybrid zone, it has been difficult to exclude alternative explanations for the observed patterns. For example, levels of genetic differentiation at the colour loci provide strong indirect evidence that selection acted in the past to generate 'coadapted' phenotypic divergence (Tavares et al., 2018). However, these analyses tell very little about how selection and gene flow have acted over the spatio-temporal scale relevant to the formation of this hybrid zone (Westram and Stankowski, 2022). Similarly, past studies of the hybrid zone have characterized the steep clines, arguing that their presence is, in itself, indicative of ongoing selection (Whibley et al., 2006). However, sharp clines can also form in absence of selection following recent secondary contact and may persist for some time if dispersal is low (Endler, 1977). In this study, we show directly that dispersal in the hybrid zone is high enough that the maintenance of the color clines ultimately requires counterbalancing selection. We also found patterns of LD that directly imply contemporary selection. Future work on this hybrid zone will estimate the direct fitness consequences of colour using the long-term pedigree.

We also found that the strength of selection varies among the color loci, with estimates of direct selection ranging from 2% to 18% (Table 5.3). This variation is most likely due to the effects that each locus has on flower colour. For example, the tightly linked loci *Rosea* and *Eluta*, which have the largest selection coefficient (18%), are the major determinants of magenta colouration. In contrast, *Rubia*, which has the smallest selection coefficient (2%), has a relatively modest effect on colouration (Field et al., in prep.), subtly increasing the intensity of magenta if it is already present (Table 5.1). Although this interpretation makes intuitive sense, because pollinator preferences are thought to be driven by the differences in coloration, other factors may also contribute to variation in our estimates of selection. For example, strength of LD between our SNPs and the causal mutations that control colour may vary among the loci. Thus, more work is needed to understand and formally test for a relationship between the strength of selection and phenotypic effects of the loci, and also, to disentagle the effects of linked SNP.

A unique aspect of our study, which was critical to our main findings, is that we inferred the full dispersal distribution using a combination of parent-offspring trios and an inference scheme to identify long-distance migrants. This showed that dispersal was highly leptokurtic, characterized primarily by short range dispersal but with a long tail of dispersal events that extend up to and beyond 1km (Fig 5.5B). Leptokurtic pollen dispersal is commonly observed in bee-pollinated plant species, and has already been documented over a smaller scale in a paternity study in snapdragons from our study area (Ellis et al., 2024) and the bees that pollinate snapdragons (Bombus spp.) tend to forage locally, but are routinely recorded moving longer distances well beyond 1km (Walther-Hellwig and

Frankl, 2000; Osborne et al., 2008). We also found that seed dispersal distribution was also leptokurtic, with a small fraction of dispersal events exceeding 1km. Although the vector for long range seed dispersal is not clear, the patchy distribution of the species implies long distance dispersal. Moreover, snapdragons are often observed growing on the roofs of buildings and high on the stone walls of churches and castles, also suggesting that long distance seed dispersal may be quite common.

Our estimates of selection do not come simply from cline width and LD, but rather from a simulation framework that allowed us to include the full dispersal kernel together with multilocus selection. Although the diffusion approximation would be accurate in the limit of weak selection (assuming finite moments), selection is moderate and long-distance dispersal is so strong that σ does not accurately describe dispersal in this system. By using the full dispersal kernel, we obtained estimates of selection that are quite different from those obtained using traditional theory, which substantially underestimated s for some loci (Table 5.3, Fig C.15). Nevertheless, our framework has some important limitations. First, it does not include the effects of epistasis, which might explain why the simulation fails to fully capture the observed cline shapes for some loci. We also assumed selection on diploids and equal haplotype frequencies in seed and pollen, but in reality, selection on flower colour is thought to act mainly through pollination and the actual life cycle depends on movements on both seed and pollen. Finally, the simulation framework ignores the effect of heterogeneous population density, which is known to shape patterns of isolation by distance in the snapdragon hybrid zone (chapter 4, Surendranadh et al. (2022)). Elaborating the model to include these factors may improve the fit between the observed and simulated data, yielding more precise estimates of selection in future.

The simulation scheme also allowed us to disentangle some potential causes of the stepped cline shapes. Specifically, we were able to separate the effects of long-range dispersal and LD on cline shape, showing that long-distance dispersal plays a much greater role. However, there are other factors that may have contributed to the stepped shape. First, the rates of introgression into the tails may be influenced by the way each allele impacts colour on the alternative background, and thus its interaction with other selected loci. This effect may be better understood by including epistasis into the simulation scheme. An additional factor that likely contributes to the stepped shapes is that our marker SNPs are only linked to the causal alleles that are under selection. Although a linked neutral locus will experience a barrier to gene flow due to its association with a selected one, it will inevitably recombine onto the alternative genetic background, where it is free to introgress across the cline. This introgression causes the cline to become stepped, with more pronounced tails of introgression observed for SNPs with increasing genetic distance (i.e., in map units) from the selected site. Future work will aim to explore this effect by studying how cline shapes vary around selected loci.

Finally, our study highlights the need for more nuanced interpretations of cline shape that go beyond the explanations established by traditional hybrid zone theory. The conventional interpretation of stepped clines is that they reflect a multilocus barrier to gene flow that impedes the exchange of neutral and selected alleles between alternative genetic backgrounds. The barrier effect is caused by associations among multiple selected loci, which increases the total effective selection at the cline centre where LD is strongest. However, in our study multilocus selection (that accounts for the effect of LD) alone failed to generate a stepped cline and did not cause positive correlations among loci outside the cline center (Fig C.15, 5.6B). Moreover, the indirect selection due to LD was relatively small in multilocus simulations with full dispersal, suggesting that the barrier to gene flow at the center of the hybrid zone is weak. Since stepped clines are primarily a result of increased gene flow between populations in our system, driven by long-range dispersal, estimates of the barrier strength (B, Szymura and Barton (1986); Westram and Stankowski (2022)) from the cline shapes would substantially overestimate the strength of the genetic barrier to gene flow between these snapdragon subspecies. Moreover, because long-distance dispersal leads to geographically widespread admixture, indirect selection due to LD is stronger in the tails of the cline than traditional theory predicts. Critically, leptokurtic dispersal is not unique to snapdragons. Many organisms show complex patterns of dispersal that are not well approximated by Gaussian diffusion. It is not clear how this assumption has affected inferences made from other hybrid zones. Regardless, more detailed empirical studies and more realistic models of hybrid zones are needed in the future.

5.4 Methods

5.4.1 Sample collection and processing

The study site was visited annually from 2009 to 2019 during the main flowering season (late May-early August), aiming to sample every flowering individual. Within the study area, the same locations were re-visited several times over the period, to ensure that we captured plants that flowered at different times during the season. The sampling area encompasses a central region where the majority of plants are hybrids, and flanking regions with mainly parental types (Fig 5.2C, D).

For each plant, we recorded its geographic location with GeoXT handheld GPS units (Trimble, Sunnydale, CA, USA). Repeated visits to the sampled plants over the years have allowed us to quantify the mean position error of these GPS units as ±3.7 meters. We also collected leaf tissue for later genotyping, preserved by desiccating in silica gel. Additionally, we collected one flower to score the intensity and pattern of the yellow and magenta pigments.

5.4.2 Scoring of flower colour

Each flower was visually scored for magenta and yellow colouration based on the intensity and spread of the pigments (Fig C.2). The magenta scores ranged from 0.5 to 5, with a score of 0.5 showing no presence of magenta pigment to 5 showing intense magenta pigmentation throughout the corolla. Similarly, each flower was scored for yellow colouration, ranging from 0.5 (no yellow or yellow colour restricted to the bee entry point) to 3 (full yellow).

5.4.3 Development of SNPs and genotyping

A panel of 248 KASP SNPs spread throughout the genome was developed previously (see Methods in Ringbauer et al. (2018); Surendranadh et al. (2022)). These include markers for divergent loci within or tightly linked to genes responsible for colour differences in snapdragons. All genotyping was conducted by LGC Genomics. We selected representative SNPs that are linked to known flower colour genes to understand the action of selection and dispersal on spatial genetic variation. We had multiple linked loci to choose from so picked the one that showed the highest allele frequency difference between magenta and yellow ends, while also having low missing data (Table 5.1).

5.4.4 Clustering individuals into demes

Rather than considering the individual position of each plant, we clustered neighboring plants into local demes. This was done for two reasons. First, the precise location of each sampled individual makes hardly any difference to the likelihood of a model in which allele frequencies change over a broad spatial scale but greatly increases the computational burden of cline-fitting. Second, we include fluctuations in the true allele frequency in our cline model, which are generated by the evolutionary process, as well as sampling error, to ensure that large demes are not given undue weight (see 'Fitting clines to genetic data').

Our clustering algorithm randomly selects the first individual, and then groups all individuals within the specified radius δ of the focal plant into a deme. It then moves on to the next available individual and repeats the process until all individuals have been grouped. We initially clustered individuals for a range of values of δ= {10, 15, 20, 25, 30, 40, 50, 75, 100, 120, 150, 175, 200, 250, 300m} and calculated the realized number and size of demes (Table C.1). From these, we chose a deme size based on two criteria. First, we checked if there were enough individuals in each deme to estimate local allele frequencies and LD. Second, we checked that F_{IS} was not inflated by the Wahlund effect (Wahlund 1928); (i.e., as δ increases, F_{IS} and F_{ST} increase, due to isolation by distance). This was assessed by checking for departures from Hardy-Weinberg equilibrium for each deme size (see section 'Deviations from HWE' below). The spatial coordinates for each deme were calculated as the average from all plants within the deme.

5.4.5 Mapping onto a 1D transect

In Val di Ribes, the population of *Antirrhinum majus* follows the south-facing slope of the valley, and so is essentially one-dimensional, with snapdragons primarily being found within 100m of two roughly parallel roads. We transform the data by mapping locations $\{x, y\}$ onto a one-dimensional transect line. The distance along the transect is denoted by z. We established the transect line by fitting polynomials of degrees 1 to 8 to the deme positions, and minimising the mean square residuals (see Appendix C.1.2). A polynomial of degree 6 was found to fit the data well (white curve in Fig. 5.2C and D), with most plants following the curve. We found little improvement from higher-order terms. For each deme, the point on the curve closest to the deme is found, and the new coordinate z is the distance to this point along the curve. The transect has an arbitrary starting point and ranges from 0 to ∼30 km.

5.4.6 Fitting clines to genetic data

We fit three cline models to the frequency of the *pseudomajus* allele along the transect for all clinal SNPs. The first model (sigmoid) has allele frequencies ranging from 0 to 1, as predicted for selection against heterozygotes (Bazykin, 1969) or negative frequency-dependent selection with no dominance, $p = \frac{1}{1+e^{-4(z-c)/w}}$ where c is the cline centre and w is the cline width; other single-locus models give a similar shape (Barton and Gale, 1993). We also include random variation in allele frequency around the cline, measured by F_{ST}; sampled frequencies are binomially distributed around the underlying frequency, which follows a Beta distribution with variance pqF_{ST} (Slatkin and Barton, 1989). Thus, we have three parameters $\{c, w, F_{ST}\}$.

For the second model (sigmoid with polymorphism), we adjust the sigmoid curve by allowing for polymorphism on either side: $p_0 + \frac{p_1-p_0}{1+e^{-4(z-c)/w}}$. This model has 5 parameters $\{c, w, p_0, p_1, F_{ST}\}$.

For the third model (stepped), we splice exponential functions onto the left and right sides, $Aexp\{-4\sqrt{\theta}(z-c)/w\}$ (Szymura and Barton, 1986, 1991). This provides a model of a stepped cline with a sigmoid shape in the centre and long exponential tails of introgression on either side. The parameter θ gives the rate of decay of the tail, which is slower than predicted by the simple sigmoid model (i.e. θ <1). The constant A is defined by the strength of the barrier to gene flow, defined by $B = \Delta p/p'$, where $\Delta p = pb_1 - pb_0$ is the difference in the allele frequencies at the transition from sigmoid to exponential and p' is the gradient in allele frequency at the edge of the central segment. Introgression may be symmetric or asymmetric, and so we estimate $\{\ \theta, B\ \}$ separately on either side. Thus, we have seven parameters $c, w, \theta_0, B_0, \theta_1, B_1, F_{ST}$.

The parameters are estimated using the Metropolis Hastings algorithm (Metropolis et al.,

1953; Szymura and Barton, 1986), which generates a random walk that follows the likelihood: random changes are made to the parameter set and the changes are accepted if they increase the likelihood or rejected otherwise. More formally, we start with an initial parameter set and its likelihood L. Uniformly random changes $[-\delta, \delta]$ are made to each parameter sequentially and the new likelihood $L'$is calculated. If L'/L is greater than a uniform random number between 0 and 1, the change is accepted and parameters and likelihood are reset to the new values. Further, we set δ to $\lambda\delta$, where λ =1.01 is the scaling factor. If L'/L is less than the random number, we reject the change but set δ to δ/λ. With enough trials, this will generate a distribution proportional to L. The size of changes, δ, is tuned for efficiency, ensuring that similar numbers of changes are accepted as rejected. Cycling through each set of parameters 3000 times assured convergence.

For each locus, this fitting process was conducted separately for each model. The best-fitting of the three cline models was determined for each locus using likelihood ratio tests (Kawakami et al., 2009). For the best model, the maximum likelihood (ML) estimates of the parameters were used to obtain the cline shape. The Metropolis Hastings algorithm not only gives ML estimates but also provides the likelihood surface.

5.4.7 Comparing clines over time

The stepped cline model was employed to fit clines for *Rosea* at various time points, since this model gave the best fit when considering all years collectively. 11 years of data are separated into three time points; 2009 to 2012, 2013 to 2015 and 2016 to 2019 with 7690, 12409 and 7384 individuals respectively. Pooling individuals from 3 to 4 years into a single time point is a valid assumption as *A. majus* has overlapping generations and average dormancy of 3-4 years (as seen from the pedigree trios).

5.4.8 Cline fit for the phenotypic data

Magenta and yellow colour scores for each individual were rescaled so that they fell between 0 (minimum) and 1 (maximum). We then calculated the mean colour score for each deme which ranged from 0 to 0.889 for magenta and 0 to 1 for yellow. We fit the same three cline models to these mean values to identify the best-fitting model and ML estimates of each parameter, as described above for the genetic data (see section 'Cline fitting for the genetic data').

5.4.9 Deviations from Hardy Weinberg equilibrium

To check for heterogeneity among demes, the deficit of heterozygotes relative to the Hardy-Weinberg expectation, F_{IS}, is calculated for each deme, as 1- the ratio of observed and expected number of heterozygotes, and each unlinked clinal locus at the clustering

scale of δ =25m. This was done by first selecting 337 of 999 demes with at least 10 individuals, and then dividing the transect into 12 bins such that each bin contains a roughly equal number of demes (see Table C.4). The mean and standard deviation (sd) of F_{IS} were computed for each bin. A null distribution of F_{IS} was also generated for each locus and deme by making 1000 replicate draws of diploid genotypes, fixing allele counts as observed, but combining alleles at random. The observed and the null distribution were then compared to check for demes (and bins) that showed significant deviations from Hardy Weinberg equilibrium. Significance was based on the p-value for each deme, i.e. the fraction of shuffled replicates with F_{IS} less than the observed if the latter is less than 0 and greater than observed otherwise.

5.4.10 Linkage disequilibrium based on hybrid index

Linkage disequilibrium (LD) between two unlinked loci was estimated from the variance in hybrid index (HI) as $\hat{R} = \frac{Vart(HI) - V_{PQ} - V_F}{V_{PQ}}$, where V_{PQ} is the variance due to heterozygosity and V_F is the variance due to heterozygote deficit. Since missing data could be correlated across loci and generate a positive bias in the estimator, $\hat{R}$ was measured based on pairs of loci. Individuals with missing data at either of the two loci in consideration were deleted and $\hat{R}_{ij}$ was estimated based on HI calculated from the 2 loci. $\hat{R}$ is then the average of $\hat{R}_{ij}$ for $i \neq j$. $\hat{R}_{ij}$ was first estimated for each locus pair for the whole population to check for significant associations between loci. Next, we looked at how average associations between loci change across the geographic range. $\hat{R}$ was found for the 12 bins as above (see section 'Deviations from Hardy Weinberg equilibrium') by considering 337 demes with at least 10 individuals. In each case, a null distribution was generated by shuffling genotypes across loci for 1000 replicates, thereby preserving HW deviations but generating linkage equilibrium. A p-value was obtained for each deme by finding the fraction of shuffled replicates that gave values less than observed if the latter is less than 0 and greater than observed otherwise.

5.4.11 Estimating parent-offspring trios

We used the SNPPIT software to find 2342 trios (i.e., an offspring with both parents), based on a panel of 98 SNP (Anderson and Garza, 2006; Anderson, 2012). These SNP included 81 non-clinal SNP, with an average $F_{ST} = 0.05$, which largely overlap with the 91 non-clinal SNP described in chapter 4 (Surendranadh et al., 2022)). The remaining 17 SNP, with $F_{ST} > 0.1$, included the clinal SNP analysed here. The rate of genotyping errors was estimated from 4300 plants that had been genotyped multiple times; there were virtually no cases in which one homozygote was mistaken for another, whilst errors mistaking homozygote for heterozygote, or vice versa, were similar; 91 SNP have mean error rates between heterozygote and homozygote 0.00040, whilst 7 SNP had higher error

rates, averaging 0.00583. All individuals were included, provided they missed data at no more than 5 SNP.

We checked for Mendelian incompatibilities in these trios- that is, cases where a parent has the opposite homozygote from the offspring. Amongst the 98 SNP used to ascertain the pedigree, we found 398 trios with at least one incompatibility. The number of incompatibilities at the SNP used to estimate the pedigree was necessarily no more than 1 per parent, since that is the threshold set by SNPPIT. However, individuals were also genotyped for additional SNP, which were not used to construct the pedigree; these showed an average of 0.0032 incompatibilities per trio per SNP, compared with 0.0020 for the SNP used to construct the pedigree. This number is much higher than expected from errors in genotyping, but much lower than expected if one of the parents were mis-assigned to a random member of the population. Moreover, the distance between parents, and between parents and offspring, hardly increases with the number of incompatibilities (Fig. C.9). This indicates that mis-assignments are to close relatives, which usually live nearby.

5.4.12 Estimating dispersal for *A. majus*

The distribution of dispersal distances between parents (pollen dispersal) and between each parent and offspring from 2342 trios is transformed into a log-probit scale such that a log-Gaussian would be a straight line. The pollen dispersal is close to this form for distances up to 600 m that can be estimated reliably from the trios (Fig 5.5A). Thus, we extrapolated the pollen dispersal r to longer distances as given by the log-Gaussian distribution; $P_{pollen}(r) = \frac{1}{r\sigma\sqrt{2\pi}} \exp\left\{-\frac{(z-log_e r)^2}{2\sigma^2}\right\}$, where MLE of mean ($z$) and standard deviation (σ) of $log_e r$ were 2.88 m and 1.64 m respectively (which corresponds to a geometric mean pollen dispersal distance of 17.8m).

Since we do not have maternally inherited markers, the mother and the father cannot be distinguished; the mother is a priori equally likely to be either of the parents. The ML estimate of seed dispersal is thus inferred from the trios using an Expectation-Maximisation (EM) algorithm which assumes that the movements of pollen and seed are independent. We start with a proposed distribution, based on the distribution of distances from offspring to the nearest parent, in those cases where the parents are closer than 20m (see Fig. C.8). This is reasonable since the two parent-offspring distances must be similar when parents are close to each other. For each trio, 100 random choices of one or the other parent are made as the mother, weighing these choices by the probability density for that distance. If the parents are close together, distances to the offspring will be similar, and the choice will be weighted roughly equally. If the parents are further apart, distances from each parent to the offspring will differ, and the algorithm will be biased towards the shortest distance with the appropriate weight. The re-weighted sample is then fitted to a smoothed distribution and used in the next iteration. This procedure converges rapidly within 5

iterations, to a distribution which somewhat deviates from a log-Gaussian; we therefore use this empirical distribution (Fig.5.5).

The actual population has a complex distribution, concentrated around two main roads, but spreading out to either side. So far, we have approximated short-range dispersal (<300m) as two-dimensional, estimating a radial distribution. From this, the one-dimensional distribution along some axis (required for fitting clines) is derived by convolving with $\frac{1}{\pi\sqrt{r^2-x^2}}$, which is the probability of moving $-r < x < r$ in some direction, given r. We can then derive the distribution of net short-range dispersal by using the fact that a gene has an equal chance of coming from the mother (via seed dispersal), or the father (via seed plus pollen dispersal). Therefore, the net distribution is an equal mixture of the seed dispersal, and the convolution of seed with pollen dispersal.

Pedigree trios give direct estimates of pollen and seed dispersal, which extend out to ∼1Km. However, longer-range dispersal will be underestimated, because parents are increasingly likely to be outside our study area. We do not make a detailed analysis of this bias, but instead, only estimate seed dispersal up to 300m, which is typically much less than the distance to the nearest populations outside our study area. Longer-range seed dispersal beyond 300m is estimated from individuals in the yellow and magenta flanks separately. We now work in one dimension, since that better describes the transect over large scales.

We estimate longer-range dispersal (up to several km) from individuals that have genotypes typical of the opposite side of the hybrid zone, based on the 5 unlinked SNP; these individuals are presumed to have arrived by seed dispersal. Table C.17 assigns individuals as immediate dispersers, F1, backcross, . . . , based on the observed allele frequencies in the flanks, and assuming HWLE (see Appendix C.1.3). However, we fit exponential tails of introgression using a more sophisticated method, based on the likelihood of the mother's location. The probability $P(X, z)$ of finding an individual with genotype X at location z is: $\int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty} f_{seed}(x_s)f_{pollen}(x_p)G(X|z+x_s, z+x_s+x_p)dx_p dx_s$. This is an integral over the probability $f_{seed}(x_s)$ that the mother was x_s away, times the probability $f_{pollen}(x_p)$ that the father was x_p away from the mother, times the probability G that the offspring has genotype X, given the allele frequencies at the parents' locations, assuming Hardy -Weinberg and linkage equilibrium; this is calculated as a product of allele frequencies across loci, with factors $\{q_y q_z, p_y q_z + p_z q_y, p_y p_z\}$, where p_y and p_z refer to the allele frequencies at locations y and z respectively predicted from the cline fits at that locus (Fig. C.10), and $q = 1-p$ (see Appendix C.1.3 for more details). f_{seed} is the seed dispersal distribution, which takes the form estimated directly from the trios up to 300m, and then spliced onto $Aexp\frac{-(r-300)}{\lambda}$ for distances greater than r= 300m. The log likelihood for in either flank is estimated by summing $log(P(X, z)$ across all individuals in the dataset.

5.4.13 Estimating strength of selection from simulated clines

Simulation framework

We consider a deterministic stepping stone model with 45 unevenly spaced demes that range from -13.5 to 13.5 km. The demes are located at $\{\ldots, -\delta\lambda^2, -\delta\lambda, 0, \delta\lambda, \delta\lambda^2, \ldots\}$ where δ =50 and λ =1.2. These parameters were chosen to have demes that span the length of the one-dimensional transect and for computational efficiency. We start with a sigmoid cline of width 4 km centered at 0, which gives the initial frequencies of 2^l haplotypes at each deme considering l biallelic loci under linkage equilibrium.

We use the estimated dispersal distribution to simulate clines along a one-dimensional transect. For this, a net dispersal was calculated since gene flow is due to both pollen and seed dispersal: the distance moved by an autosomal gene is given by seed dispersal if it derives from the ovule, and by seed and pollen dispersal if it derives from pollen. Thus, dispersal is a mixture of seed dispersal (probability 1/2) and the convolution of seed and pollen dispersal (probability 1/2). The backwards migration matrix is calculated from the CDF of net dispersal, by calculating the fraction of genes in each deme that came from every other deme in the previous generation. Each deme is centred at a point, and extends half-way towards its neighbours on either side. We assume uniform density, so the number in each deme is proportional to its extent; migration probabilities are calculated directly from the CDF. Because we estimated the exponential tails differently on the two sides, we interpolate their distance scale using a logistic function, such that at distance z along the transect from the centre, $\lambda = \frac{(\lambda_0+\lambda_1 e^z)}{(1+e^z)}$. The asymmetric dispersal causes the clines to move due to a lack of density gradient. To overcome the computational burden of recalculating the migration matrix, we fix the locations of demes and shift the clines to be centred at 0 every 10 generations. The cline centre is defined as the location at which the gradient of the interpolated cline is the steepest.

The simulation follows the haplotype frequencies at each deme for 500 generations where the order of events is the random union of gametes, selection on diploid genotype, free recombination and migration of haploids. Note that the simulations only follow haplotype frequencies under HW equilibrium and assumes the same allele frequencies in pollen and ovule. We first simulated clines at a single locus with heterozygote disadvantage and Gaussian dispersal distribution to check whether simulations gave results as expected from theory.

Estimating selection coefficient:

This model was used to simulate clines, either considering a single locus (with fitness $1 : 1 - s : 1$) or 5 unlinked loci (denoting 5 unlinked flower colour loci), giving allele frequencies at simulated loci. For multilocus cline simulations, we considered asymmetric

selection at each locus, where fitness follows $1 + s_{l,left} : 1 : 1 + s_{l,right}$ and is multiplicative across loci. Assuming that sampled frequencies are binomially distributed with the underlying frequencies following a Beta distribution as above (see 'Fitting clines at genetic loci'), this model has parameters F (variance around the cline), c (cline centre) and s (selection coefficient), giving $5 * 3 = 15$ parameters for 5 single locus clines and 12 parameters for multilocus cline simulations, where we assume a single value of c and F for all loci.

For single-locus clines, we first find maximum likelihood estimates of c and F for a range of selection coefficients and then find s that maximizes the total likelihood. With multiple loci, MLE of parameters at each locus depend on the allele frequencies at all other loci. We start with the MLE of s from the single locus fits and then follow the single locus procedure to find the MLE estimates of parameters at any one locus under multilocus cline model. The selection coefficient at that locus is updated to the new value and this procedure is repeated for each locus multiple times until the parameters converge.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank all the interns and volunteers who have helped us with field work and processing over the years and Eva Salmerón Mateu for her help with logistics at the field station in El Serrat, Planoles. We thank the funding agencies, Austrian Science Fund FWF (grant P32166) and ERC grant (HaplotypeStructure 101055327) for supporting this study.

Author Contributions

PS, DF and NB conceived the idea, DF generated the SNP panel and inferred the pedigree, PS performed the descriptive cline fitting, and simulations. NB inferred dispersal and checked the pedigree generation (together with DF). PS, SS and NB wrote the manuscript.

References

Anderson, E. C. (2012). Large-scale parentage inference with snps: An efficient algorithm for statistical confidence of parent pair allocations. *Statistical Applications in Genetics and Molecular Biology*, 11(5).

Anderson, E. C. and Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, 172(4):2567–2582.

Ashley, M. V. (2010). Plant parentage, pollination, and dispersal: How DNA microsatellites have altered the landscape. *CRC Crit. Rev. Plant Sci.*, 29:148–161.

Barton, N. H. (1979a). The dynamics of hybrid zones. *Heredity*, 43:341–359.

Barton, N. H. (1979b). Gene flow past a cline. *Heredity*, 43(3):333–339.

Barton, N. H. (1983). Multilocus clines. *Evolution*, 37:454–471.

Barton, N. H. and Gale, K. S. (1993). Genetic analysis of hybrid zones. In *Hybrid Zones and the Evolutionary Process, ed. R. Harrison*, pages 13–45. Oxford University Press.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.*, 16:113–148.

Barton, N. H. and Shpak, M. (1990). The stability of symmetric solutions in evolutionary models. *Theoretical Population Biology*, 38:210–224.

Bazykin, A. D. (1969). Hypothetical mechanism of speciation. *Evolution*, 23:685–687.

Bradley, D., Xu, P., Mohorianu, I.-I., Whibley, A., Field, D., Tavares, H., Couchman, M., Copsey, L., Carpenter, R., Li, M., Li, Q., Xue, Y., Dalmay, T., and Coen, E. (2017). Evolution of flower color pattern through selection on regulatory small RNAs. *Science*, 358:925–928.

Cain, M. L., Milligan, B. G., and Strand, A. E. (2000). Long-distance seed dispersal in plant populations. *Am. J. Bot.*, 87:1217–1227.

Cayuela, H., Rougemont, Q., Prunier, J. G., Moore, J. S., Clobert, J., Besnard, A., and Bernatchez, L. (2018). Demographic and genetic approaches to study dispersal in wild animal populations: A methodological review. *Molecular Ecology*, 27(20):3976–4010.

Ellis, T. J., Field, D. L., and Barton, N. H. (2024). Joint estimation of paternity, sibships and pollen dispersal in a snapdragon hybrid zone. *bioRxiv*.

Endler, J. A. (1977). *Geographic Variation, Speciation, and Clines*, volume 10 of *Monogr. Popul. Biol.* books.google.com.

Haldane, J. B. S. (1948). The theory of a cline. *J. Genet.*, 48:277–284.

Harrison, R. G. (1993). Hybrids and hybrid zones: Historical perspective. In *Hybrid Zones and the Evolutionary Process*, pages 3–12. books.google.com.

Hewitt, G. M. (1988). Hybrid zones - natural laboratories for evolutionary studies. *Trends Ecol. Evol.*, 3:158–167.

Hudson, A., Critchley, J., and Erasmus, Y. (2008). The genus *Antirrhinum* (snapdragon): A flowering plant model for evolution and development. *CSH Protoc.*, 2008:db.emo100.

Kawakami, T., Butlin, R. K., Adams, M., Paull, D. J., and Cooper, S. J. B. (2009). Genetic analysis of a chromosomal hybrid zone in the australian morabine grasshoppers (*Vandiemenella, viatica* species group). *Evolution*, 63:139–152.

Kruuk, L. E. B., Baird, S. J. E., Gale, K. S., and Barton, N. H. (1999). A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, 153:1959–1971.

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

Mallet, J., Barton, N., Lamas, G., Santisteban, J., Muedas, M., and 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., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Nagylaki, T. (1976). Clines with asymmetric migration. *Genetics*, 83:867–886.

Nathan, R., Schurr, F. M., Spiegel, O., Steinitz, O., Trakhtenbrot, A., and Tsoar, A. (2008). Mechanisms of long-distance seed dispersal. *Trends Ecol. Evol.*, 23:638–647.

Osborne, J. L., Martin, A. P., Carreck, N. L., Swain, J. L., Knight, M. E., Goulson, D., Hale, R. J., and Sanderson, R. A. (2008). Bumblebee flight distances in relation to the forage landscape. *J. Anim. Ecol.*, 77:406–415.

Phillips, B. L., Baird, S. J. E., and Moritz, C. (2004). When vicars meet: A narrow contact zone between morphologically cryptic phylogeographic lineages of the rainforest skink, *Carlia rubrigularis*. *Evolution*, 58:1536–1548.

Pinho, C. and Hey, J. (2010). Divergence with gene flow: Models and data. *Annu. Rev. Ecol. Evol. Syst.*, 41:215–230.

Porter, A. H., Wenger, R., Geiger, H., Scholl, A., and Shapiro, A. M. (1997). The *Pontia daplidice-edusa* hybrid zone in northwestern Italy. *Evolution*, 51:1561–1573.

Raufaste, N., Orth, A., Belkhir, K., Senet, D., Smadja, C., Baird, S. J., and Boursot, P. (2005). Inferences of selection and migration in the Danish house mouse hybrid zone. *Biological Journal of the Linnean Society*, 84(3):593–616.

Ringbauer, H., Kolesnikov, A., Field, D. L., and Barton, N. H. (2018). Estimating barriers to gene flow from distorted isolation-by-distance patterns. *Genetics*, 208(3):1231–1245.

Searle, J. B. (1991). A hybrid zone comprising staggered chromosomal clines in the house mouse (*Mus musculus domesticus*). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 246(1315):47–52.

Slatkin, M. (1973). Gene flow and selection in a cline. *Genetics*, 75:733–756.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, 236:787–792.

Slatkin, M. and Barton, N. H. (1989). A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, 43(7):1349–1368.

Surendranadh, P., Arathoon, L., Baskett, C. A., Field, D. L., Pickup, M., and 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. and 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*, 40:1141–1159.

Szymura, J. M. and 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*, 45:237–261.

Tastard, E., Andalo, C., Burrus, M., Gigord, L., and Thébaud, C. (2014). Effects of floral diversity and pollinator behaviour on the persistence of hybrid zones between plants sharing pollinators. *Plant Ecol. Divers.*, 7:391–400.

Tastard, E., Andalo, C., Giurfa, M., Burrus, M., and Thébaud, C. (2008). Flower colour variation across a hybrid zone in *Antirrhinum* as perceived by bumblebee pollinators. *Arthropod Plant Interact.*, 2:237–246.

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., Elleouet, J., Burrus, M., Andalo, C., Li, M., Li, Q., Xue, Y., Rebocho, A. B., Barton, N. H., and Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proc. Natl. Acad. Sci. U. S. A.*, 115:11006–11011.

Walther-Hellwig, K. and Frankl, R. (2000). Foraging habitats and foraging distances of bumblebees, *Bombus spp. (Hym., Apidae)*, in an agricultural landscape. *J. Appl. Entomol.*, 124:299–306.

Westram, A. M. and Stankowski, S. (2022). What is reproductive isolation? *Journal of.*

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

Chapter 6

General Discussion

In this thesis, I looked at the role of population structure on genetic variation in multiple contexts when populations evolve in the face of gene flow. This thesis focuses on two aspects of population structure, one generated due to assortative mating in a mainland-island model and the other, under patterns of isolation by distance driven by non- uniform density and dispersal, with a specific focus on hybrid zones, thereby bridging together the fields of spatial population genetics and speciation.

After giving a quantitative definition of RI, chapter 2 gives analytical predictions for RI in simple scenarios and lays out difficulties in measuring it in practise. Chapter 3 quantified RI in terms of effective migration rate m_e in a more realistic scenario that considers a polygenic "magic trait", influenced by large number of unlinked loci, which is under both divergent selection across two populations and mediates assortative mating and thereby sexual selection. We obtain accurate theoretical predictions for trait divergence in terms of m_e, which is shown to depend on measurable fitness components of migrants (or F1 hybrids, etc). In the next two chapters (4 and 5), I shift my focus to the effects of heterogeneous population density and dispersal, specifically using a long-term dataset of 22500 individuals from a hybrid zone of *Antirrhinum majus* (snapdragons). Chapter 4 showed that the observed population density and leptokurtic dispersal in snapdragons shape the distribution of heterozygosity and patterns of isolation by distance. Next, we analysed how selection and dispersal shape allele frequency clines at 6 focal flower colour loci. We observed sharp stepped clines at all loci that are primarily shaped by patterns of dispersal with a small contribution from linkage disequilibrium (LD) between loci. Below I discuss broader conclusions, limitations and possible directions for future work.

Beyond simple models of population structure

While uniformly continuous populations are rather well studied (Wright, 1943; Malécot, 1948; Epperson, 2003; Rousset, 2004), it is essential to better understand the role of realistic demography on genetic variation, both theoretically and empirically. This also involves building on existing inference methods and generating novel measures to detect signatures of demographic and evolutionary processes from genetic datasets. Recent years have seen the emergence of geographically extensive datasets (Aguillon et al., 2017; Leslie et al., 2015), new simulation tools (Haller and Messer, 2019; Nelson et al., 2020) and several studies focusing on inferences from spatially continuous populations, specifically focusing on visualizing and inferring population structure from geo genetic maps (Petkova et al., 2016; Al-Asadi et al., 2019; Bradburd et al., 2016), and utilizing signals from covariance of allele frequencies and shared long identity-by-descent blocks (Ringbauer et al., 2017; Ralph and Coop, 2013; Barton et al., 2013).

In this thesis, I focused on the hybrid zone of *Antirrhinum majus*, which behaves as a single population, except for sharp clines at a few known genomic regions that mediate flower colouration. Extensive field sampling for over a decade revealed that snapdragons have a heterogeneous and patchy distribution with complex life history: they are perennial, have a seed bank and a complex self-incompatibility mating system. Furthermore, thorough sampling enabled us to generate a spatial pedigree consisting of $\sim$ 2300 trios (Field et. al in prep), which was utilized to obtain an empirical dispersal distribution by combining with genotype information far from the cline centre. Thus, we have a unique system in which the observed population density and dispersal are known, together with genomic regions influencing flower colouration; these are marked by 91 'neutral' and 6 'selected' SNPs. Thus, snapdragons provide a powerful test bed to understand the effects of realistic demography on genetic variation and to validate existing methods.

By analysing the effects of density and dispersal in snapdragons, we conclude that demography plays a key role in structuring genetic variation. Firstly, examining 'neutral' genetic variation demonstrated that isolation by distance (IBD) and distribution of heterozygosity in snapdragons are shaped by their heterogeneous population density and leptokurtic dispersal (Ch.4). Moreover, the empirical (leptokurtic) dispersal distribution by itself, while neglecting density variation across the population does not explain IBD and underestimates variance in inbreeding. Secondly, the sharp stepped clines and patterns of associations across the hybrid zone can only be explained by inclusion of inferred dispersal, with a small fraction of long-range migrants. Both these conclusions were obtained from two kinds of simulations: the neutral simulations used a spatially explicit model, that simulated the population conditioned on known locations of sampled individuals. This generated a spatial pedigree and genotype for 91 unlinked loci which was then fit to the empirical F_{ST} to infer N_e. The clinal simulations however relied on a stepping stone

model but included the effects of known dispersal, and obtained selection estimates by fitting simulated and observed allele frequencies. Together, both studies showcase how inferences can be biased when methods with oversimplified assumptions are applied to natural systems.

While we present a general simulation framework (in Ch.4) that can be applied to any system with known population density, it comes with limitations. Firstly, we were unable to test isolation by distance (IBD) patterns in snapdragons due to lack of theory for heterogeneous populations. The latter requires extending Wright-Malecot theory on how relatedness changes across space (Wright, 1943; Malécot, 1948; Maruyama, 1972) in order to account for effects of non-Gaussian dispersal (as done in the recent study by Smith and Weissman (2023)) and non-uniform density (Barton et al., 2010; Etheridge et al., 2023). Secondly, the simulation excludes the effect of seed bank and overlapping generations. While these effects can be included into our framework without much difficulty, discerning the interplay of seed bank and patchy population structure is more challenging, since relatedness is now a function of both time and space (Duforet-Frebourg and Slatkin, 2016). Thirdly, we condition the simulation on known spatial structure in a rather simple manner, thereby raising the question of whether and how we can track extinction and recolonization of patches over time, and how correlated movements of genes can be detected. Finally, including the effect of clinal selection in the spatial pedigree together with 2D heterogeneous density is challenging, both theoretically and computationally (Julseth, 2023). Further work is required to answer whether inferences from traditional cline analysis are robust to heterogeneous population density.

Shedding new light on hybrid zones

Hybrid zones are common in nature and act as natural laboratories to study how evolutionary forces maintain them. Allele frequencies at divergent trait loci are often studied as geographic clines giving estimates of selection and dispersal (Barton and Hewitt, 1985). Moreover, the stepped cline shape is usually thought to signal a barrier to gene flow between the two subspecies (Barton and Gale, 1993). We go beyond this 'traditional' cline analysis in two ways. Firstly, we showed that in the snapdragon hybrid zone, where dispersal is long-tailed selection is strong, inferences based on the diffusion approximation are unreliable. Secondly, unlike the expectation when genes diffuse, correlations between unlinked loci (LD) are found to be positive throughout the hybrid zone, not just at the cline centre. We showed that long-range dispersal, together with selection at a few flower colour loci explains the stepped cline shape and the associations, with the former contributing primarily to the step. Below, I discuss broader conclusions for the study of hybrid zones.

Understanding the causes of the stepped clines has important implications for hybrid zone

studies. When multiple loci influence the trait under selection, strong LD in the centre may generate stepped cline shape indicating a genetic barrier to gene flow (Szymura and Barton, 1986; Macholán et al., 2007). This is usually inferred from cline fits as the parameter B, the strength of the barrier to gene flow. As mentioned in Chapter 2, this quantifies the strength of RI in spatially continuous populations. However, B is confounded when additional forces are also in play and may not indicate the reduction in gene exchange between hybridising populations. For example, a physical barrier to gene flow can generate stepped clines (Jackson, 1992), which can be amplified when combined with a genetic barrier. The step due to the physical barrier can be calculated (Ch.2) and accounted for to get the strength of RI solely due to the genetic barrier. Additionally, long-range dispersal can generate stepped clines even when a single locus is under selection (as in the snapdragon hybrid zone). When the step is due to both dispersal and LD due to multilocus selection, B inferred from cline fits overestimates barrier to gene exchange. On the contrary, long-distance dispersal increases gene flow between populations by directly bringing in foreign genotypes to the alternate genetic background. In such scenarios, disentangling different factors and their contribution to the stepped shape is crucial. In these scenarios we show that the genetic barrier to gene flow can be quantified either as selection experienced by any trait locus due to its effect from all other trait loci (i.e indirect selection due to LD) or estimated from multilocus cline simulations with appropriate gaussian dispersal, thereby excluding the effects of long-range dispersal (see Ch.5).

Apart from the above-mentioned scenarios, tight linkage to a causal locus can also generate stepped clines. The strength of barrier to neutral gene flow will now depend on the rate of recombination between neutral and selected sites; the closer it is to the causal mutation, the stronger the reduction in gene exchange it experiences (Barton and Bengtsson, 1986; Barton, 1983; Tavares et al., 2018). In chapter 5, we focused on a few focal genes influencing flower colouration. While the expected cline shape for a locus under direct selection is sigmoid, we find supporting evidence to conclude that stepped clines indeed form at these causal loci due to long-range migration. However, the current genotyping (KASP SNPs) confidently identifies markers linked to causal mutation only for some genes (like Rosea and Eluta). Thus, to see how linkage to the causal mutation affects cline shape, we next plan to fit clines along the genome. We expect a pattern with sharp stepped clines for those genes in tight linkage to the causal mutation and shallower clines at increasing distance from it. This allows us to quantify how RI varies along the genome by finding how B changes with distance to the focal trait locus (see Ch.2). This can also be compared to F_{ST} scans to further distinguish 'islands of divergence' from false positives (Cruickshank and Hahn, 2014; Westram et al., 2018). Additionally, we now have whole genome sequences of samples from the snapdragon hybrid zone (obtained using the haplotagging technique; Meier et al. (2021)) thereby giving more genomic markers with better resolution. Comparing clines between different sequencing methods (SNP vs whole

genome) can help us dissect more focal genes and see if our current estimates are robust.

While our study gives new insights on hybrid zones, getting a complete picture of the genotype-phenotype map is non-trivial even in traits with relatively simple genetic basis (mediated through a few major effect loci). This will require further work on including the effects of dominance and epistasis, based on how flower colouration depends on the different genetic backgrounds. One of the factors that was essential to our conclusion was the detailed spatial and temporal sampling. This gave use the power to detect rare long-range migrants and infer the full dispersal distribution. While this might often not be possible, it highlights the importance of sampling scheme on the accuracy of inferences. Additionally, future work on LD based signals in the tails of clines might provide new ways to infer long range dispersal. I also emphasize the need to carefully examine whether assumptions of cline analysis are met before applying them to natural systems and combine them with simulations to validate inferences.

Feedback between theory, simulations and data

Evolutionary genetics uses a broad range of techniques ranging from field work, lab experiments, data analysis and modelling. With advances in sequencing technology, we now have access to large genomic datasets together with records of life history or demographic features. To best utilize these datasets, we need to identify gaps in current methods, develop new theoretical predictions, which can then be applied on genetic datasets (experimental or empirical) and validated using simulations. This constant feedback between theory and data is essential and valuable. In this thesis, I have combined some of these approaches. In chapter 3, we validated our theoretical predictions for divergence at a trait locus against simulations. Besides showing that our predictions are accurate, simulations also showed that the sexual selection and not assortment per se promotes speciation in our model (see Discussion, Ch.3). Similarly, chapters 4 and 5 highlighted the need to go beyond existing theory, and suggested new directions to improve inferences by incorporating either realistic density and/or dispersal. Together, these show that careful analysis of simulation-based approaches is a powerful tool to get valuable insights into questions of interest.

Measures of genetic variation from polygenic traits

Genetic variation gives valuable insights into the mechanism and strength of processes that shaped them over timescales ranging from a few generations to longer evolutionary timescales. Evolutionary biologists are therefore interested in devising new measures of genetic variation and their application to data. Some of these like F-statistics, π, site frequency spectrum, LD have been widely used and we have theoretical expectations for

these under "null models" that include drift, population structure, etc. In this thesis, we extend these measures and suggest possible directions for future studies.

One main aspect of this thesis specifically in light of speciation was the consideration of polygenic traits and utilizing signals from LD (Ch. 2,3,5). While traits are commonly influenced by multiple genes, such traits are not widely studied in the speciation literature. A key feature in the study of polygenic traits is to examine when trait loci evolve independently vs when sets of loci are coupled due to the effects of LD, making them behave like a single unit (Feder et al., 2012; Barton and De Cara, 2009; Butlin and Smadja, 2018). In the latter case, allele frequencies at a single trait locus are influenced by all other trait loci, so that one needs to examine how alleles evolve in different genetic backgrounds, thus making it more challenging to study. In this regard, effective migration rate is a well-defined yet pragmatic measure, that encapsulates the effect of LD between all trait loci on the focal locus with just a single quantity. We now have analytical predictions for m_e for scenarios involving multiple habitats, drift (Sachdeva, 2022), dominance (Zwaenepoel et al., 2024), sex chromosomes (Fraïsse and Sachdeva, 2021), while other studies aim to infer m_e from demographic models (Aeschbacher et al., 2017; Laetsch et al., 2023; Fraïsse et al., 2021; Rougemont et al., 2017).

In this thesis, we extend this notion of m_e to a polygenic magic trait to study the effect of both viability selection, assortment and sexual selection on barriers to gene flow and show how/when it can be estimated from genetic data. We demonstrate its utility by showing that predictions for m_e and divergence depend on measurable components of fitness of migrants (or F1 hybrids, etc), thereby merging phenotypic and genetic divergence even in complex scenarios when multiple barriers are in play. We further encourage researchers to utilize m_e based approaches while analysing polygenic selection. However, current analytical predictions for effective migration rate assume weak migration and a natural question is whether we can predict m_e when the population consists of a variety of recombinants (such as F2's, etc.). One possible future direction is how one can extend this notion to spatially continuous populations, for example hybrid zones, where B is the appropriate measure (see Ch.2). However, when long-distance migration is rare, we can still estimate effective migration rate in the tails of the cline, using either estimates of fitness directly inferred from the pedigree or indirectly from simulations (see Ch.5). We can further ask how this contributes to gene flow relative to diffusion of genes across the hybrid zone.

A second conclusion from this thesis is how (co)variances can be used to infer quantities of interest. Specifically, we found that distribution of heterozygosity (from neutral sites) in snapdragons had higher variance than expcted in a panmictic population, suggesting a small fraction of mating with relatives; we used signals from covariance in heterozygous state between loci, i,e identity disequilibrium (g_2), to quantify variance in inbreeding.

However, since evolutionary processes are highly stochastic, it is challenging to distinguish actual signal from noise due to the effect of drift. This is shown using simulations in chapter 4, where measures of g_2 were highly variable between replicate pedigrees and between replicate genotypes within a single pedigree. In this regard, Buffalo and Coop (2020) utilizes signals of temporal covariance in allele frequency changes across generations to distinguish signals of selection from drift. These suggest that measures based on covariance in allele frequency might be a promising future direction, for example to detect correlated movements in spatially continuous populations, or to understand local fluctuations in allele frequency, effects of strong assortment. This will be challenging in natural populations with migration and overlapping generations.

To conclude, this thesis takes a step forward by studying the effects of realistic density, dispersal and assortative mating on genetic variation in polygenic traits, specifically in the context of speciation. We show that heterogeneous population density and dispersal structures genetic variation in space and emphasize the need to account for them in inference methods. We also suggest ways to analyse the effects of polygenic traits on RI either based on effective migration rates, or by quantifying indirect selection or barriers to gene flow. We highlight when and where care should be taken when applying existing theory and inferences on empirical data. Together, this study seeks to bridge the gap between theoretical and empirical research, and incorporates the effect of demography in speciation studies.

References

Aeschbacher, S., Selby, J. P., Willis, J. H., and Coop, G. (2017). Population-genomic inference of the strength and timing of selection against gene flow. *Proceedings of the National Academy of Sciences of the United States of America*, 114:7061–7066.

Aguillon, S. M., Fitzpatrick, J. W., Bowman, R., Schoech, S. J., Clark, A. G., Chen, N., and Reed, F. A. (2017). Deconstructing isolation-by-distance: the genomic consequences of limited dispersal. *PLOS Genetics*, 13:e1006911.

Al-Asadi, H., Petkova, D., Stephens, M., and Novembre, J. (2019). Estimating recent migration and population-size surfaces. *PLOS Genetics*, 15:e1007908.

Barton, N. H. (1983). Multilocus clines. *Evolution*, 37:454–471.

Barton, N. H. and Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57:357–376.

Barton, N. H. and De Cara, M. A. R. (2009). The evolution of strong reproductive isolation. *Evolution*, 63(5):1171–1190.

Barton, N. H., Etheridge, A. M., Kelleher, J., and Véber, A. (2013). Inference in two dimensions: allele frequencies versus lengths of shared sequence blocks. *Theoretical population biology*, 87:105–119.

Barton, N. H. and Gale, K. S. (1993). Genetic analysis of hybrid zones. In Harrison, R. G., editor, *Hybrid Zones and the Evolutionary Process*, pages 13–45. Oxford University Press, Oxford.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, 16:113–148.

Barton, N. H., Kelleher, J., and Etheridge, A. M. (2010). A new model for extinction and recolonization in two dimensions: Quantifying phylogeography. *Evolution*, 64:2701–2715.

Bradburd, G. S., Ralph, P. L., and Coop, G. M. (2016). A spatial framework for understanding population structure and admixture. *PLOS Genetics*, 12:e1005703.

Buffalo, V. and Coop, G. (2020). Estimating the genome-wide contribution of selection to temporal allele frequency change. *Proceedings of the National Academy of Sciences of the United States of America*, 117(34):20672–20680.

Butlin, R. K. and Smadja, C. M. (2018). Coupling, reinforcement, and speciation. *The American Naturalist*, 191(2):155–172.

Cruickshank, T. E. and Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23:3133–3157.

Duforet-Frebourg, N. and Slatkin, M. (2016). Isolation-by-distance-and-time in a stepping-stone model. *Theoretical Population Biology*, 108:24–35.

Epperson, B. (2003). *Geographical Genetics*. Princeton University Press, Princeton, NJ.

Etheridge, A. M., Letter, I., Kurtz, T. G., Ralph, P. L., and Lung, T. T. H. (2023). Looking forwards and backwards: Dynamics and genealogies of locally regulated populations.

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

Fraïsse, C., Popovic, I., Mazoyer, C., Spataro, B., Delmotte, S., Romiguier, J., Loire, E., Simon, A., Galtier, N., Duret, L., Bierne, N., Vekemans, X., and Roux, C. (2021). Dils: Demographic inferences with linked selection by using abc. *Molecular Ecology Resources*, 21:2629–2644.

Fraïsse, C. and Sachdeva, H. (2021). The rates of introgression and barriers to genetic exchange between hybridizing species: Sex chromosomes vs autosomes. *Genetics*, 217(2):iyaa025.

Haller, B. C. and Messer, P. W. (2019). Slim 3: Forward genetic simulations beyond the wright–fisher model. *Molecular Biology and Evolution*, 36(3):632–637.

Jackson, K. S. (1992). *The Population Dynamics of a Hybrid Zone in the Alpine Grasshopper Podisma pedestris: An Ecological and Genetic Investigation*. PhD thesis, University of London, University College London.

Julseth, M. (2023). The effect of local population structure on genetic variation at selected loci in the *A. majus* hybrid zone. Master's thesis, Institute of Science and Technology Austria. https://doi.org/10.15479/at:ista:12800.

Laetsch, D. R., Bisschop, G., Martin, S. H., Aeschbacher, S., Setter, D., and Lohse, K. (2023). Demographically explicit scans for barriers to gene flow using gimble. *PLOS Genetics*, 19(10):e1010999.

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., et al. (2015). The fine-scale genetic structure of the british population. *Nature*, 519:309–314.

Macholán, M., Munclinger, P., Šugerková, M., Dufková, P., Bímová, B. V., Božíková, E., Zima, J., and Piálek, J. (2007). Genetic analysis of autosomal and x-linked markers across a mouse hybrid zone. *Evolution*, 61:746–771.

Malécot, G. (1948). *Les Mathématiques de l'hérédité*. Masson, Paris.

Maruyama, T. (1972). Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics*, 70:639–651.

Meier, J. I., Salazar, P. A., Kučka, M., Davies, R. W., Dréau, A., Aldás, I., and Chan, Y. F. (2021). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences*, 118(25):e2015005118.

Nelson, D., Kelleher, J., Ragsdale, A. P., Moreau, C., McVean, G., and Gravel, S. (2020). Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS Genetics*, 16(5):e1008619.

Petkova, D., Novembre, J., and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, 48:94–100.

Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across europe. *PLOS Biology*, 11:e1001555.

Ringbauer, H., Coop, G., and Barton, N. H. (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205:1335–1351.

Rougemont, Q., Gagnaire, P.-A., Perrier, C., Genthon, C., Besnard, A.-L., Launey, S., and Evanno, G. (2017). Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. *Molecular Ecology*, 26:142–162.

Rousset, F. (2004). *Genetic Structure and Selection in Subdivided Populations*, volume 40. Princeton University Press, Princeton, NJ.

Sachdeva, H. (2022). Reproductive isolation via polygenic local adaptation in sub-divided populations: Effect of linkage disequilibria and drift. *PLoS Genetics*, 18(9):e1010297.

Smith, T. B. and Weissman, D. B. (2023). Isolation by distance in populations with power-law dispersal. *G3: Genes, Genomes, Genetics*, 13:jkad023.

Szymura, J. M. and 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*, 40:1141–1159.

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., and 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.

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

Wright, S. (1943). Isolation by distance. *Genetics*, 28:114–138.

Zwaenepoel, A., Sachdeva, H., and Fraïsse, C. (2024). The genetic architecture of polygenic local adaptation and its role in shaping barriers to gene flow. *Genetics*, page iyae140.

Appendix A

SUPPLEMENTARY INFORMATION for What is reproductive isolation?

A.1 Simulation results

Flow into a single deme

We simulate a mainland-island model considering a diploid population with either two (one selected and one neutral) or three (two selected and one neutral) loci and two alleles (namely 1 or 2) at each locus. We assume that the haplotype 11 (or 111) is fixed in the island and the alternative haplotype 22 (or 222) is initially fixed in the mainland. Every generation a small fraction of migrants enter the island from the mainland with rate m. The selection coefficient, s, is constant for each selected locus, with the diploid genotype fitnesses following heterozygote disadvantage in the ratio $1 : 1 - s : 1 - 2s$ and with multiplicative fitness. The population is simulated until an equilibrium is reached. The haplotype frequencies are followed and the strength of reproductive isolation (RI) is calculated for the neutral locus. We also consider different recombination rates, r, between the neutral and selected loci, so as to calculate how the strength of RI varies along the genome.

A specific example is shown first, with one selected and one neutral locus with $m = 0.01, s = 0.1$ and $r = 0.05$. Fig. A.1 shows the allele frequency of the selected and neutral locus as a function of time (on a log scale). The allele frequency at the neutral locus shows a steady state decrease, the slope of which gives the effective migration rate (m_e). This is found to be -0.0031. Since the allele frequency decreases over time, the slope is negative, but we use only the absolute value to calculate RI as $1 - m_e/m$. Furthermore, m_e is

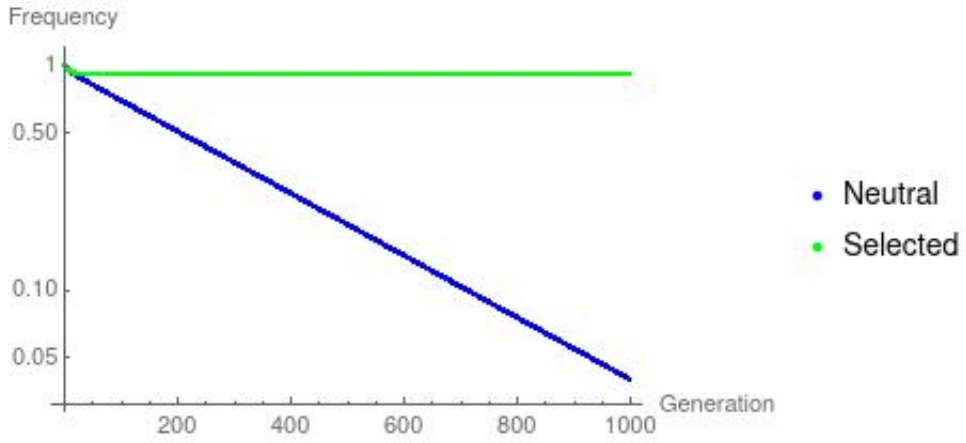


Figure A.1: The allele frequency over time of the neutral (blue) and the selected (green) loci from a mainland-island model with $s = 0.1, r = 0.05$ and$m = 0.01$ plotted on a log scale. The rate of change of the allele frequency of the neutral locus is 0.0031.

simply the allele frequency changes in a generation (Δu), relative to the allele frequency difference between demes (δu). A plot of Δu vs. δu gives a straight line, suggesting that m_e is constant over time (see Fig. 2.5A in chapter 2).

Next, we simulate the population for different recombination rate between the neutral and selected loci with $s = 0.1$ and $m = 0.01$ throughout and calculate strength of RI for each case. The strength of RI is also calculated for the case with two selected and one neutral locus, again with $s = 0.1$ for both loci and $m = 0.01$. We consider the neutral loci to be at either ends of the selected loci or between them to see how RI varies along the genome. We also verify the results from the simulation to the analytical result of RI (see chapter 2). These results are shown in Fig. 2.3 of chapter 2.

One-dimensional hybrid zone

Here, we assume a 1D stepping stone model with 101 demes and nearest neighbour migration with rate $m = 0.5$. The haplotype 11 (or 111) is fixed in the first 50 demes and the alternative haplotype 22 (or 222) is fixed in the remaining demes. At the selected site, the alternative alleles are selected in the first 50 and the last 51 demes. As before we assume heterozygote disadvantage and multiplicative fitness. For a simplified calculation of RI, we keep the haplotype frequencies always fixed for the first and last demes (111 is fixed in deme 1 and 222 in deme 101). The population is simulated until neutral allele shows a stable pattern.

Fig. A.2 shows allele frequency over space for different generation for the case with $s = 0.1$ and $r = 0.05$ between neutral and selected locus (in blue). It takes about 5000 to 10000 generations for the allele frequency of the neutral locus to stabilize. The barrier strength is then calculated as $B = \Delta p/p'$, where the numerator is the step or the difference in the allele frequency between the tails and denominator is the rate of change of allele frequency

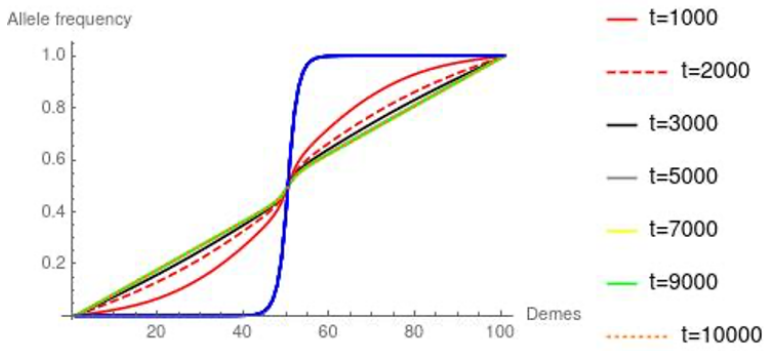


Figure A.2: Allele frequency of the neutral locus for different generations over space

in the tails. To calculate this from the simulations, we need to define a cut-off for the tails. We use demes 1 to 40 at one end and 61 to 101 at the other, and use regression to find their gradients, the mean of which gives p'. The predicted allele frequencies at deme 50 is calculated from both tails, the difference between which gives Δp. A barrier strength of 8 is obtained for this case. As before, we now calculate the strength of the barrier, B, for different positions of the neutral loci along the genome, with either one or two selected loci. This is shown in Fig. 2.4 in chapter 2.

Next, the allele frequencies at the first and last deme is let to vary to represent a more plausible biological scenario. 201 demes are used instead of a 100 along with stronger selection of $s = 0.2$ and lower recombination rate ($r = 0.01$). The hybrid zone is simulated for 30000 generation. The allele frequency changes over space along with the associated barrier strength is shown in Fig. 2.5B in chapter 2.

Hybrid zone with a physical barrier

Here, we consider the 1D stepping stone model as before, with an additional assumption of a physical barrier between demes 50 and 51. The migration rate is $\epsilon/2$ between the demes, where $\epsilon = 0.1$. As before, we calculate RI along the genome for the following cases 1. with just a physical barrier as mentioned above with no selection (i.e no genetic barrier) shown in grey
2. with only a genetic barrier (in black)
3. with both physical barrier and selection (in red)
4. when we add the effects of physical (1) and genetic barrier (2) (in blue)
and compare RI among them. This is shown is figure 2.6 in chapter 2. We see that the barrier strength with both physical and genetic barrier is greater than their effects combined.

A.2 Theoretical predictions

Flow into a single deme

Unlinked loci

Suppose that all loci are unlinked, and that migrants backcross into the recipient population; there is no selfing. The expected number of F1 offspring of a migrant individual, relative to a resident, we denote $\bar{W}_0$, and the expected relative number of offspring of the F1 individual is $\bar{W}_1$; similarly, the relative fitness of the k'th backcross generation is $\bar{W}_k$ ($k \geq 2$). If the population is at equilibrium, then each resident is expected to produce two offspring, and with Mendelian segregation, each gene is expected to transmit one copy. Therefore, the relative contribution of an incoming migrant is $\bar{W}_0\bar{W}_1\bar{W}_2\ldots$; each component of this product is a relative fitness that includes viability, mating success, and fecundity. There may be arbitrary epistasis and dominance: all that matters is the mean fitness of each backcross generation. Thus:

$$\mathrm{RI}_{2D} = 1 - \frac{m_e}{m} = 1 - \bar{W}_0\bar{W}_1 \prod_{k=2}^{\infty} \bar{W}_k \tag{A.1}$$

The relative fitnesses in the first and the F1 generations will be idiosyncratic: for example, there may be heterosis, such that $\bar{W}_1>1$. In the short term, Eq. A.1 still applies, and may give $m_e > m$, $\mathrm{RI}_d < 0$. If heterosis is due to divergent frequencies of recessive deleterious alleles, these are expected to equilibrate, and heterosis would dissipate. However, if other processes maintain different frequencies of deleterious recessives, or if there is a continual introduction of beneficial alleles via immigrants, then heterosis might persist, weakening RI.

Approximating later backcross generations In the k'th backcross generation, we expect $2^{-(k-1)}$ heterozygous loci, and so we expect $\bar{W}_k$ to approach 1 as introgressing loci are diluted out. If fitnesses are multiplicative, then we can write $\bar{W}_k = \exp[-S_k]$, with $S_k = S_2 2^{-(k-2)}$, which approaches zero as k increases. Therefore:

$$\mathrm{RI}_{2D} = 1 - \frac{m_e}{m} = 1 - \bar{W}_0\bar{W}_1 \prod_{k=2}^{\infty} \bar{W}_k = \bar{W}_0\bar{W}_1 \exp\left[-\sum_{k=2}^{\infty} S_2 2^{-(k-2)}\right] = \bar{W}_0\bar{W}_1\bar{W}_2^2 \tag{A.2}$$

In this simplest case, therefore, we can estimate the effective migration rate, a measure of the strength of reproductive isolation, from the mean fitnesses of successive backcross generations. It is determined primarily by the first few generations: the influx of genes is reduced by the same factor in the first backcross generation as in all subsequent backcrosses. Epistasis (i.e., a deviation from multiplicative effects) may increase or decrease the rate of gene flow. If even a small fraction of foreign alleles reduces fitness (i.e., $S_k > S_2 2^{-(k-2)}$), then gene flow would be reduced by more - and in principle, to zero if there were an

indefinitely large number of selected loci. However, this seems an extreme scenario. To the extent that loci are unlinked, and introgression is via backcrossing, the mean fitnesses of successive backcross generations give a simple estimate of the strength of isolation.

Effects of purging Eq. A.1 gives a simple formula for the effective rate of gene flow under continued backcrossing: $\frac{m_e}{m} = \bar{w}_0\bar{w}_1\ldots$ However, it is important to realise that these are the mean fitnesses of the *actual* backcrosses, which will have been subject to selection. Here, we examine how our estimate of m_e is affected if we use the mean fitnesses of randomly generated backcrosses, generated with no selection, as compared with the true situation, where selection reduces the proportion of introgressing alleles. We take a simple example, where the mean fitness, w_j, defined relative to the native genotype, depends on the number of heterozygous loci, j. With multiplicative fitnesses, $w_j = (1-s)^j$, for example. We have:

$$f_{0,j} = \delta_{j,n} f_{t+1,j} = \sum_{k=0}^{n} w_j B_{j|k} f_{t,k} \tag{A.3}$$

where $f_{t,j}$is the distribution of the # of heterozygous loci in generation t, $B_{j,k}$ is the binomial distribution, and $\delta_{j,n} = 0\text{for}j < n$, and 1 for $j = n$, where n is the number of loci. Note that there is no normalisation, since we are counting the actual number of backcross individuals.

Assume that $\bar{W}_0 = 1$, and $\bar{W}_1 = w_n$ is the backcross fitness. From Eq. A.1, the net gene flow is the product :

$$\frac{m_e}{m} = w_n \left(\sum_{j=0}^{n} w_j B_{j|n}\right)\left(\sum_{k=0}^{n}\sum_{i=0}^{k} w_i B_{i|k} w_k B_{k|n}\right)\ldots \tag{A.4}$$

Fig. A.3a shows how, over successive backcross generations, the number of introgressed alleles (blue) decreases slightly faster than in the absence of purging (red). Fig. A.3b shows that the selected distributions (blue) have slightly higher mean fitness; for these parameters (s=0.1 on 20 loci), the net gene flow is $m_e/\,m$ =0.021 rather than $m_e/\,m$ =0.016 calculated from raw backcross fitnesses. With $s = 0.2$, the net gene flow is about 2.7 times higher than predicted. Table S2.1 makes this comparison for $n = 20$ and 100, with selection being scaled such that the total selection $n\,s$ is comparable between the two tables. Purging does alleviate RI, but this effect is only apparent which the total selection is strong (implying low F1 fitness), and is less marked when selection is spread over more loci.

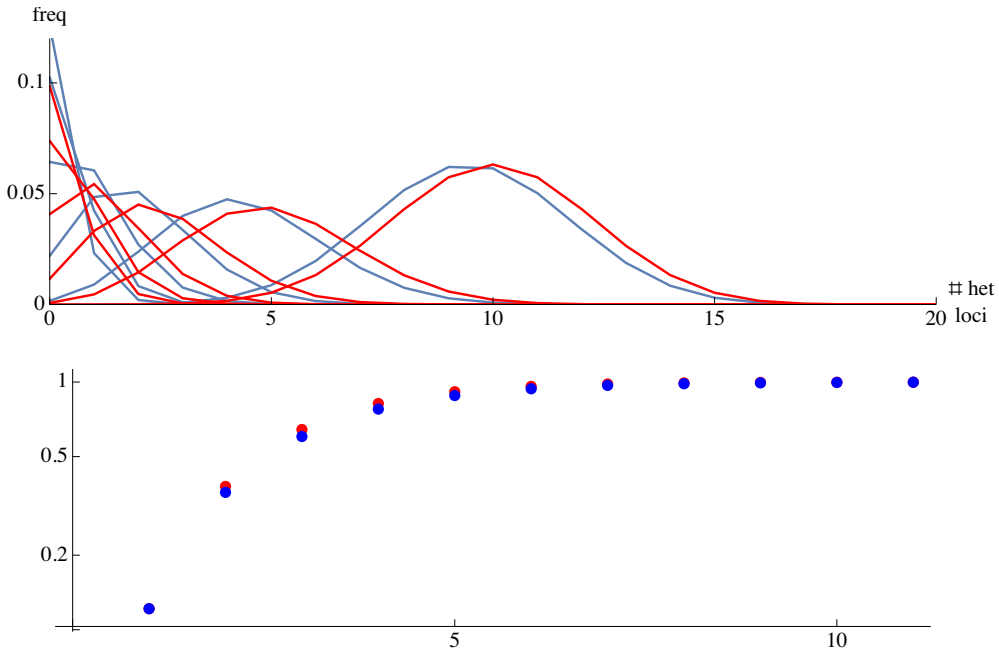


Figure A.3: a) Distribution of the number of heterozygous loci selected (red) and unselected (blue) distributions, assuming multiplicative selection $s = 0.1$ on $n = 20$ loci. b) Mean fitnesses over successive generations ($s = 0.1, n = 20$), for the selected (red) and unselected (blue) distributions.

s	(1-s)^{20}	m_e/m, with purging	m_e/m, with no purging
0.	1.	1.	1.
0.05	0.358486	0.139716	0.130897
0.1	0.121577	0.0206073	0.0159296
0.15	0.0387595	0.00317261	0.00178946
0.2	0.0115292	0.000504032	0.000184015
0.25	0.00317121	0.0000816492	0.0000171517
0.3	0.000797923	0.000013313	$1.43 * 10^{-6}$
0.35	0.000181245	$2.15 * 10^{-6}$	$1.06 * 10^{-7}$
0.4	0.0000365616	$3.40 * 10^{-7}$	$6.75 * 10^{-9}$

s	(1-s)^{20}	m_e/m, with purging	m_e/m, with no purging
0.	1.	1.	1.
0.01	0.366032	0.136348	0.134562
0.02	0.13262	0.0188277	0.017863
0.03	0.0475525	0.00263174	0.00233879
0.04	0.0168703	0.000372209	0.000301944
0.05	0.00592053	0.0000532393	0.0000384282
0.06	0.00205487	$7.7 * 10^{-6}$	$4.82 * 10^{-6}$
0.07	0.000705172	$1.12 * 10^{-6}$	$5.96 * 10^{-7}$
0.08	0.000239212	$1.66 * 10^{-7}$	$7.25 * 10^{-8}$
0.09	0.0000801935	$2.47 * 10^{-8}$	$8.67 * 10^{-9}$
0.1	0.0000265614	$3.72 * 10^{-9}$	$1.03 * 10^{-9}$

Table A.1: Fitnesses of the F1 $(1-s)^n$, $m_e/\,m$ with purging, and $m_e/\,m$ without accounting for purging. a) $n = 20$ loci, b) $n = 100$ loci.

Linked loci

We use Eq. A6 from Barton and Bengtsson (1986) to find the effective migration rate, m_e, and hence RI_{2d}, where there are multiple selected loci on either side of the focal neutral locus. These can readily be calculated for any given set of selection coefficients and recombination rates, even including epistasis; however, we could find no simple explicit formula for the general case. Here, we consider the simple symmetric case, where selected loci are evenly spaced, with recombination rate r between adjacent loci, selection is s for all loci, and their effects on fitness multiply. Suppose that there are J loci to the left, and K to the right. The neutral locus is $\alpha\, r$ from the selected locus on the left, and $(1-\alpha)$r from that on the right; we let $\theta = s/r$. Then:

$$R)_{2d} = 1 = \frac{m_e}{m} = 1 - \prod_{j=0}^{J-1} \frac{\alpha + j(1+\theta)}{\alpha + j(1+\theta) + \theta} \prod_{k=0}^{K-1} \frac{(1-\alpha) + k(1+\theta)}{(1-\alpha) + k(1+\theta) + \theta} \tag{A.5}$$

This equation contains the product of the effects of each selected locus on the focal neutral locus, with RI_{2d} increasing with the number of selected loci but the effect of each individual selected locus decreasing with distance. It implies that if selection is strong relative to recombination ($\theta >> 1$), the barrier in-between sets of selected loci will be much stronger than the barrier due to selected loci that are only on one or on the other side. The strength of the barrier will primarily decrease with $r^2\alpha(1-\alpha)$, i.e. it decreases the further the selected loci are apart and (as in the previous example) the further away

the neutral locus is from the nearest selected locus. The square appears in this expression because introgression requires two recombination events to move the neutral allele onto the new background.

Relative contributions of unlinked vs linked loci

Consider a very large number of loci, spread over a long genetic map, of total length R. To take the simplest case, assume multiplicative selection s on each of n loci, which are evenly spaced on the genetic map; there are J loci to the left of the focal neutral locus, and K to the right. The total selection is $S = n\,s$, the fitness of an F1 being e^{-S},and the ratio of selection to recombination is $\theta = s/r \sim S/R$. If we ignore linkage, gene flow is reduced by a factor $m_e/\,m = e^{-2S}$. We wish to compare this with the reduction due to linked loci, which is given by Eq. A.2. Using Stirling's approximation, we find that for large numbers of loci:

$$\frac{m_e}{m} = (JK)^{-\theta/(1+\theta)} f[\alpha, \theta] \tag{A.6}$$

where f[α,θ] depends on α (the position of the neutral locus relative to the selected loci on either side) and θ. Taking J, K to be proportional to the number of loci, n, we have $m_e/m \sim n^{-2\theta/(1+\theta)}$, as in chapter 2.

Unlinked loci reduce gene flow by e^{-2S} (Eq. A.3), setting a lower bound on the barrier strength. Linked loci reduce gene flow by a factor $\exp(C - 2\log(n)\theta/(1+\theta))$, where $C \sim \log(f)$ is a constant that depends on the relative positions of the genes, but not on the number of loci. Unless hybrids are extremely unfit ($S > R$), we expect that $\theta = S/R$ is small. Then, assuming $\theta << 1$, unlinked loci will make a stronger contribution to RI if $S > \theta \log(n)$, which is equivalent to $R > \log(n)$. Thus, with a long genetic map, global RI due to unlinked loci is more important than local RI due to linked loci, even when very many loci are involved.

Effect of a physical barrier

Barton (1986) showed that the barrier due to a local reduction in density or dispersal is:

$$B_{\text{phys}} = \int \left(\frac{\rho_0^2 \sigma_0^2}{\rho^2 \sigma^2} - 1 \right) dx \tag{A.7}$$

Here, $\rho(x), \sigma^2(x)$ are the density and dispersal rate (i.e. variance of parent-offspring distance) across a one -dimensional habitat; ρ_0, σ_0^2 are the values outside the barrier, assumed the same in the symmetric case.

Relation between m_e and B in one dimension

Consider two regions, each of size x, on either side of a strong barrier ($B >> \sigma$) at zero. (Throughout, we use lower case x to denote actual distance, and capital $X = x \Big/ \left(\sigma\sqrt{t}\right)$ the scaled distance). Two divergent populations, fixed for different alleles at a neutral locus, meet, and the allele frequency difference decreases as genes flow across the barrier. If we define the two populations as the regions on either side of the barrier, then the mean allele frequency decreases at a rate proportional to the flux of genes across the barrier, which in turn is proportional to the gradient immediately adjacent to the barrier. The rate of change of mean frequency in the region to the left is:

$$\frac{\partial \bar{p}_-}{\partial t} = \frac{1}{x}\int_{-X}^{0} \frac{\partial p}{\partial t} dy = \frac{\sigma^2}{2x}\frac{\partial p}{\partial y} \tag{A.8}$$

where we used the diffusion $\partial_t p = (\sigma^2/2)\,\partial^2 p/\partial x^2)$, and integrated; the gradient is evaluated next to the barrier at $y = 0$. The barrier is defined by $\Delta \mathrm{p}_B = B(\partial p/\partial y)$, where $\Delta \mathrm{p}_B$ is the difference in allele frequency immediately across the barrier. Therefore:

$$\frac{\partial \bar{p}_-}{\partial t} = \frac{\sigma^2}{2x}\frac{\Delta \mathrm{p}_B}{B} \tag{A.9}$$

Now, the effective migration rate is defined as $m_e = (\partial \bar{p}_-/\partial t)/\Delta \mathrm{p}_T = (\Delta \mathrm{p}_B/\Delta \mathrm{p}_T)\,\sigma^2/(2B\,x)$. For a considerable time ($\sim (x/\sigma)^2$ generations), the allele frequency will change near the barrier, so that $\Delta \mathrm{p}_B < \Delta \mathrm{p}_T$, and m_e will increase towards $\sigma^2/(2B\,x)$ over the long time needed for allele frequencies to become homogeneous on either side. During this period, m_e would vary, if we defined different regions; only after the two regions had become homogeneous would m_e approach a constant value, independent of the region that defines it.

The corresponding migration rate in the absence of a barrier, m,is less straightforward. The rate of change of mean allele frequency in the regions on either side is still proportional to the gradient at zero (Eq. A.6), because this determines the flux of allele frequencies in either direction. First, consider the initial period, before introgression has reached the boundaries ($\sigma^2 t << X$). The allele frequency at position $y$following secondary contact is then $P_t\left[y \Big/ \left(\sigma\sqrt{t}\right)\right]$, where P_t is the integral of a normal distribution with variance $\sigma^2 t$, from -∞ to y. The gradient of this at zero is just the value of the normal distribution at that point, $1 \Big/ \sqrt{2\pi\,\sigma^2 t}$. Therefore:

$$\frac{\partial \bar{p}_-}{\partial t} = \frac{1}{x}\int_{-X}^{0} \frac{\partial \tilde{P}}{\partial t} dy = \frac{\sigma^2}{2x}\frac{\partial p}{\partial y} = \frac{\sigma^2}{2x\sqrt{2\pi\,\sigma^2 t}} \tag{A.10}$$

During this initial period, $\Delta \mathrm{p}_T \sim 1$, and so m is equal to the above value, and decreases

with $\sqrt{t}$ - reflecting the decreasing gradient, and hence the decreasing flux of genes across the contact region. To understand the behaviour at later times, we need to account for the boundaries. There is no gene flow outside $\{-x, x\}$, and so the gradients at $\pm x$ are zero. The solution is just:

$$\tilde{P}[Y] = P[Y] + \sum_{k=1}^{\infty} \left(P\left[-2kX + (-1)^k Y\right] - P\left[-2k\,X - (-1)^k Y\right] \right)$$

which satisfies the boundary conditions; here, X, Y denote distances scaled by $\sigma\sqrt{t}$

Because there is no flow across the boundaries, the total mass within $\{-x, x\}$ is unaffected. However, the total mass within $\{-x, 0\}$ is now given by an alternating series. In unscaled units:

$$\bar{p} = \frac{1}{x}\int_{-x}^{0} \tilde{P}[y]\, dy = \frac{\sigma\sqrt{t}}{x}\int_{-X}^{0} \tilde{P}[Y]\, dX = \frac{\sigma\sqrt{t}}{x}\sum_{k=0}^{\infty}(-1)^k C[-2(k+1)X, -2kX]$$
$$\text{where } C[a,b] = \int_a^b P[Y]\, dY, X = \frac{x}{\sigma\sqrt{t}} \quad \text{(A.11)}$$

The rate of change of mean frequency can be expressed as the difference in flux across the boundaries, which in turn is proportional to the difference in gradients at the boundaries. The gradient at -X is zero, and so we just need the gradient at 0, which in turn is a sum of Gaussians.

$$\frac{\partial \bar{p}_-}{\partial t} = \frac{1}{x}\int_{-x}^{0} \frac{\partial \tilde{P}}{\partial t} dy = \frac{\sigma^2}{2x}\frac{\partial \tilde{P}[0]}{\partial y} = \frac{\sigma}{2x\sqrt{t}}\frac{\partial \tilde{P}[0]}{\partial Y} = \frac{\sigma}{2x\sqrt{t}}\sum_{k=-\infty}^{\infty}(-1)^k P'[-2k\,X]$$
$$= \frac{\sigma}{2x\sqrt{2\,\pi\,t}}\sum_{k=-\infty}^{\infty}(-1)^k e^{-2(k\,X)^2} = \frac{\sigma}{2x\sqrt{2\,\pi\,t}}\partial_4\left[0, e^{-2X^2}\right] \quad \text{(A.12)}$$

where ∂_4 is the elliptic theta function.

We now need to calculate $m_{e,N} = \partial_t p / (1 - 2\bar{p}) = \frac{1}{2}\partial_t \log(\Delta \mathrm{p})$; this is an effective migration rate for the *neutral* case, defined by treating the region $\{-x, 0\}$ as a "population". The left plot in Fig. A.4 shows this, for $x = 100\,\sigma$, and the middle plot shows Δp (red) and $\partial_t \bar{p}$ (blue), which is the gradient of the top curve. The right plot shows the actual clines at times 10, 100, 1000, 10^4. $m_{e,N}$ gradually declines, with $1/\sqrt{t}$, up until ~3000, when introgressing alleles reach the boundary. It then quickly approaches a limit $0.617/(x/\sigma)^2$ (blue line). Thus, once allele frequencies are almost homogeneous, the effective migration rate in the absence of a barrier can be defined. However, it depends on x in a *different* way than the effective migration rate defined in the absence of a barrier (which is proportional

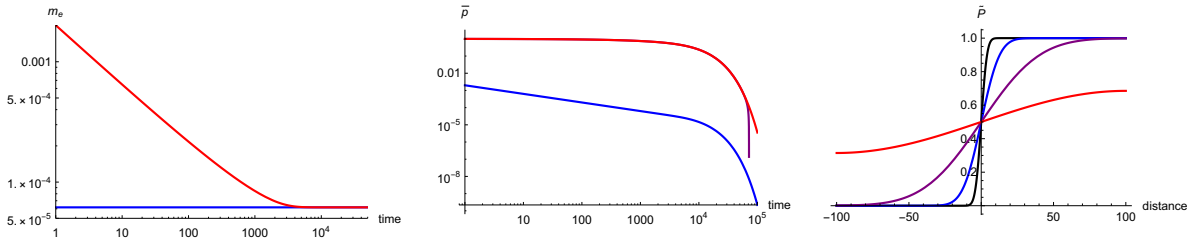


Figure A.4: Left: the effective migration rate, m_e, defined for the population between the boundary at $x = -100$, anda barrier at $x = 0$, plotted against generations since secondary contact, t. The blue line shows the limit for large t. Middle: The step, Δp (red) and the gradient, $\partial_t \bar{p}$ (blue). Right: Clines at $t = 10,\ 100, 1000, 10^4$.(black, . . . , red).

to $1/x$). Thus, RI $= 1 - m_e/m_{e,N}$ would depend on how one defines the "population", even in the limit of large times.

Detailed derivation The solution relies on the fact that any solution of the form $A\,P[C \pm y]$ satisfies the diffusion equation in the interior. We require 1/2 at X=0 and zero gradient at ±X. The first constraint can be satisfied by dividing by a constant, and the second, by summing a series of corrections of the form $P[-2k\,X \pm y]$ for integer k>=0. We start with an initial trial solution $P[y]$. This breaks the boundary conditions, because $P[\pm X] \neq 0$, and so we add two compensating terms, $P[-2X - y]$, $-P[-2X + y]$. These must in turn be compensated at the opposite boundary. The first term has gradient P[-3X] at +X, and is compensated by $+P[-4X + y]$, and the second term is compensated by $-P[-4X - y]$. Thus:

$$\begin{aligned}
\tilde{P}[y] &= P[y] + (P[-2X - y] - P[-2X + y]) + (P[-4X + y] + P[-6X - y] + P[-8X + y] \ldots) \\
&\quad - (+P[-4X - y] + P[-6X + y] + P[-8X - y] \ldots) \\
\tilde{P}'[y] &= P'[y] + (-P'[-2X - y] + P'[-4X + y] - P'[-6X - y] + P'[-8X + y] \ldots) \\
&\quad - (P'[-2X + y] - P'[-4X - y] + P'[-6X + y] - P'[-8X - y] \ldots) \quad \text{(A.13)}
\end{aligned}$$

As a check, the gradients at $\pm X$ are:

$$\begin{aligned}
\tilde{P}'[-X] &= P'[-X] + (-P'[-X] + P'[-3X] - P'[-5X] + P'[-7X] \ldots) \\
&\quad - (P'[-3X] - P'[-5X] + P'[-7X] - P'[-9X] \ldots) = 0 \\
\tilde{P}'[+X] &= P'[X] + (-P'[-3X] + P'[-3X] - P'[-7X] + P'[-7X] \ldots) \\
&\quad - (P'[-X] - P'[-5X] + P'[-5X] - P'[-9X] \ldots) = 0 \quad \text{(A.14)}
\end{aligned}$$

The total mass in {-X,0} is now $\{-2X, 0\}, \{-6X, -4X\}, \{-10X, -8X\}$...−($\{-4X, -2X\}$, $\{-8X, -6X\}$, ..)). The rate of change is necessarily proportional to the gradient at zero, which is Gaussian:

$$\tilde{P}'[0] = P'[0] - 2P'[-2X] + 2P'[-4X] - 2P'[-6X] + 2P'[-8X] = P'[0] + 2\sum_{k=1}^{\infty}(-1)^k P'[-2kX]$$

In unscaled units, therefore:

$$\frac{\partial}{\partial t}\int_{-X}^{0}\tilde{P}[y]\,dy = \frac{\sigma^2}{2}\frac{\partial \tilde{P}'[0]}{\partial x} = \frac{\sigma}{2\sqrt{t}}\sum_{k=1}^{\infty}(-1)^k\frac{\partial P[-2kX]}{\partial X}\Bigg)$$

where $P'[y] = \frac{e^{-y^2/\left(2\sigma^2 t\right)}}{\sqrt{2\pi\sigma^2 t}}$. Thus:

$$\frac{\partial}{\partial t}\int_{-X}^{0}\tilde{P}[y]\,dy = \frac{\sigma^2}{2}\frac{1}{\sqrt{2\pi t}}\sum_{k=0}^{\infty}(-1)^k a^{k^2} = \frac{\sigma}{2}\frac{1}{\sqrt{2\pi t}}\partial_4[0,a]$$

$$\text{where } a = e^{-(2X)^2/\left(2\sigma^2 t\right)} \text{and } \partial_4(0,a) \text{is the elliptic theta} \quad \text{(A.15)}$$

Ecogeographic isolation

Suppose that the two populations have contiguous ranges, which overlap in a narrow region along a one-dimensional continuum (Fig. 2.8c); within this region they remain distinct, and only a small proportion of genes m is exchanged between the populations per generation; this is an effective rate, which includes the obstacles to introgression that allow the populations to remain distinct in sympatry. Here, we derive the barrier to the flow of genes from population 1 into population 2, via interbreeding in the region of overlap.

We define the density of individuals in the two populations as $n_1(x), n_2(x)$ respectively. We assume that the density of population 1 is zero on the left, and approaches n_1^* on the right. Conversely, the density in population 2 approaches a density n_2^* on the left, and zero on the right. The two densities overlap in a limited region, within which allele frequencies are approximately constant at $p_1, p_2; p_2 - p_1 = \Delta\mathrm{p}$. (This assumption is accurate for small m). Over a small interval δ, a fraction m of genes from population 1 move into population 2, and so a number $m\,n_1\delta$ move into population 2, which includes $n_2\delta$ individuals in this interval. Thus, migrants make up a proportion $m\,n_1/n_2$ of population 2, and the allele frequency changes at rate $m\,(n_1/n_2)\,\Delta\mathrm{p}$. The change of allele frequency within population 2 is thus:

$$\frac{\partial p_2}{\partial\, t} = \frac{\sigma^2}{2}\frac{\partial^2 p_2}{\partial x^2} + \sigma^2\frac{\partial p_2}{\partial x}\frac{\partial n_2}{\partial x} + m\,\frac{n_1}{n_2}\Delta\mathrm{p} \quad \text{(A.16)}$$

The first term on the right describes diffusion, at a rate given by the variance of the distance between parent and offspring, σ^2. The second term describes the flow of genes from dense to less dense regions (Nagylaki 1976). Assuming equilibrium, so that this

equation is zero, and multiplying through by n_2^2 gives:

$$0 = \frac{\sigma^2}{2}\frac{\partial}{\partial x}\left(n_2^2\frac{\partial p_2}{\partial x}\right) + m\, n_1 n_2 \Delta \mathrm{p} \tag{A.17}$$

Integrating across the overlap region gives:

$$\frac{\sigma^2}{2}\left(n_2^2\frac{\partial p_2}{\partial x}\right)_{x_0} = m \int n_1 n_2 \Delta \mathrm{p} \tag{A.18}$$

where the term on the left is evaluated at the left of the overlap. Since we define $B = \Delta \mathrm{p}/(\partial p/\partial x)$, we have:

$$B = \frac{\sigma^2 n_2^{*2}}{2m \int n_1 n_2 dx} \tag{A.19}$$

The barrier is stronger for flow from a larger population into a smaller, being proportional to n_2^*. In chapter 2, we assume symmetry, leading to the equation in Box 2.

Consequences of barriers

We have argued that reproductive isolation should be defined by the effects of genetic differences on the flow of *neutral* genes. Crucially, we do not include the direct effects of selection, which can readily keep populations distinct, even with no barrier to gene flow. Nevertheless, we are ultimately concerned with the effects of reproductive isolation on selected alleles, and so we summarise here the effects of barriers on various kinds of allele. Our key points are i) that results for neutral alleles still apply if selection on the focal allele is weak, relative to the selection that maintains the barrier, ii) the effects of a barrier depend on both the spatial context, and how selection acts on the allele concerned, and iii) even strong reproductive isolation may have little long-term effect on divergence, especially in spatially extended populations. In this section, we first consider the effect of an idealised barrier on various kinds of allele. Our summary is based on simple heuristic approximations; Piálek and Barton (1997) discuss their accuracy, assuming an idealised barrier in a spatially continuous population.

Flow into a single deme

The simplest case is where alleles flow into a single deme, at an effective rate m_e. Neutral alleles accumulate steadily, converging to the source frequency, u_s, over a timescale $1/m_e : u \sim u_s(1 - e^{-m_e t})$. Alleles with selective advantage s will increase much more rapidly, as $u \sim m_e\left(1 - e^{-(m_e+s)t}\right)/\left(m_e + se^{-(m_e+s)t}\right)$, reaching 50% after $\log\left[2 + s/m_e\right]/(m_e + s)$ generations, if $u_s = 1$. Because favoured alleles increase exponentially, the delay caused by a reduction in gene flow only depends logarithmically on the strength of reproductive isolation: $\sim \log\left[s/m_e\right]/s \mathrm{for} s >> m_e$.

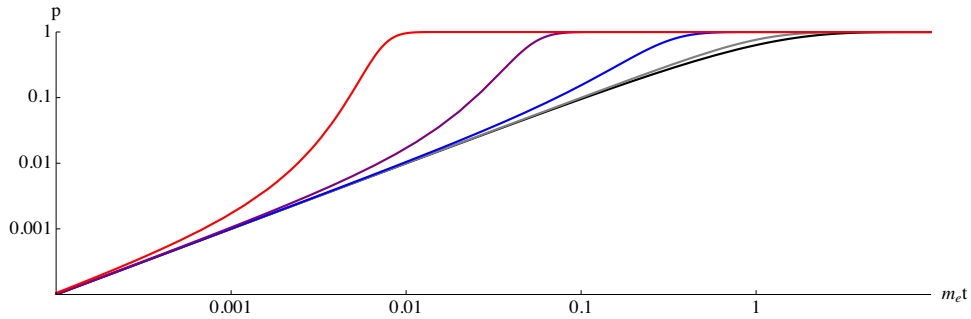


Figure A.5: Increase of alleles through migration and selection. Allele frequency is plotted against scaled time, $m_e t$, for $s/m_e = 0, 1, 10, 100, 1000$ (black,. . .,red). If migration is very low, neutral alleles will take a very long time to become common ($\sim 1/m_e$). However, if selection is much stronger than migration ($s >> m_e$), the time to reach high frequency is proportional to $1/s$, and depends only logarithmically on m_e.

Locally deleterious alleles will reach an equilibrium at $u \sim m_e/s$, but will swamp local adaptations if $m_e > s$. In such a case, genetic incompatibilities that reduce m_e below the threshold required for local adaptation lead to a qualitatively difference. Similarly, if incoming alleles can only increase above a threshold allele frequency (for example, if heterozygotes are less fit, or if there is positive frequency-dependence), then a genetic barrier can prevent their spread.

Random drift can substantially amplify the effect of a barrier. In a deme with effective size N_e diploid individuals, $2N_e m_e$ alleles enter in every generation, and each have a probability $2s$ of establishing. Therefore, there is a delay of $\sim 1/(4N_e m_e s)$ generations before the first such allele establishes and increases. Thus, the delay now increases linearly with the strength of the barrier, rather than logarithmically. Conversely, if increase is possible only above a threshold frequency (as, for example, with underdominance), then rather than being trapped indefinitely, an allele may establish and spread via random drift. In both cases, the effect of reproductive isolation depends on the demographic context.

Flow across one dimension

In a one-dimensional population, the effects of a barrier on the spread of different kinds of allele are similar: neutral alleles are delayed for $\sim (B/\sigma)^2$ generations, but favourable alleles are delayed much less, by only $\sim \log\left[(B/\sigma)\sqrt{2s}\right]/s$ in a dense population. In a finite population, the delay is much longer, by $\sim B/(2s\,\rho_e\sigma^2)$ (Piálek and Barton 1997). Perhaps the most important point to make is that even when a barrier has a dramatic effect on cline shape, inducing a sharp step, the delay in spread of favourable alleles is negligible, on evolutionary timescales. For example, the hybrid zone between *Bombina bombina* and *B. variegata* is ∼6Km wide, but poses a barrier to gene flow of $B\sim$ 100Km, relative to an estimated dispersal rate of σ ∼1Km. Thus, neutral alleles could be delayed

by $\sim (B/\sigma)^2 \sim 10^4$ generations, as is reflected in the limited introgression since postglacial secondary contact. However, an allele with a 1% advantage would be delayed by only 0.5 $\log\left[0.7(B/\sigma)\sqrt{2s}\right]\Big/ s$∼100 generations, assuming a dense population (Piálek and Barton 1997). This compares with the wave of advance of such an allele through unimpeded habitat of only $\sigma\sqrt{2s} \sim 0.14$Kmpergeneration; simple physical distance has a stronger effect than a local genetic barrier, simply because just a few introgressing alleles can initiate an invasion.

Comparing the effects of barriers between two demes vs across one dimension

Comparing Figs. 2.3 and 2.4 shows that although different measures of barrier strength are appropriate for a single deme versus a linear habitat (namely, m/m_e vs. B/σ), these vary in a similar way along the genome. However, this does not imply that their observed consequences will be the same. Figure A.6 shows how F_{st} varies along the genome, when measured between two demes (blue), versus between two points immediately on either side of a hybrid zone (red), assuming an IM model with a population split $0.5N_e$ generations ago. F_{st} falls away much more sharply along the genome in a linear transect, simply because a given barrier has a weaker effect on divergence in a spatial continuum (Barton 2008). In this example, the population has approached an equilibrium between drift and gene flow, so that for two demes, $F_{st} \sim 1/(1+4N_em)$. However, in a spatially continuous population, F_{st} at equilibrium has a more complex form, which depends on the distance from the barrier and on the spatial dimension (Nagylaki 1993, Barton 2008). The population history may well be more complex; in the extreme case of recent secondary contact, F_{st} will reflect the deeper history of the divergent populations, and may hardly be influenced by current gene flow. As is well known, gene flow can only be inferred from F_{st} under restrictive assumptions (Whitlock and McCauley 1999).

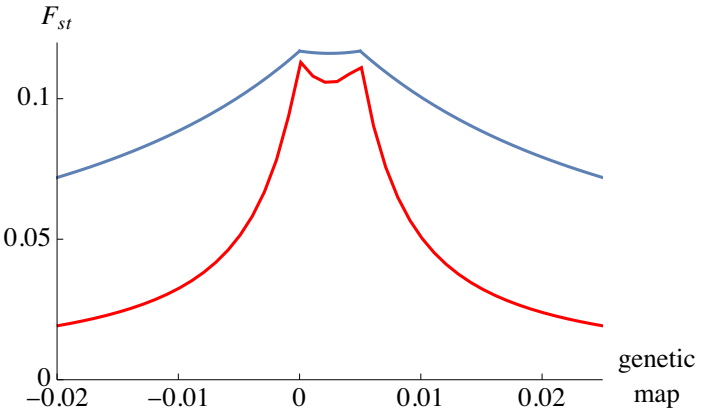


Figure A.6: Comparison between the effects of a barrier on F_{st} between two demes (blue), and immediately on either side of a barrier in one dimension (red). Two loci, each under selection $s = 0.05$, and r=0.5cM apart, are assumed. For one dimension, values are for 200 demes, each of 150 diploid individuals, with a reduction in gene flow between demes 100, 101; results are shown at 2000 generations. For flow between two demes (blue), $N_e m = 6.4$, and $T = 0.5N_e$. Here, F_{st} is calculated by comparing the probability of coalescence more recently than 2000 generations, between two genes on different sides of the barrier, P_b, withthatbetween two genes from within the same location, P_w. Assuming that lineages that do not coalesce in this recent interval share ancestry much further back, $F_{st} = \frac{P_w - P_b}{P_w + P_b}$. The probability of coalescence before time t between genes in demes i, j, $P_{t,i,j}$, is calculated from the recursion $P_{t+1} = \frac{\mathcal{I}}{2N} + \left(1 - \frac{\mathcal{I}}{2N}\right) M.P_t.M^T$, where $\mathcal{I}$ is the identity matrix, and M the migration matrix, which includes the effects of the barrier. (From Tavares et al., (2018) Supplementary text 5.4, Fig. S16).

References

Barton, N., & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. Heredity, 57(3), 357-376.

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

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

Piálek, J., & Barton, N. H. (1997). The spread of an advantageous allele across a barrier: the effects of random drift and selection against heterozygotes. Genetics, 145(2), 493-504.

Barton, N. H. (2008). The effect of a barrier to gene flow on patterns of geographic variation. Genetics research, 90(1), 139-149.

Nagylaki, T. (1993). The evolution of multilocus systems under weak selection. Genetics, 134(2), 627-647.

Whitlock, M. C., & McCauley, D. E. (1999). Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. Heredity, 82(2), 117-125.

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., Elleouet, J., Burrus, M., Andalo, C., Li, M., Li, Q., Xue, Y., Rebocho, A. B., Barton, N., and 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.

Appendix B

SUPPLEMENTARY INFORMATION FOR EFFECT OF ASSORTATIVE MATING AND SEXUAL SELECTION ON POLYGENIC BARRIERS TO GENE FLOW

B.1 Derivation of gene flow factor

Let h_i represent the fraction of individuals in the population with a migrant ancestor precisely i generations ago. Thus, h_1 represents the fraction of $F1$ individuals and h_i the fraction of $(i-1)^{th}$ generation backcrosses (denoted as BC_{i-1}). As specified in chapter 3, an individual with *no* migrant ancestor in the last n generations is designated as a 'resident', where n is an arbitrary threshold. Thus, by definition, the fraction of residents in the population is $h_r = 1 - \sum_{i=1}^{n} h_i$. We further assume that trait values are normally distributed *within* any population subgroup: thus the distribution $P_r(z)$ of trait values among residents is normal with mean $\overline{z}_r$ and variance V_r; similarly, the distribution $P_i(z)$ amongst i^{th} generation descendants of migrants is normal with mean z_i and variance V_i (for any i). The trait value distribution across the entire population is thus the sum of $n+1$ normal distributions and is given by $P(z) = h_r P_r(z) + \sum_{i=1}^{n} h_i P_i(z)$. For simplicity, we will assume that the distribution of trait values amongst migrants is also normal with mean $\overline{z}_0$ and variance V_0; however, this assumption is not crucial. Finally, we will assume that $m \ll 1$ and perform various computations to first order in m. This is justified on the grounds that non-zero adaptive divergence between mainland and island requires m to be comparable to the typical *per locus* selection coefficient s, where $s \ll 1$ for a polygenic

trait influenced by many small-effect loci.

We now consider how the proportions h_i and the distributions $P_i(z)$ associated with different population subgroups change within a single generation due to migration, selection and assortative mating.

Migration from the mainland results in a fraction m of each subgroup being replaced by migrants. Since the proportions h_i of $F1$s, BC_1s, etc., are themselves $\mathcal{O}(m)$ (see below), to lowest order in m these are unchanged. The main effect of migration is thus to reduce the proportion of residents by m; this is balanced by the emergence of a new sub-group— migrants, who constitute a proportion $h_0^{'} = m$ of the population. Thus, following migration, we have (to first order in m): $h_0^{'} = m$, $h_i^{'} = h_i$ for $i = 1, \ldots n$ and $h_r^{'} = h_r - m$. Note that migration does not change the trait value distributions of existing sub-groups, so that: $P_r^{'}(z) = P_r(z)$ and $P_i^{'}(z) = P_i(z)$.

Migration is followed by *viability selection* on the trait, wherein an individual with trait value z contributes to the pool of reproducing adults in proportion to $e^{-\beta z}/\overline{W}_{vs}$. Here, $\overline{W}_{vs}$ is the viability component of mean fitness and is given by: $\overline{W}_{vs} = \int dy\, e^{-\beta y} \left[h_r^{'} P_r^{'}(y) + \sum_{i=0}^{n} h_i^{'} P_i^{'}(y) \right]$. Thus, following viability selection, the proportions of the different subgroups in the pool of reproducing adults are:

$$h_i^{''} = h_i^{'} \frac{\int dz\, e^{-\beta z} P_i^{'}(z)}{\overline{W}_{vs}} \qquad h_r^{''} = h_r^{'} \frac{\int dz\, e^{-\beta z} P_r^{'}(z)}{\overline{W}_{vs}}$$

Assuming that all h_i are $\mathcal{O}(m)$ (see below), it follows that to first order in m, we have:

$$h_i^{''} = h_i^{'} \frac{\int dz\, e^{-\beta z} P_i(z)}{\int dz\, e^{-\beta z} P_r(z)} \qquad h_r^{''} = 1 - \sum_{i=0}^{n} h_i^{''} \tag{B.1a}$$

$$P_i^{''}(z) = \frac{e^{-\beta z} P_i^{'}(z)}{\int dy\, e^{-\beta y} P_i(y)} \qquad P_r^{''}(z) = \frac{e^{-\beta z} P_r^{'}(z)}{\int dy\, e^{-\beta y} P_r(y)} \tag{B.1b}$$

Equation (B.1b) specifies the distribution of trait values among different population subgroups following viability selection. These distributions are still normal, but with a shifted mean $\overline{z}_i^{''} = \overline{z}_i - \beta V_i$ for sub-group i (and $\overline{z}_r^{''} = \overline{z}_r - \beta V$ for residents); the variances within sub-groups remain unchanged ($V_i^{''} = V_i$ and $V_r^{''} = V_r$). Thus, the trait value distribution across the entire population is the sum of $n+2$ normal distributions: $P^{''}(z) = h_r^{''} P_r^{''}(z) + \sum_{i=0}^{n} h_i^{''} P_i^{''}(z)$. Note that this sum now includes the group '0', i.e., individuals who have immigrated at the start of the generation. In the following we will show that βV is $\mathcal{O}(m)$; this also implies that the various βV_i are $\mathcal{O}(m)$ (assuming V_i to be comparable to V). Thus, to first order in m, we can ignore the shifts in trait means (due to viability selection) within subgroups, i.e., within migrants, $F1$s, BC_i and so on. This is because the proportions of these subgroups are $\mathcal{O}(m)$; the shifts in trait mean among these, being equal to βV_i, is also $\mathcal{O}(m)$, so that the contribution of these shifts to

the shift in the overall trait mean of the population is $\mathcal{O}(m^2)$. For the same reason, we cannot ignore the shift in trait mean among residents, since $h_r^{''}$ is $\mathcal{O}(1)$.
Finally, consider the effects of *assortative mating* and *sexual selection* on population subgroups. As described in chapter 3, the mating function $M(x, y)$ specifies the probability of mating between a male and a female with trait values x and y respectively. Since an individual with a given trait value is equally likely to be male or female, the probability of mating between two randomly chosen individuals with trait values x and y is $[M(x, y) + M(y, x)]/2$. Under the weak migration ($m \ll 1$) assumption, migrants mate predominantly with residents to produce $F1$s, which mate with residents to produce BC_1 and so on. Further, under our definition of resident, mating between BC_n and residents also produces residents. Thus, after mating, we have:

$$h_r^{'''} = \int dy \int dx \, \frac{M(x,y) + M(y,x)}{2} \left[\left(h_r^{''}\right)^2 P_r^{''}(x) P_r^{''}(y) \; + \; 2h_r^{''} h_n^{''} \, P_r^{''}(x) P_n^{''}(y) \right] \quad \text{(B.2a)}$$

$$h_i^{'''} = \int dy \int dx \, \frac{M(x,y) + M(y,x)}{2} \left[2h_r^{''} h_{i-1}^{''} \, P_r^{''}(x) P_{i-1}^{''}(y) \right] \qquad i = 1, \ldots, n \quad \text{(B.2b)}$$

Further, the trait value distribution among the various population sub-groups are given by:

$$P_r^{'''}(z) \; = \; \int dy \int dx \, \frac{M(x,y) + M(y,x)}{2} \left[\frac{\left(h_r^{''}\right)^2}{h_r^{'''}} P_r^{''}(x) P_r^{''}(y) \frac{e^{-\frac{\left(z - \frac{x+y}{2}\right)^2}{2V_{r,r}}}}{\sqrt{2\pi V_{r,r}}} \; + \; \frac{2h_r^{''} h_n^{''}}{h_r^{'''}} \, P_r^{''}(x) P_n^{''}(y) \frac{e^{-\frac{\left(z - \frac{x+y}{2}\right)^2}{2V_{r,n}}}}{\sqrt{2\pi V_{r,n}}} \right] \quad \text{(B.3a)}$$

$$P_i^{'''}(z) = \int dy \int dx \, \frac{M(x,y) + M(y,x)}{2} \left[\frac{2h_r^{''} h_{i-1}^{''}}{h_i^{'''}} \, P_r^{''}(x) P_{i-1}^{''}(y) \frac{e^{-\frac{\left(z - \frac{x+y}{2}\right)^2}{2V_{r,i-1}}}}{\sqrt{2\pi V_{r,i-1}}} \right] \qquad i = 1, \ldots, n \quad \text{(B.3b)}$$

Equation B.3 assumes that the trait is sufficiently polygenic that the offspring of any two individuals are normally distributed about the midparent value with a within-family variance that depends on the relatedness between parents, as in the standard infinitesimal model (Barton et al., 2017). As discussed in chapter 3, we further assume that the within-family variance only depends on the subgroups to which the two parents belong. Thus, for example, the within-family variance of the offspring of any resident and BC_i pair is assumed to be $V_{r,i+1}$.

At equilibrium, the trait value distributions and the proportions of various subgroups at

the end of each generation must be the same as at the beginning, so that $h''' = h$ and $P'''(z) = P(z)$. This allows us to use equations (B.1)-(B.3) to derive expressions for the equilibrium proportions h_i and the means and variances $\overline{z}_i$ and V_i of the sub-group trait value distributions (by assuming that these are approximately normal) to first order in m. In the following, we present these derivations separately for Models I and II of assortative mating.

Under **Model I**, we have:

$$M(x,y) \;=\; M(y,x) \;=\; \frac{e^{-\gamma(x-y)^2}}{\int\, d\tilde{x}\, \int\, d\tilde{y}\, P''(\tilde{x})P''(\tilde{y})e^{-\gamma(\tilde{x}-\tilde{y})^2}}$$

where $P''(x)$ is the distribution of trait values in the population just after viability selection. As described above, to first order in m, this is simply: $P^{''}(x) = (1 - \sum_{i=0}^{n} h_i^{''})P_r^{''}(x) + \sum_{i=0}^{n} h_i^{''} P_i^{''}(x)$. Substituting into the mating function $M(x,y)$ and using eq. (B.2a), it follows that to lowest order in m, we have: $h_i = h_i^{'''} = 2h_{i-1}^{''} \frac{\int d\tilde{x} \int d\tilde{y}\, e^{-\gamma(\tilde{x}-\tilde{y})^2} P_r''(\tilde{x})P_{i-1}''(\tilde{y})}{\int d\tilde{x} \int d\tilde{y}\, e^{-\gamma(\tilde{x}-\tilde{y})^2} P_r''(\tilde{x})P_r''(\tilde{y})}$ and $h_r = 1 - \sum_{i=1}^{n} h_i$. Using eq. (B.1) for $P_i^{''}(z)$ and h_i'' finally gives the following expressions for the equilibrium proportions of the various subgroups:

$$h_i = 2h_{i-1}\overline{W}_{i-1} \quad (\text{for} \quad i = 1, \ldots, n) \quad \text{where} \quad h_0 = m; \qquad\qquad h_r = 1 - \sum_{i=1}^{n} h_i$$

where,

$$\begin{aligned} \overline{W}_i \;&=\; \frac{\int\, d\tilde{x}\, \int\, d\tilde{y}\, e^{-\gamma(\tilde{x}-\tilde{y})^2}\, e^{-\beta(\tilde{x}+\tilde{y})} P_r(\tilde{x})P_i(\tilde{y})}{\int\, d\tilde{x}\, \int\, d\tilde{y}\, e^{-\gamma(\tilde{x}-\tilde{y})^2}\, e^{-\beta(\tilde{x}+\tilde{y})} P_r(\tilde{x})P_r(\tilde{y})} \\ &=\; \sqrt{\frac{1+4\gamma V_r}{1+2\gamma(V_r+V_i)}} \exp\left[-\frac{\beta(1+4\gamma V_r)(\overline{z}_i - \overline{z}_r) + \gamma(\overline{z}_i - \overline{z}_r)^2 - \frac{\beta^2}{2}(1+4\gamma V_r)(V_i - V_r)}{1+2\gamma(V_r+V_i)} \right] \end{aligned} \tag{B.4}$$

It follows from eq. (B.4) above that $h_1 = 2m\overline{W}_0$ (since $h_0 = m$), $h_2 = 4m\overline{W}_0\overline{W}_1$, and (more generally) that: $h_i = 2^i m \prod_{k=1}^{i-1} \overline{W}_k$. This establishes that the various h_i are all $\mathcal{O}(m)$.

Equation (B.4) simply states that the equilibrium proportion of i^{th} generation descendants of migrants (e.g., BC_1 for $i = 2$) is equal to two times the equilibrium proportions of the two parental subgroups (e.g., residents and $F1$s for BC_1s) multiplied by the mean relative fitness of the two parental subgroups. The proportion of residents is $h_r = 1 - \mathcal{O}(m)$ and their relative fitness is $1 + \mathcal{O}(m)$, whereas the various h_i are $\mathcal{O}(m)$. Thus, to first order in m, we have: $h_i = 2h_{i-1}\overline{W}_{i-1}$, as above. Here, $\overline{W}_i$ denotes the mean fitness of i^{th} generation descendants of migrants relative to residents, and is thus the same $\overline{W}_i$ that

enters the expression for the gene flow factor g (equation 3.2 of chapter 3). From the above equation, we see that $\overline{W}_i$ is influenced by the strength of (viability and sexual) selection as well as the trait value distributions within subgroups.

However, eq. (B.4) by itself does not allow us to calculate g, as the means ($\overline{z}_r$ and $\{\overline{z}_i,\, i = 1, \ldots, n\}$) and variances ($\overline{V}_r$ and $\{\overline{V}_i,\, i = 1, \ldots, n\}$) of the sub-group trait value distributions are unknown. These can be determined using eq. (B.3). To do this, we will further assume that the threshold n in eq. (B.3a) is sufficiently large that $\overline{z}_n - \overline{z}_r \ll \sqrt{V_r}$, so that $P_r(z)$ (which is strictly a mixture of normal distributions; see eq. (B.3a) can be approximated by a single normal distribution with moments equal to the weighted sum of moments of the constituent distributions. Recall that n here is an arbitrary threshold such that an individual with *no* migrant ancestor in the last n generations is designated as a 'resident'.

Also note that we are interested in deriving expressions for $\overline{z}_r$, V_r and $\{\overline{z}_i, V_i\}$ that are correct to first order in migration rate m. Since the effective migration rate m_e is simply m multiplied by the gene flow factor g, this means that we need to derive expressions for g or alternatively for $\{\overline{z}_i, V_i\}$) (which determine $\{\overline{W}_i\}$, which determine g) that are correct to *zeroth* order in m. To do this, it is useful to first enumerate all the quantities that are $\mathcal{O}(m)$. As argued above, this includes all the equilibrium proportions h_i. Further, βV is also $\mathcal{O}(m)$: this will be shown for the special case $\gamma V \ll 1$ below, but follows more generally from the fact that under migration-selection balance, the change in trait mean due to selection (which is equal to βV) must be balanced by the change due to migration (which is proportional to m). Further, since $V_i \sim V$ (see below), it follows that all βV_i are also $\mathcal{O}(m)$. Thus, assuming that $\{h_i\}$, $\{\beta V_i\}$ and βV are all $\mathcal{O}(m)$, and using eq. (B.3), we obtain the following equations for $\overline{z}_r$, V_r, and $\{\overline{z}_i, V_i\}$ (to lowest order in m):

$$\beta V_r \;=\; 2h_n \frac{(1+4\gamma V_r)}{1+2\gamma(V_r+V_n)} \frac{\overline{z}_n - \overline{z}_r}{2} \qquad\qquad V_r \;=\; 2V_{r,r} \tag{B.5a}$$

$$\overline{z}_i - \overline{z}_r \;=\; \frac{1+4\gamma V_r}{1+2\gamma(V_r+V_{i-1})} \frac{\overline{z}_{i-1} - \overline{z}_r}{2} \qquad V_i \;=\; \frac{V_r + V_{i-1} + 8\gamma V_r V_{i-1}}{4(1+2\gamma(V_r+V_{i-1}))} + V_{r,i-1} \tag{B.5b}$$

$$\overline{W}_i \;=\; \sqrt{\frac{1+4\gamma V_r}{1+2\gamma(V_r+V_i)}} \exp\left[-\frac{\beta(1+4\gamma V_r)(\overline{z}_i - \overline{z}_r) + \gamma(\overline{z}_i - \overline{z}_r)^2}{1+2\gamma(V_r+V_i)} \right] \tag{B.5c}$$

Equation (B.5a) can be derived by integrating eq. (B.3a) to compute the various moments of $P_r'''(z)$, then using the equilibrium condition $\overline{z}_r''' = \overline{z}_r$ and $V_r''' = V$, and finally retaining terms that are lowest order in m. Equation (B.5b) follows similarly from eq. (B.3b). Equation (B.5c) is simply obtained from the expression for $\overline{W}_i$ in eq. (B.4) by retaining only $\mathcal{O}(1)$, i.e., lowest (zeroth) order terms in m.

Note that eq. (B.5) depends on the various segregation (within-family) variances V_{rr}

and V_{ri}, which are (so far) unknown, but must depend on allele frequencies $\{p_{i,j},\, j = 1,\ldots,L\}$ (and $\{p_{r,j},\, j = 1,\ldots,L\}$) at the various trait loci in the various population sub-groups. Here, $p_{i,j}$ (or $p_{r,j}$) represents the allele frequency at the j^{th} locus among the i^{th} generation descendants of migrants (or among residents). More concretely, we have: $V_{rr} = \sum_{j=1}^{L} \frac{\alpha_j^2}{2}\, p_{r,j}(1-p_{r,j})$ and $V_{ri} = \sum_{j=1}^{L} \frac{\alpha_j^2}{2}\, [p_{r,j}(1-p_{i,j}) + p_{i,j}(1-p_{r,j})]$. In the absence of assortment, allele frequencies among i^{th} generation descendants will simply be the average of the parental allele frequencies (i.e., of residents and $(i-1)^{th}$ generation descendants). With assortment, $\{p_{i,j}\}$ are the weighted average of parental allele frequencies, which gives (under Model I): $p_{i,j} - p_{r,j} = \frac{1+4\gamma V_r}{1+2\gamma(V_{i-1}+V_r)} \frac{p_{i-1,j}-p_{r,j}}{2}$ (this is analogous to equation (B.5b) for $\overline{z}_i - \overline{z}_r$). From this, we finally have:

$$V_{ri} = V_{rr} + \left(\prod_{k=0}^{i-1} \frac{1+4\gamma V_r}{1+2\gamma(V_k+V_r)}\right) \sum_{j=1}^{L} \alpha_j^2 \frac{p_{0,j}-p_{r,j}}{2^{i+1}}(1-2p_{r,j}), \quad \text{where} \quad V_{rr} = \sum_{j=1}^{L} \frac{\alpha_j^2}{2}\, p_{r,j}(1-p_{r,j}) \tag{B.6}$$

Thus, we can express all segregation variances in terms of the (as yet) unknown allele frequencies $\{p_{r,j}\}$ in the resident pool.

The main idea behind our theoretical approach is that if trait loci are unlinked (or loosely linked), then the allele frequency at any one locus is still (approximately) predicted by the theory for a *single* locus, but with the raw migration rate m replaced by an effective migration rate m_e, which depends on allele frequencies at all the other trait loci. More concretely, in section B.2, we argue that at migration-selection balance (and neglecting drift), allele frequencies in the resident pool satisfy: $s_j p_{r,j}(1-p_{r,j}) + m_e(p_{0,j} - p_{r,j})$, where $s_j = \alpha\beta_j$ and $p_{0,j}$ is the allele frequency among migrants at the j^{th} locus. If the mainland is fixed for alleles that are locally deleterious on the island (which is the scenario we consider in chapter 3), we have $p_{0,j} = 1$, which yields $p_{r,j} = m_e/s_j$.

We now have all the ingredients necessary for self-consistently and iteratively determining allele frequencies and mean trait divergence. The iterative procedure can be implemented as follows: one starts with an arbitrary 'guess' for the allele frequencies $\{p_{r,j}\}$; this allows one to determine the segregation variance V_{rr} (of offspring produced by mating between residents) using eq. (B.6), as well as the trait mean ($\overline{z} = \sum_{j=1}^{L} \alpha_j p_{r,j}$) and trait variance ($V_r = 2V_{rr}$) among residents. These can then be used to determine the segregation variance V_{r1} (of $F1$ offspring produced by resident $\times$ migrant mating) using eq. (B.6), as well as $\overline{z}_1$ and V_1 using eq. (B.5b). These can now be used to determine V_{r2} (using eq. (B.6)), which can be used to determine $\overline{z}_2$ and V_2 (using eq. (B.5b)), allowing us to sequentially determine $\{\overline{z}_i, V_i, V_{ri}\}$ for $i = 1 \ldots n$. The trait means and variances can now be substituted into eq. (B.5c) to obtain $\{\overline{W}_i,\, i = 1 \ldots n\}$, the product of which yields the gene flow factor g. This can be used to calculate the new m_e, which yields the new allele frequencies via $p_{r,j} = m_e/s_j$. This procedure can be iterated until the allele frequencies

$\{p_{r,j}\}$ converge to a fixed set of values.

However, we can obtain a simple expression for g if γV_r and γV_i are $\ll 1$ (see below for a discussion of this condition). Then, equation (B.5) reduces to:

$$\beta V_r = h_n(\overline{z}_n - \overline{z}_r) \qquad V_r \;=\; 2V_{r,r} \tag{B.7a}$$

$$\overline{z}_{i+1} - \overline{z}_r \;=\; \frac{\overline{z}_i - \overline{z}_r}{2} \qquad V_{i+1} \;=\; \frac{V_r + V_i}{4} + V_{r,i} \tag{B.7b}$$

$$\overline{W}_i \;=\; \exp\left[-\beta(\overline{z}_i - \overline{z}_r) - \gamma(\overline{z}_i - \overline{z}_r)^2\right] \tag{B.7c}$$

In other words, the mean phenotypic difference between each successive backcross generation and residents is half that of the previous backcross generation (eq. (B.7b)), so that: $\overline{z}_i - \overline{z}_r \;=\; (\overline{z}_0 - \overline{z}_r)/2^i$. Substituting into eq. (B.7c) gives: $\overline{W}_i \;=\; \exp\left[-\beta\frac{(\overline{z}_0 - \overline{z}_r)}{2^i} - \gamma\frac{(\overline{z}_0 - \overline{z}_r)^2}{4^i}\right]$, which can be substituted into $g \;=\; \lim_{n\to\infty} \prod_{i=1}^{n} \overline{W}_i \;=\; \prod_{i=1}^{\infty} \overline{W}_i$. This gives the approximate expression for g under Model I in equation 3.3a of chapter 3.

A second observation is that substituting $h_n = 2^n m \prod_{k=1}^{n-1} \overline{W}_k$ and $\overline{z}_n - \overline{z}_r \;=\; (\overline{z}_0 - \overline{z}_r)/2^n$ into eq. (B.7a) yields $\beta V_r = m\left(\prod_{k=1}^{n-1} \overline{W}_k\right)(\overline{z}_0 - \overline{z}_r)$. If we now take the limit $n \to \infty$, we have: $\beta V_r = m_e(\overline{z}_0 - \overline{z}_r)$. This is precisely what one would obtain by multiplying our heuristic equation $s_j p_r(1 - p_{r,j}) = m_e(p_{0,j} - p_{r,j})$ for allele frequency dynamics by the effect size α_j (where $s_j = \beta\alpha_j$), and summing over j. This demonstrates the self-consistent nature of our solution.

In summary, our simple approximation for the gene flow factor g (eq. 3.4a of chapter 3) follows from the more general equations (B.2) and (B.3) by making two additional assumptions– first, that migration rates are sufficiently low that we need only consider $\mathcal{O}(m)$ terms, and second, that γV_r , $\gamma V_i \ll 1$. As discussed above, the low migration assumption implies that: (i) the probability of mating between individuals with recent immigrant ancestry (e.g., $F1 \times F1$ or $F1 \times BC_1$ mating) is negligible, thus allowing us to express g in terms of the average fitness of successive backcrosses (see also Westram et al.,2022, and (ii) the effect of viability selection within any sub-group of i^{th} generation descendants can be neglected, since this only has an $\mathcal{O}(m^2)$ effect on the population trait mean.

We now turn to the second assumption, namely that $\gamma V \ll 1$. To understand what this implies, note that in a population with normally distributed values (with variance V), the phenotypic correlation between mating individuals is equal to $\frac{2\gamma V}{1+2\gamma V}$ under Model I and $\frac{2\gamma V}{\sqrt{1+2\gamma V(1+2\gamma V)}}$ under Model II. Thus, the correlation between mates (which can also be used as a measure of the strength of assortative mating) is $\approx 2\gamma V$, if $\gamma V \ll 1$,

under both models. The condition that $\gamma V_r \ll 1$ and $\gamma V_i \ll 1$ can thus be interpreted as assortative mating being sufficiently weak that mate choice is ineffective and phenotypic correlation between mates is insignificant for mating *within* any subgroup. However, as long as $\gamma(\overline{z}_0 - \overline{z}_r)^2$ is not too small, mating *between* subgroups (e.g., between residents and $F1$s) is still much less likely than resident×resident mating.

We now outline the derivation of the gene flow factor under **Model II**. This follows the same logic as above but involves somewhat more cumbersome expressions. Under Model II, the mating function is:

$$M(x,y) \;=\; \frac{e^{-\gamma(x-y)^2}}{\int\, d\tilde{x}\, P''(\tilde{x}) e^{-\gamma(\tilde{x}-y)^2}} \;=\; \frac{e^{-\gamma(x-y)^2}}{\int\, d\tilde{x}\, P_r''(\tilde{x}) e^{-\gamma(\tilde{x}-y)^2}} \;+\; \mathcal{O}(m)$$

Note that in this case, $M(x,y) \neq M(y,x)$, i.e., the probability of mating between a male and female with trait values x and y respectively is different from that of a female-male pair with values x and y.
As before, from eq. (B.2), it follows that: $h_i = h_{i-1}'' \left[1 + \int\, dy\, P_r''(y) \frac{\int dx\, e^{-\gamma(x-y)^2} P_{i-1}''(x)}{\int d\tilde{x}\, e^{-\gamma(\tilde{x}-y)^2} P_r''(\tilde{x})}\right]$ and $h_r = 1 - \sum_{i=1}^{n} h_i$ to lowest order in m. Using equation (B.1) for $P_i''(z)$ and h_i'' then yields the following expressions for the equilibrium proportions of the various subgroups:

$$h_i = 2h_{i-1}\overline{W}_{i-1} \quad (\text{for} \quad i = 1, \ldots, n) \quad \text{where} \quad h_0 = m; \qquad\qquad h_r = 1 - \sum_{i=1}^{n} h_i$$

$$\begin{aligned}
\overline{W}_i \;&=\; \frac{\int\, dx\, e^{-\beta x} P_i(x) \left[\frac{1}{2} + \frac{1}{2} b(x)\right]}{\int\, dz\, e^{-\beta z} P_r(z)} \quad \text{where} \quad b(x) = \int\, dy\, e^{-\beta y} P_r(y) \frac{e^{-\gamma(x-y)^2}}{\int\, d\tilde{x}\, e^{-\gamma(\tilde{x}-y)^2}\, e^{-\beta\tilde{x}} P_r(\tilde{x})} \\
&= \exp\left[-\beta\left(\overline{z}_i - \overline{z} - \frac{\beta}{2}(V_i - V)\right)\right] \\
&\qquad \left(\frac{1}{2} + \frac{1}{2}\frac{1 + 2\gamma V}{\sqrt{1 + 2\gamma(V + 2\gamma V^2 + V_i)}} \exp\left[-\frac{\gamma(\overline{z}_i - \overline{z} - \beta(V_i - V))^2}{1 + 2\gamma(V + 2\gamma V^2 + V_i)}\right]\right) \qquad \text{(B.8)} \\
&= \exp\left[-\beta\left(\overline{z}_i - \overline{z}\right)\right] \left(\frac{1}{2} + \frac{1}{2}\frac{1 + 2\gamma V}{\sqrt{1 + 2\gamma(V + 2\gamma V^2 + V_i)}} \exp\left[-\frac{\gamma(\overline{z}_i - \overline{z})^2}{1 + 2\gamma(V + 2\gamma V^2 + V_i)}\right]\right) \\
&\qquad + \mathcal{O}(m)
\end{aligned}$$

As in the case of Model I, this yields: $h_i = 2^i m \prod_{k=0}^{i-1} \overline{W}_k$.
We can now use eq. (B.3) together with the function $M(x,y)$ for Model II (as given on the previous page) to derive expressions relating $\{\overline{z}_i, V_i\}$ and $(\overline{z}_r, V_r)$ to each other, restricting ourselves (as before) to computations that are correct to lowest order in m. However, there is now an important difference (as compared to Model I): under Model II, if trait values are normally distributed among residents and (say) among $F1$s, then the

distribution of trait values among their (BC_1) offspring will be the *sum* of two different normal distributions– one corresponding to offspring of resident father and F1 mothers and the other to offspring of resident mothers and F1 fathers. These distributions may be slightly shifted away from the midparent value towards the F1 mean (for offspring of F1 mothers and resident fathers) or towards the resident mean (for offspring of resident mothers and F1 fathers, and provided $V_1 > V$) and thus are not identical. However, one can show that the squared difference between the mean of these two distributions relative to the variance is proportional to γV, and will thus be small for $\gamma V \ll 1$. Then, we can approximate the mixture of two normal distributions by a single normal distribution with moments equal to the weighted sum of the constituent distributions. This is the additional assumption that we need to make in order to derive expressions for $\{\overline{z}_i, V_i\}$ and $(\overline{z}_r, V_r)$ under Model II (in analogy to eq. (B.5) for Model I).

However, these expressions are somewhat cumbersome and we do not provide them here. Instead we focus on a parameter regime in which γV_r, γV_i are $\ll 1$. Then, as in Model I, the expressions for the trait means and variances and average relative fitness of the subgroups simplify considerably, and we have:

$$\beta V_r = h_n(\overline{z}_n - \overline{z}_r) \qquad V_r = 2V_{r,r} \tag{B.9a}$$

$$\overline{z}_{i+1} - \overline{z}_r = \frac{\overline{z}_i - \overline{z}_r}{2} \qquad V_{i+1} = \frac{V_r + V_i}{4} + V_{r,i} \tag{B.9b}$$

$$\overline{W}_i = e^{-\beta(\overline{z}_i - \overline{z}_r)} \frac{1 + e^{-\gamma(\overline{z}_i - \overline{z}_r)^2}}{2} \tag{B.9c}$$

Note that equations (B.9a) and (B.9b) are identical to eqs. (B.7a) and (B.7b). However, the mean fitness of the i^{th} subgroup is different under the two models (eq. (B.7c) vs. eq. (B.9c)), reflecting the fact that mating success is reduced for both male and female non-resident individuals under Model I but only for non-resident males under Model II. It follows from eq. (B.9b) that $\overline{z}_i - \overline{z}_r = (\overline{z}_0 - \overline{z}_r)/2^i$. Substituting into eq. (B.9c) gives: $\overline{W}_i = \exp[-\beta(\overline{z}_0 - \overline{z}_r)/2^i] \frac{(1+\exp\left[-\gamma(\overline{z}_0 - \overline{z}_r)^2/4^i\right])}{2}$, which can be substituted into $g = \lim_{n\to\infty} \prod_{i=1}^{n} \overline{W}_i = \prod_{i=1}^{\infty} \overline{W}_i$. This gives the approximate expression for g under Model II (equation 3.4b of chapter 3).

B.2 Allele frequency dynamics at individual trait loci under Linkage Equilibrium (LE)

In this section, we derive the equation describing the evolution of allele frequencies at any trait locus by neglecting statistical associations or LD between loci, i.e., under

Linkage Equilibrium (LE). The key result of this section is that under LE, sexual selection contributes negligibly to *direct* selection at any locus as long as the trait is sufficiently polygenic (see below for a more precise statement). Instead, its main effect is to increase indirect selection (due to LD), which we describe in terms of a reduction in the effective migration rate m_e of deleterious alleles (section B.1).

Consider the change in allele frequency $(\Delta p)_{sel}$ due to a single generation of natural and sexual selection at a focal locus with additive effect α and allele frequency p. This can be expressed as:

$$(\Delta p)_{sel} \;=\; p(1-p)\frac{\int dz'\, W_1(z')P_*(z') \;-\; \int dz'\, W_0(z')P_*(z')}{p\,\int dz'\, W_1(z')P_*(z') \;+\; (1-p)\,\int dz'\, W_0(z')P_*(z')} \tag{B.10}$$

where $W_1(z')$ (or $W_0(z')$) is the fitness of an individual that carries the '1' (or '0') allele at the focal locus and has trait value $z = z' + \alpha$ (or $z = z'$). Thus, z' is the contribution to trait value due to all loci minus the focal locus, and $P_*(z')$ is the distribution of z' in the population. Note that under the LE assumption, this distribution must be the same for individuals carrying the '0' vs. '1' allele at the focal locus. Now, if L is large, then $P_*(z')$ is approximately normal with mean $\overline{z}_* = \overline{z} - \alpha p$ and variance $V_* = V - \alpha^2 p(1-p)$, where $\overline{z}$ and V are the mean and variance of trait values (as determined by all L loci) in the population.

Under **Model I**, an individual with trait value z has fitness $W(z) \;\propto\; e^{-\beta z}\int dy\, e^{-\beta y - \gamma(z-y)^2} P(y)$, where $P(z) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{(z-\overline{z})^2}{2V}}$ is the distribution of trait values in the population. This gives: $W(z) \;\propto\; \exp\Big[-\frac{1+4\gamma V}{1+2\gamma V}\beta z - \frac{\gamma}{1+2\gamma V}(z-\overline{z})^2\Big]$. Now substituting $W_1(z') = W(z'+\alpha)$ and $W_0(z') = W(z')$ into eq. (B.10), performing the integrals over the (normal) distribution $P_*(z')$, and Taylor expanding in powers of α, one obtains:

$$(\Delta p)_{sel} \;= -\beta\alpha\, p(1-p) + \left(\frac{\beta^2\alpha^2}{2} - \frac{\gamma\alpha^2}{1+4\gamma V_r}\right)\, p(1-p)(1-2p) + \mathcal{O}(\alpha^3) \tag{B.11}$$

Under **Model II**, the fitness of an individual with trait value z is: $W(z) \;\propto\; e^{-\beta z}\int dy\, e^{-\beta y}\left[1 + \frac{e^{-\gamma(z-y)^2}}{\int d\tilde{z} e^{-\beta\tilde{z}} e^{-\gamma(\tilde{z}-y)^2} P(\tilde{z})}\right] P(y) \;=\; e^{-\beta z}\left(1 + \frac{1+2\gamma V}{\sqrt{1+2\gamma V+4\gamma^2 V^2}}\, e^{-\frac{\gamma(z-\overline{z}+\beta V)^2}{2\gamma V(1+2\gamma V)}}\right)$. As before, we can substitute $W_1(z') = W(z'+\alpha)$ and $W_0(z') = W(z')$ into eq. (B.10), integrate over $P_*(z')$, and Taylor expand in powers of α. This gives:

$$(\Delta p)_{sel} \;= -\beta\alpha\, p(1-p) + \left(\frac{\beta^2\alpha^2}{2} - \frac{\gamma\alpha^2}{2(1+2\gamma V_r)^2}\right)\, p(1-p)(1-2p) + \mathcal{O}(\alpha^3) \tag{B.12}$$

Thus, if α is sufficiently small (see below for a more precise condition), then to leading order in α, we have: $(\Delta p)_{sel} \;\approx -\beta\alpha\, p(1-p)$ under both models of assortative mating. If we now assume the mainland to be fixed for alleles that are locally deleterious on the

island, then the per generation change in allele frequency due to migration is simply $(\Delta p)_{mig} = m(1-p)$. Then, at migration-selection balance (and assuming LE): $\Delta p = (\Delta p)_{sel} + (\Delta p)_{mig} = -sp(1-p) + m(1-p) = 0$, where $s = \beta\alpha$. Thus, the equilibrium allele frequency is $p = m/s$ if $m < s$, and $p = 1$ otherwise.

How small does α need to be for us to be able to neglect $\mathcal{O}(\alpha^2)$ terms in equations (B.11) and (B.12)? This requires that both $\beta^2\alpha^2$ and $\gamma\alpha^2$ be much smaller than $\beta\alpha$, which in turn requires: $\beta\alpha \ll 1$ and $\alpha \ll \beta/\gamma$. The first condition, namely, $s = \beta\alpha \ll 1$ simply means that the change in allele frequency in any generation due to selection (or migration) should be sufficiently small that allele frequency dynamics can be approximated by continuous time equations (as is standard in population genetics). The second condition $\alpha \ll \beta/\gamma$ can be rewritten as $\frac{\alpha}{\bar{z}_0 - \bar{z}} \ll \frac{\beta(\bar{z}_0 - \bar{z})}{\gamma(\bar{z}_0 - \bar{z})^2}$: this translates into the requirement that the relative contribution of any locus to mean trait divergence is much smaller than the ratio of the sexual selection component of migration load relative to the viability selection component. This condition is satisfied as long as the two components are comparable in magnitude.

B.3 Deterministic simulations under the hypergeometric model

We perform deterministic simulations of the mainland-island model based on the hypergeometric model. The order of events on the island is migration, viability selection, assortative mating (which generates sexual selection), and free recombination between parental genotypes. The mainland is assumed to be fixed for the alleles that are locally deleterious on the island. Additionally, unless specified otherwise, the island is initially fixed for the favorable allele at all trait loci, as in a secondary contact scenario.

Let $P(z)$ be the probability distribution of trait values $z = \sum_{i=0^L} \alpha_i X_i$ on the island, where α_i is the effect size and $X_i = 0, 1$ denote alternate alleles at locus i. Under the hypergeometric model, all L loci have equal effect (denoted by α) on the trait and all multi-locus genotypes corresponding to a given trait value are equally abundant in the population. Thus, we can express the trait value of an individual as $z = \alpha k$, where k is the number of '1' alleles carried by the individual. Further, the distribution $P(z)$ is equivalent to the distribution $P(k)$ on the island. Starting with a distribution $P_t(k)$ at the end of generation t, the distribution is evolved under migration, viability selection and assortative mating as follows: :

Migration: $P'(k) = (1-m)P_t(k) + m\delta_{k,L}$, where $\delta_{k,L} = 1$ if $k = L$ and is 0 otherwise.

Viability selection: $P''(k) = \dfrac{W(k)P'(k)}{\sum\limits_{j=0}^{L} W(j)P'(j)}$, where $W(k) = e^{-\beta\alpha k}$

Assortative mating and free recombination: $P'''(k) = \sum_{i=0}^{L}\sum_{j=0}^{L} P''(i)P''(j)M(x = i\alpha, y = j\alpha)R_{i,j\to k}$. Here, $R_{i,j\to k}$ is the probability that parents carrying i and j alleles (say, of type '1') produce an offspring with k '1' alleles. This is given by: $R_{i,j\to k} = \sum_{h=0}^{min(j,k,i+j-k)} \frac{\binom{i}{h}\binom{L-i}{j-h}}{\binom{L}{j}}\binom{i+j-2h}{k-h}(\frac{1}{2})^{i+j-2h}$, where $j \leq i$, and $max(0, i+j-L) < k < min(i+j, L)$, and L is the number of loci (see equation A2 in Barton 1992). Also, $M(x, y)$ is the probability of mating between a male and female with trait values x and y respectively, where $x = i\alpha$ and $y = j\alpha$ (see section B.1 for concrete formulae under Models I and II). Thus, at the end of the $(t+1)^{th}$ generation, we have: $P_{t+1}(k) = P'''(k)$.

The steps above are iterated in each generation until equilibrium is attained, i.e., the average allele frequency $\overline{p} = \frac{1}{L}\sum_{k=0}^{L} kP(k)$ (averaged over all trait loci) no longer changes with time. In the following, we will depict our results in terms of the average allele frequency divergence per trait locus, $Y = 1 - \overline{p}$.

B.4 Critical divergence thresholds and migration rates

As described in chapter 3 (see also section B.2), the dynamics of the average allele frequency divergence per locus, $Y = 1 - \overline{p}$ can be approximated as:

$$dY/dt \approx sY(1-Y) - m_e\, Y \tag{B.13}$$

This is the same as the equation for a single selected locus, but with m replaced by the effective migration m_e, which captures the effect of LD between the focal locus and all other trait loci. Furthermore, $m_e = m\, g(Y)$, where g is the gene flow factor (given by equation 3.3 in chapter 3), which itself depends on the average divergence level (and on $\beta\Delta_0$ and $\gamma\Delta_0^2$, which quantify respectively the strength of viability and sexual selection). Thus, at migration-selection-assortment equilibrium, we have $dY/dt = 0$, so that the equilibrium divergence level is given by the solution(s) of:

$$f(Y) = Y(1-Y) - \frac{m}{s}g(Y)Y = 0 \tag{B.14}$$

This yields one or more equilibria Y^*, which are stable provided $f'(Y^*) < 0$. This can be visualised by plotting $f(Y)$ vs Y (see figure B.1). For any m/s, the point at which the curve intersects the horizontal axis corresponds to an equilibrium; it is stable if the curve is downward-sloping at that point and unstable otherwise. $Y^* = 0$ is always an equilibrium and is stable if $m/s > 1$. Additionally, there may be other equilibria with $Y^* > 0$ which satisfy $1 - Y^* = \frac{m}{s}g(Y^*)$.

For weak selection, say when $\beta\Delta_0 = 0.2$, $\gamma\Delta_0^2 = 0.5$ (see figure B.1A), there is an unstable equilibrium at 0 and a non-zero stable equilibrium (as marked by a cross) when $m/s < 1$. Increasing m/s causes this equilibrium to shift to smaller values; in other words, there is a smooth decline in adaptive allele frequencies until divergence is lost at the critical migration threshold, $m_c/s = 1$. For $m/s \geq 1$, the only stable equilibrium is at $Y^* = 0$.

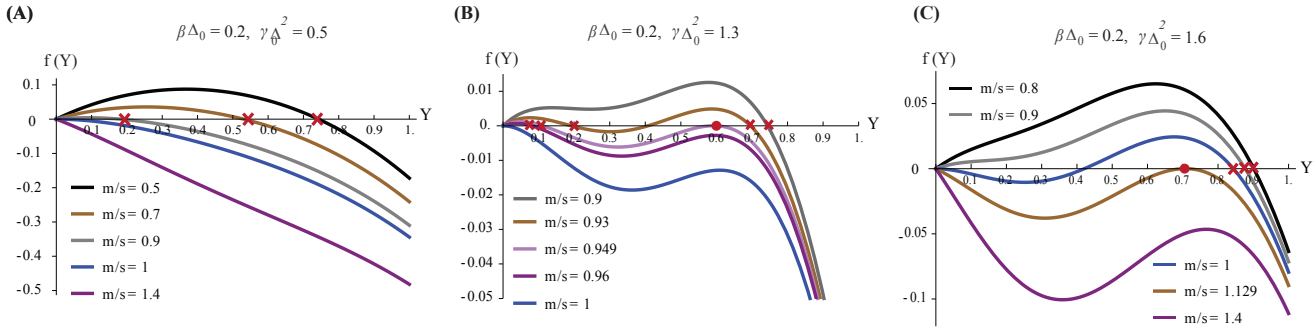


Figure B.1: Different behaviours of the equilibria are visualized by plotting $f(Y)$ vs. Y for Model I of assortative mating, where Y is the mean allele frequency divergence per locus and $f(Y) = \frac{1}{s}\, dY/dt$ is given by equation B.14. For a given m/s, equilibria Y^* are the points of the curve that intersect the horizontal axis, i.e., for which $f(Y^*) = 0$, and are stable if the curve is downward sloping, i.e., if $f'(Y^*) < 0$. (A) Weak selection ($\beta\Delta_0 = 0.2$, $\gamma\Delta_0^2 = 0.5$): there is a non-zero stable equilibrium (marked by a cross) if $m/s < 1$, which disappears at a critical migration threshold, $m_c/s = 1$, beyond which 0 is the only stable equilibrium. (B) Intermediate selection ($\beta\Delta_0 = 0.2$, $\gamma\Delta_0^2 = 1.3$): there are two non-zero stable equilibria for certain migration thresholds, e.g., for $m/s = 0.93$ (see crosses), separated by an unstable equilibrium. With increasing m/s, the unstable and stable equilibria come closer and collide at a critical divergence threshold Y_c (marked by a circle for $m/s = 0.949$). This equilibrium vanishes with a slight increase in m/s, say when $m/s = 0.95$, resulting in a sudden drop in divergence to a low but non-zero value (marked by a cross). A further increase in m causes divergence to drop gradually until 0 becomes the only stable equilibrium beyond the critical migration threshold $m_c/s = 1$. (C) Strong selection ($\beta\Delta_0 = 0.2$, $\gamma\Delta_0^2 = 1.6$): As with weak and intermediate selection, there is one non-zero stable equilibrium when $m/s < 1$ and 0 is an unstable equilibrium. When $m/s \geq 1$, there are two stable equilibria– one at 0, and the other at a high divergence level, separated by an unstable equilibrium. At a critical migration threshold $m_c/s = 1.129$, the stable and unstable equilibria collide– this occurs at a divergence level Y_c (marked by a circle). Beyond this critical migration threshold, 0 is the only stable equilibrium. Note that the curve $f(Y)$ always has a maximum at $Y = Y_c$ (marked by circles in B and C).

The behaviour of equilibria is qualitatively different if the strength of (viability and/or sexual) selection is intermediate or high. For example, for $\beta\Delta_0 = 0.2$ and $\gamma\Delta_0^2 = 1.3$ (figure B.1B), there is a single non-zero stable equilibrium at low m/s, as with weak selection. Increasing m/s, say to 0.93 gives rise to two stable non-zero equilibria (marked as crosses), separated by an unstable equilibrium (in addition to an unstable equilibrium at 0). As m/s increases further (to 0.949), the stable 'high-divergence' equilibrium collides with the unstable equilibrium at a critical divergence threshold Y_c (marked by a circle).

A further increase in m/s causes this high-divergence equilibrium to vanish, so that only the alternative 'low-divergence' stable equilibrium remains. We refer to this abrupt collapse of the high-divergence equilibrium as a 'tipping point'. A further increase in m/s causes the low-divergence equilibrium to approach 0, until $Y^* = 0$ becomes the only stable equilibrium at $m_c/s \geq 1$. Thus, with intermediate selection, we observe a tipping point (at which divergence collapses abruptly from high to low but non-zero levels) at a migration threshold m^*/s that is slightly below $m_c/s = 1$, which is the critical migration threshold for complete loss of divergence.

With higher selection, say for $\beta\Delta_0 = 0.2$ and $\gamma\Delta_0^2 = 1.6$ (see figure B.1C), there is a single stable high-divergence equilibrium for $m/s < 1$. As m/s crosses 1, the equilibrium $Y^* = 0$ also becomes stable, so that there are now two stable equilibria– the zero divergence and the high-divergence equilibrium separated by an unstable equilibrium. A further increase in m/s causes the unstable and stable (high-divergence) equilibria to approach each other, until at a critical migration threshold (here, $m_c/s = 1.129$), the two equilibria collide– this occurs at a critical divergence value Y_c (marked by a circle). Beyond this migration threshold, $Y^* = 0$ becomes the only stable equilibrium. Thus, in the strong selection regime, there is a critical migration threshold ($m_c/s > 1$) at which a tipping point occurs such that divergence collapses abruptly from a high level to zero.

Below, we use equation B.14 to derive the divergence threshold (Y_c) beyond which there occurs a tipping point or sudden drop in divergence, and the critical migration threshold m_c/s, beyond which divergence goes to zero. From figures B.1B and B.1C, we can see that the critical divergence threshold Y_c (at which the high-divergence equilibrium suddenly vanishes) has the property that $f(Y_c) = 0$ and $f'(Y_c) = 0$. This gives:

$$1 - Y_c = \frac{m}{s} g(Y_c) \tag{B.15a}$$

$$1 - 2Y_c - \frac{m}{s} g(Y_c) - \frac{m}{s} Y_c g'(Y_c) = 0 \tag{B.15b}$$

Note that m/s in equation B.15 is the migration threshold associated with Y_c. This migration threshold is less than m_c/s (which is equal to 1) when assortment is weak or moderate (see figure B.1B) and the same as m_c/s under strong assortment (see figure B.1C).

Random mating: With no assortment ($\gamma = 0$), we have: $g(Y) = e^{-2\beta\Delta_0 Y}$, so that $g'(Y_c) = -2\beta\Delta_0\, g(Y_c)$, which when substituted into eq. B.15 gives: $Y_c = \frac{2\beta\Delta_0 - 1}{2\beta\Delta_0}$.

Model I of assortative mating: In this case, we have: $g(Y) = e^{-2\beta\Delta_0 Y - \frac{4}{3}\gamma\Delta_0^2 Y^2}$, so that $g'(Y_c) = (-2\beta\Delta_0 - \frac{8}{3}\gamma\Delta_0^2 Y_c)\, g(Y_c)$. Substituting this into eq. B.15b, and using eq. B.15a gives:
$\frac{8}{3}\gamma\Delta_0^2 Y_c^2 + (2\beta\Delta_0 - \frac{8}{3}\gamma\Delta_0^2) Y_c + (1 - 2\beta\Delta_0) = 0$. This can be solved to get two roots for

Y_c of which

$$Y_c = \frac{4\gamma\Delta_0^2 - 3\beta\Delta_0 + \sqrt{9(\beta\Delta_0)^2 + 16(\gamma\Delta_0^2)^2 + 24\beta\Delta_0\gamma\Delta_0^2 - 24\gamma\Delta_0^2}}{8\gamma\Delta_0^2} \tag{B.16}$$

is the biologically meaningful solution.

Model II of assortative mating: In this case, we have: $g(Y) = e^{-2\beta\Delta_0 Y}\prod_{i=0}^{\infty}\left(\frac{1+e^{-\gamma\Delta_0^2\frac{Y^2}{4^i}}}{2}\right)$, so that: $g'(Y_c)/g(Y_c) = \frac{d\log g(Y)}{dY}\Big|_{Y=Y_c} = -2\beta\Delta_0 - 2\gamma\Delta_0^2 Y_c\left[\sum_{i=0}^{\infty} 4^{-i}\left(1+e^{\gamma\Delta_0^2\frac{Y_c^2}{4^i}}\right)^{-1}\right]$. Substituting this into eq. B.15b and using eq. B.15a gives the following equation for Y_c:

$$Y_c = 1 - \frac{1}{2\beta\Delta_0 + 2\gamma\Delta_0^2 Y_c\left[\sum_{i=0}^{\infty} 4^{-i}\left(1+e^{\gamma\Delta_0^2\frac{Y_c^2}{4^i}}\right)^{-1}\right]} \tag{B.17}$$

which is solved numerically by approximating the infinite sum by the first 20 terms (which is sufficient for convergence).

For both models of assortative mating, Y_c can be substituted in equation B.15a to get m^*/s, which is the migration threshold associated with the tipping point. This is the same as m_c/s (at which divergence is completely lost) under strong selection (e.g., in figure B.1C). However, under more moderate selection, tipping points occur at $m_*/s < m_c/s$, where $m_c/s = 1$ (e.g., in figure B.1B). Thus, we have: $m_c/s = \text{Max}[1, m^*/s]$. These predictions for Y_c and m_c/s are depicted in figure 3.3A-D of chapter 3.

We can also use the above predictions for Y_c and m_c/s to compute the minimum strength of sexual selection required for: (a) tipping points in divergence (associated with a non-zero Y_c), and (b) shifted critical migration thresholds, i.e., $m_c/s > 1$, given a certain strength of viability selection. These predictions are shown in figure 3.3E-F of chapter 3 . With random mating, $Y_c > 0$ requires $\beta\Delta_0 > 1/2$ and is also associated with $m_c/s > 1$ (also see Sachdeva 2022). Put another way, if the strength of viability selection exceeds $\beta\Delta_0 = 1/2$, then we have $Y_c > 0$ and $m_c/s > 1$ even with zero sexual selection. Thus, we will focus on the parameter regime with $\beta\Delta_0 < 1/2$.

Under Model I of assortative mating, $Y_c > 0$ requires that the term under the square root in equation C be positive. This gives the minimum strength of sexual selection required to observe a tipping point (given $\beta\Delta_0$) as:

$$\frac{3}{4}(1 + \sqrt{1 - 2\beta\Delta_0} - \beta\Delta_0) \tag{B.18}$$

To find the minimum strength of sexual selection required to shift the critical migration threshold beyond 1, we substitute $m/s = 1$ into eq. B.15a, which gives $g(Y_c) = 1 - Y_c$.

If we now substitute the expression for Y_c under Model I (as given by eq. B.16), then we obtain an equation that only involves $\beta\Delta_0$ and $\gamma\Delta_0^2$. This can be solved numerically to obtain the value of $\gamma\Delta_0^2$ at which m_c/s just exceeds 1, for any given value of $\beta\Delta_0$. Obtaining similar predictions under Model II is non-trivial. Thus, in this case, we numerically solve equation B.17 to find the minimum value of $\gamma\Delta_0^2$ at which $Y_c > 0$ and similarly for $m_c/s > 1$. These predictions are depicted in figure 3.3E-F of chapter 3.

B.5 Dependence of divergence per locus on the number of loci

The theoretical predictions in equations 3.3 and 3.4 of chapter 3 (see also section B.1) are independent of L, the number of trait loci, and depend on only three composite parameters— the rate of migration relative to per locus selection m/s (where $s = \beta\alpha$), the strength of viability selection on migrants (as quantified by $\beta\Delta_0$) and the strength of sexual selection on migrants (as quantified by $\gamma\Delta_0^2$). These predictions are correct to first order in small parameters m, s etc., and are thus expected to become increasingly inaccurate as selection per locus becomes stronger. In particular, for a given strength of viability selection $\beta\Delta_0$ and sexual selection $\gamma\Delta_0^2$ (where $\Delta_0 = \alpha L$ is the maximum possible trait divergence between mainland and island), we expect our predictions to become more accurate as we consider traits determined by larger and larger number of loci of weaker effect, i.e., as we approach the limit $\alpha \to 0$, $L \to \infty$ with $\Delta_0 = \alpha L$ held constant.

We now verify these expectations by comparing theoretical predictions with the results of deterministic simulations of traits influenced by $L = 10, 40, 100$ loci, for a specific strength of viability and sexual selection ($\beta\Delta_0 = \gamma\Delta_0^2 = 1$) for the two models of assortative mating. In order to keep the maximum possible divergence $\Delta_0 = \alpha L$ (as well as $\beta\Delta_0$ and $\gamma\Delta_0^2$) constant, we simultaneously decrease α while increasing L. For small L, say $L = 10$, the effect of LD at any locus is weaker than what is predicted by our expressions, so that divergence is lost at lower values of m/s. However, as expected, there is better agreement between simulations and theory with increasing L, so that m_c/s for a trait with $L = 100$ loci only slightly lower than the analytical prediction (compare purple lines with black line in figure B.2). More generally, we observe larger deviations from theory when $\beta\Delta_0$ and/or $\gamma\Delta_0^2$, L are large. Thus, if net selection is strong, then L must be correspondingly large to observe good agreement between theory and simulations.

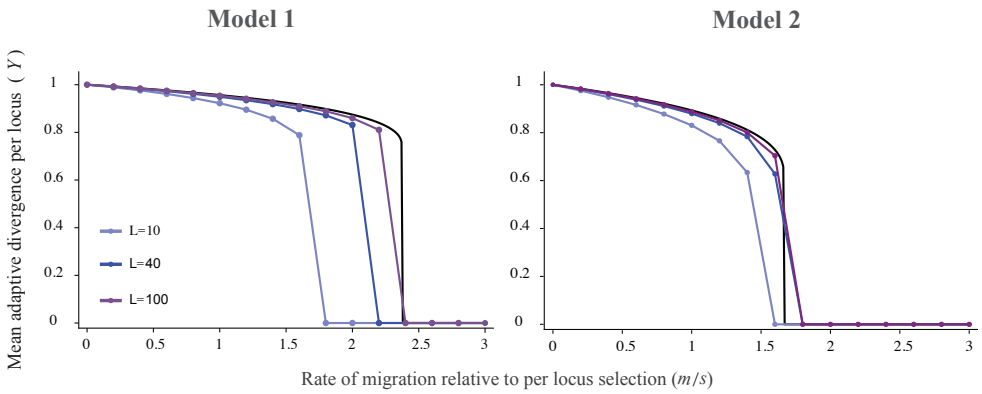


Figure B.2: Mean allele frequency divergence per locus (Y) between mainland and island vs. migration rate relative to per locus selection (m/s), for Models I and II of assortative mating, with $\beta\Delta_0 = \gamma\Delta_0^2 = 1$ in both panels. Different colours show the results from the deterministic simulation with $L = 10$, 40, and 100 loci, where α is decreased as L is increased to keep $\Delta_0 = \alpha L$ constant. The black curve shows the analytical (large L or, equivalently, small s) predictions for the two models, obtained by numerically solving eq. 3.4 in chapter 3.

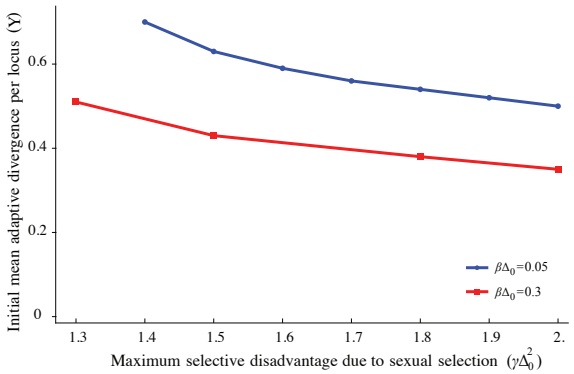


Figure B.3: Minimum initial allele frequency divergence per locus required for populations to evolve towards the high-divergence equilibrium, plotted against the strength of sexual selection ($\gamma\Delta_0^2$) for Model I of assortative mating. Different colours show results for two different strengths of viability selection $\beta\Delta_0 = 0.05$ and 0.3. All results are obtained from deterministic simulations with $L = 40$ loci.

B.6 Dependence of mean divergence on initial divergence levels.

Throughout chapter 3, we have considered a secondary contact scenario with high initial divergence between mainland and island. However, under intermediate or strong assortment, where two stable equilibria are possible (marked by crosses in figure B.1B and C), the island population can evolve towards either equilibrium under gene flow, depending on initial conditions. More specifically, starting with high initial divergence ($Y \sim 1$), the population approaches the stable equilibrium associated with the higher level of divergence, provided $m < m_c$. By contrast, if Y is initially low, the population

approaches the low (or zero) divergence equilibrium. In other words, in this regime, gene flow is strong enough to prevent genetically similar populations from diverging (despite divergent selection), but not strong enough to erode divergence in populations that are already strongly diverged.

Figure B.3 explores how equilibrium divergence depends on initial divergence using deterministic simulations initialised with different values of Y, the mean per locus allele frequency divergence. Note that under the hypergeometric model (where all trait loci are exchangeable), this amounts to assuming an allele frequency divergence Y at each locus. The population is then allowed to equilibrate under gene flow. We consider assortment levels at which alternative stable equilibria exist, so that populations can evolve towards either equilibrium, depending on the initial divergence Y. Figure B.3 shows the minimum initial divergence per locus required for populations to attain the 'high-divergence' equilibrium, as a function of $\gamma\Delta_0^2$, the strength of sexual selection, for two different values of $\beta\Delta_0$. As expected, long-term divergence can be maintained more easily, i.e., even with moderate levels of initial divergence, if net (viability or sexual) selection is strong.

References

Barton, N. (1992). On the spread of new gene combinations in the third phase of Wright's shifting- balance. Evolution, pages 551–557.

Barton, N. H., Etheridge, A. M., and Veber, A. (2017). The infinitesimal model: Definition, derivation, and implications. Theoretical Population Biology, 118:50–73.

Sachdeva, H. (2022). Reproductive isolation via polygenic local adaptation in sub-divided populations: Effect of linkage disequilibria and drift. PLOS Genetics, 18(9):e1010297.

Westram, A. M., Stankowski, S., Surendranadh, P., and Barton, N. (2022). What is reproductive isolation? Journal of Evolutionary Biology, 35(9):1143–1164.

Appendix C

SUPPLEMENTARY INFORMATION for Genetic Analysis of flower colour clines in *Antirrhinum majus*

C.1 Supplementary Text

C.1.1 Finding the 1D polynomial transect

To perform cline fitting, we converted the 2D spatial positions of the demes into 1D by fitting a polynomial transect of degree n, $y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n$. We fit polynomials of different degrees to the data. The curve satisfying the polynomial of degree 6 best fits the data (by calculating deviations between the position on the curve and the observed position). For each deme, we find the point on the curve that minimises the distance between the deme and the curve. Further, the distance of the point along the curve was obtained as $z = \int_0^x \sqrt{1 + (\frac{\partial y}{\partial x})^2}dx$, where x is the position of the deme on the curve.

C.1.2 F_{IS} at different clustering radii

To examine deviations from Hardy Weinberg equilibrium and choose a spatial scale that minimised departures from random mating, the maximum likelihood (ML) estimate of F_{IS} is found at different clustering scales for the clinal SNPs (excluding *Eluta*, being linked with *Rosea*) and the 91 non-clinal SNPs (as in chapter 4 Surendranadh et al, 2022), assuming a uniform value across all demes. F_{IS} depends on the allele frequencies (p, q) and is bounded by $\text{Max}[\{-p/q, -q/p\}, 1]$. F_{IS} increases with deme size due to the Wahlund effect, with similar slopes for both clinal and non-clinal loci as expected from isolation by distance. Heterozygote deficit was greater for clinal SNPs when compared to the mean

from non-clinal SNPs (coloured vs black curves in Fig. C.2). The magnitude of F_{IS} differs between clinal loci and is somewhat consistent with the difference in cline width; *Rosea* shows the largest heterozygote deficit, and Rubia together with the non-clinal SNP, is the smallest. We see some decline in F_{IS} over small scales probably due to greater small-scale local fluctuations at loci with steeper clines, or assortative pollination for flower colouration. However, at the 25m scale, F_{IS} is close to 0 for all loci except *Rosea* (see Fig C.2). Additionally, multilocus heterozygosity from 91 non-clinal SNPs was consistent with matings from parents 10m apart (see Appendix C.3 Fig. D.1). Taken together, 25m is a reasonable choice for the clustering scale.

C.1.3 Evidence of long-range migrants in the flanks

For each individual in the yellow and magenta flank, the full genotypes are examined at five unlinked loci, to identify recent migrants and their offspring. The individuals are assumed to either come from a local population (i.e deme in which the individual belongs), which is assumed to be well-mixed, or from a well-mixed distant source (i.e. the magenta flank if the focal individual comes from the yellow flank and vice versa) or are F1 or backcross hybrids between these sources.

The probability $P_{genotype}$ of a diploid genotype in a native population that is in Hardy-Weinberg proportions and linkage equilibrium (HWLE) is the product of allele frequencies across loci, with factors $\{q_i^2, 2p_iq_i, p_i^2\}$ according to diploid genotype (denoted $X_i = 0$,1 or 2). We estimate $log(P_{genotype})$ as a sum over loci, and for missing values, add the average $q_i^2 log(q_i^2) + 2p_iq_i log(p_iq_i) + p_i^2 log(p_i^2)$ (4.1% of values are missing at these loci). This extends in a simple way to F1s and backcross hybrids derived from a cross with a source population with allele frequencies p_i^*. The frequencies of the three diploid genotypes are now $\{q_iq_i^*, q_ip_i^* + p_iq_i^*, p_ip_i^*\}$, and the net probability is again a product over loci. As before, we calculate $log(P)$, add the expected contribution when there is a missing value, and account for the error rates in this procedure. Note that this algorithm does capture the linkage disequilibrium generated by admixture, even though it is based on a product of factors across loci, and extends to individuals from any two locations.

For each flank, we first identified individuals with $P_{genotype} < 2 * 10^{-4}$ of coming from their location, with $P_{genotype}$ calculated as described above. Here, the allele frequency is based on the best estimate from the cline fitting for that location (see 'Cline fitting for the genetic data'). For the 'improbable individuals' that are unlikely to come from the local populations, we generated the probability that it came from the opposite flank, or is an F1 or backcross derived from that source. The allele frequency of the source population is calculated directly from the genotypes of all individuals in that flank. Table C.17 shows the improbable individuals from the yellow and magenta flanks.

C.2 Supplementary Tables

δ (Cluster Radius)	Number of Demes	Mean Size	SD of Size	Minimum Size	Maximum Size
10	10	2352	10	22	266
15	15	1640	14	34	362
20	20	1279	18	47	486
25	25	999	23	60	540
30	30	858	26	69	711
40	40	682	33	92	1062
50	50	564	40	116	1444
75	75	381	59	169	2063
100	100	303	74	216	2389
120	120	258	87	254	2503
150	150	223	101	293	2692
175	175	195	115	324	2841
200	200	177	127	366	3111
250	250	132	170	508	3355
300	300	117	192	605	4026

Table C.1: Number of demes, mean size, standard deviation of size, minimum size, and maximum size of the demes corresponding to each cluster radius δ. The size of a deme denotes the number of individuals within a deme.

Models	dof	Rosea	Eluta	Sulfurea	Flavia	Rubia	Cremosa
1 and 3	4	483.849	700.51	401.41	363.93	153.07	330.16
2 and 3	2	86.76	19.64	43.01	176.41	41.32	83.32

Table C.2: Test statistic from the likelihood ratio test for comparison between models 1 (sigmoid), 2 (sigmoid with flanking polymorphism), and 3 (stepped) cline models. Models 1 and 3 differ by 4 degrees of freedom (dof) and models 2 and 3 by 2 degrees of freedom. The critical values associated with the 95% confidence intervals are 9.448 and 5.991 respectively. The more complex model is chosen since the test statistic is greater than the critical value.

Time Point	B_0	B_1	θ_0	θ_1	w (m)	y (m)	F_{ST}
2009-2012	7887.1 (9941.2 – 55445.6)	7003.5 (1582.7 – 29970.7)	0.0017 (0.0002 – 0.0038)	0.0249 (0.0042 – 0.0960)	861.98 (726.8 – 1014.6)	14932.5 (14894.5 – 14961.3)	0.1156 (0.096 – 0.147)
2013-2015	5220.2 (9417.22 – 56324)	8064.24 (2290.39 – 36674.7)	0.0026 (0.0003 – 0.0047)	0.0242 (0.0050 – 0.0617)	758.01 (629.301 – 854.07)	14923.8 (14898.6 – 14950.2)	0.0796 (0.0670 – 0.1016)
2016-2019	38422.6 (2725.2 – 287780)	25539.5 (1990.45 – 276979)	0.0004 (7.22×10^{-6} – 0.0019)	0.0024 (0.0000396 – 0.0495)	785.00 (600.03 – 926.71)	14923.4 (14896.0 – 14951.7)	0.0883 (0.0698 – 0.1167)

Table C.3: MLE together with the 95% confidence interval (CI) for the parameters from the asymmetric cline model for ROSEA at three time points: 2009-2012, 2013-2015, and 2016-2019.

Interval (km)	Number of demes	Total Number of individuals
0 to 12.5	9	146
12.5 to 13.25	29	580
13.25 to 13.75	25	621
13.75 to 14.25	28	992
14.25 to 14.5	33	1420
14.5 to 14.75	29	1473
14.75 to 15	37	3714
15 to 15.25	39	4322
15.25 to 15.75	42	5710
15.75 to 16.25	27	786
16.25 to 18.5	27	632
> 18.5	12	211

Table C.4: 1D transect divided into different bins (or intervals) with the number of demes (with at least 10 individuals) in each bin and the total number of individuals in all demes.

SNP	Mean F_{IS}	Mean $F_{IS,null}$	Fraction of demes with significantly positive F_{IS}	Fraction of demes with significantly negative F_{IS}
Rosea*	0.034	-0.021	0.083	0
Sulfurea	-0.018	-0.022	0.036	0.024
Flavia	-0.016	-0.021	0.032	0.019
Rubia	-0.053	-0.035	0.012	0.03
Cremosa	-0.008	-0.022	0.046	0.012

Table C.5: Mean F_{IS} across demes at each locus compared with the mean from 100 shuffled replicates ($F_{IS,null}$) at 25m cluster diameter. The latter represents a null model wherein each deme is at HW equilibrium. The last two columns show the fraction of significantly positive and negative F_{IS} as compared to the null model. Significance is based on p-value obtained for each deme, which is the fraction of replicates that gives F_{IS} greater than observed (when $F_{IS} > 0$) and less than observed otherwise. Significant heterozygote deficit is observed only at Rosea (marked by an asterisk) with more than 5% demes with observed F_{IS} greater than mean $F_{IS,null}$.

Interval (km)	Mean F_{IS}	sd(F_{IS})	Mean $F_{IS,null}$	sd($F_{IS,null}$)	P_{mean}	P_{sd}	Mean $\hat{R}$	sd($\hat{R}$)	Mean $\hat{R}_{null}$	sd($\hat{R}_{null}$)	P_{mean}	P_{sd}
0 to 12.5	-0.111	0.09	-0.04	0.132	0.948	0.953	0.054	0.131	-0.000262	0.055	0.005	0.004
12.5 to 13.25	-0.011	0.116	-0.03	0.101	0.144	0.164	0.086	0.142	0.0000291	0.065	0	0
13.25 to 13.75	0.015	0.168	-0.028	0.097	0.01	0.001	0.128	0.193	0.0000732	0.069	0	0
13.75 to 14.25	-0.032	0.084	-0.022	0.09	0.713	0.654	0.019	0.087	-0.0002	0.058	0.044	0.002
14.25 to 14.5	-0.049	0.143	-0.034	0.107	0.798	0.01	0.02	0.092	0.000136	0.080	0.089	0.179
14.5 to 14.75	-0.008	0.114	-0.018	0.081	0.25	0.013	0.027	0.068	-0.00062	0.061	0.013	0.24
14.75 to 15	-0.024	0.116	-0.028	0.093	0.353	0.052	0.049	0.076	0.00042	0.06	0	0.006
15 to 15.25	-0.0002	0.090	-0.016	0.079	0.092	0.163	0.029	0.074	-0.00052	0.062	0.003	0.114
15.25 to 15.75	0.020	0.087	-0.015	0.075	0.003	0.153	0.017	0.059	0.00026	0.056	0.041	0.361
15.75 to 16.25	-0.014	0.104	-0.024	0.098	0.303	0.334	0.02	0.082	-0.000095	0.059	0.047	0.021
16.25 to 18.5	-0.032	0.143	-0.03	0.112	0.526	0.055	0.052	0.133	-0.00054	0.064	0	0
> 18.5	-0.038	0.078	-0.032	0.091	0.559	0.619	0.006	0.055	-0.0004	0.045	0.304	0.171

Table C.6: F_{IS} and LD across hybrid zone at 25m cluster diameter: Mean F_{IS} across 5 unlinked loci and correlations across 10 unlinked locus pairs for different intervals. The table also shows the standard deviations (sd) within each bin, mean and sd from the null distribution for F_{IS} and LD obtained from 100 shuffled replicates. P_{mean} and P_{sd}refer to the p-value comparing the means and sd respectively between the observed and null distribution.

SNP 1	SNP 2	Mean $\hat{R}$	Mean $\hat{R}_{null}(*10^{-4})$	Fraction of demes with significantly positive $\hat{R}$	Fraction of demes with significantly negative $\hat{R}$
Rosea	Sulfurea	0.049	2.15	0.063	0.021
Rosea	Flavia	0.089	-1.14	0.141	0.183
Rosea	Rubia	0.029	6.33	0.082	0.034
Rosea	Cremosa	0.064	1.41	0.105	0.012
Sulfurea	Flavia	0.022	0.70	0.069	0.03
Sulfurea	Rubia	-0.002	-1.83	0.045	0.048
Sulfurea	Cremosa	0.031	2.14	0.065	0.039
Flavia	Rubia	0.035	1.82	0.073	0.033
Flavia	Cremosa	0.066	-0.74	0.121	0.027
Cremosa	Rubia	0.031	-1.9	0.045	0.039

Table C.7: Mean correlations $\hat{R}$ between each pair of loci across demes compared with the mean from 100 shuffled replicates ($\hat{R}_{null}$) at 25m cluster diameter. Shuffling is done to remove any associations between loci but preserves observed Hardy-Weinberg deviations. The last two columns show the fraction of demes that show significant positive and negative correlations compared to the null. All locus pairs have significantly positive observed correlations except for *Sulfurea-Rubia* and *Cremosa- Rubia*, while no locus pairs show significantly negative correlations.

C.3 Supplementary Figures

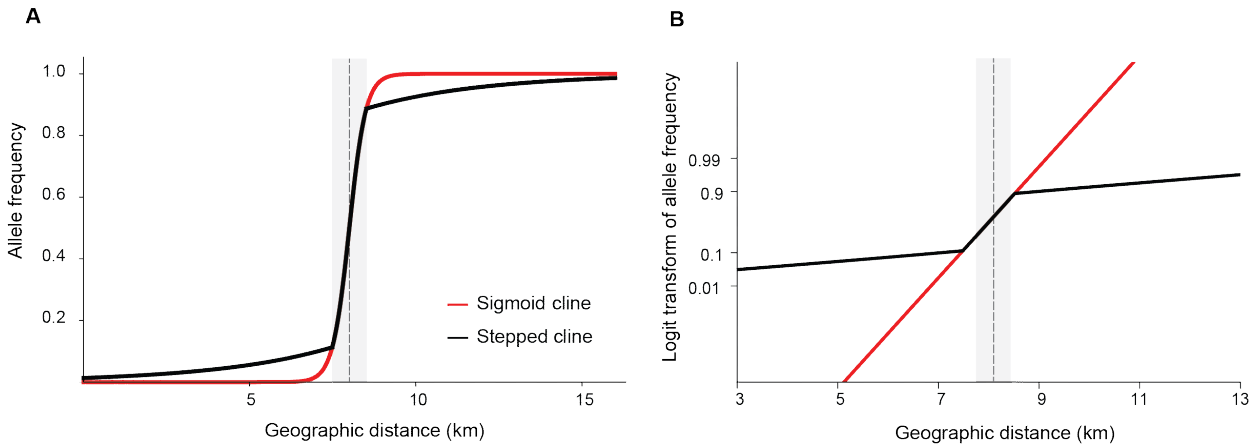


Figure C.1: A sigmoid (red) and stepped (black) cline shapes of the same width (1 km) and centre (8 km) plotted on normal scale (A) and logit transformed (B). The dotted line represents the cline centre and the grey bar denotes the cline width. (A) A stepped cline has a sigmoid shape in the centre and overlaps with the sigmoid cline of the same width in the centre. (B) Logit transformation plots $\log(p/q)$, where p is the allele frequency and $q = 1 - p$. Here, the sigmoid shape appears as a straight line and the step is easily distinguishable.

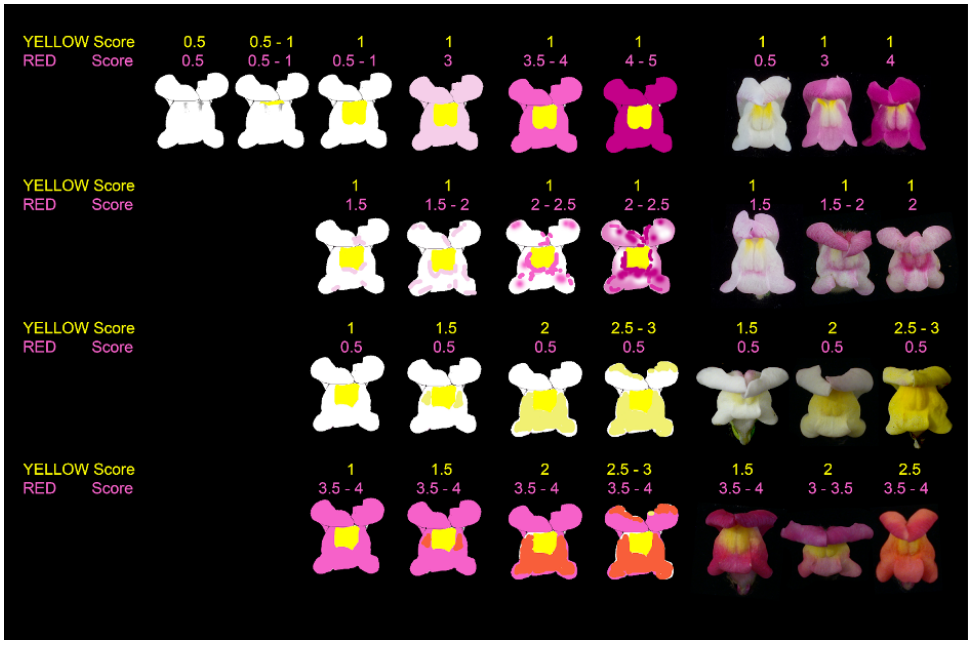


Figure C.2: Flower scoring guide used to score the magenta and yellow flower colouration. For each flower, magenta and yellow flower scores range from 0.5 to 5 or 3 depending on the intensity of the pigmentation and how it spreads across the flower.

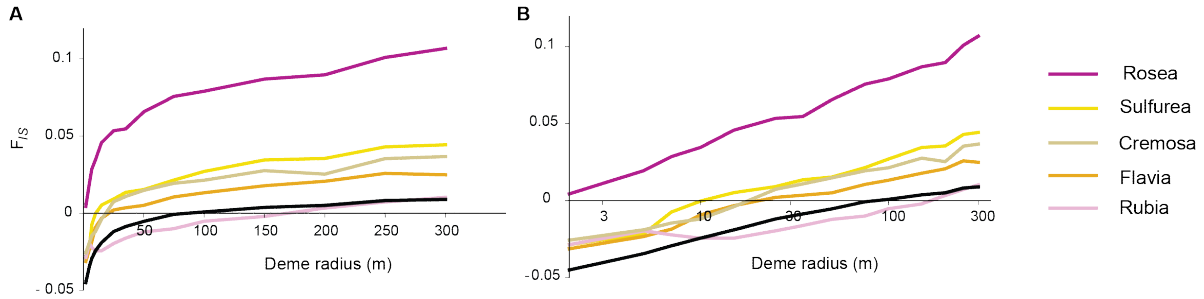


Figure C.3: Mean heterozygote deficit (F_{IS}) across 999 demes for different deme radii shown for each clinal loci on the normal scale (A) and (B) log scale. The mean F_{IS} from 91 non-clinal SNPs is shown in black for comparison.

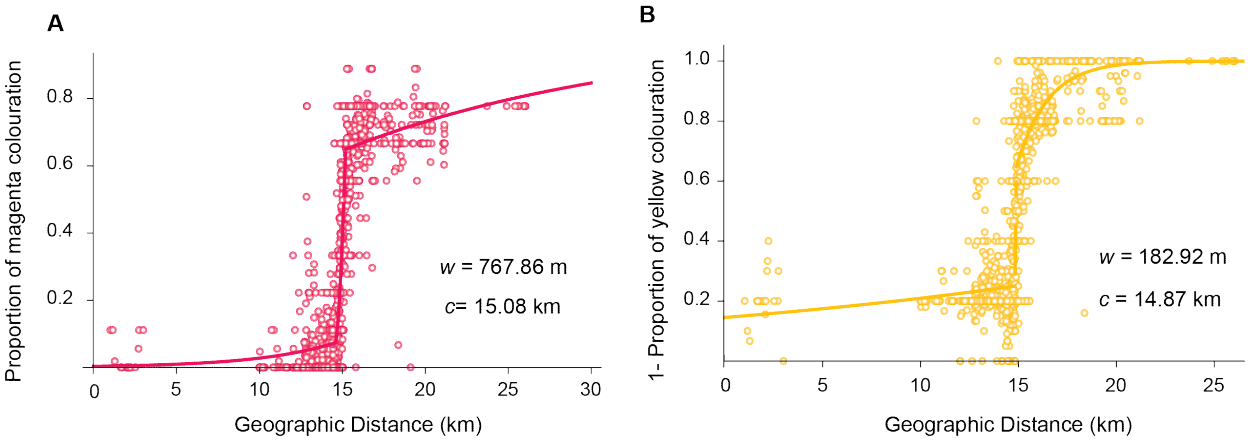


Figure C.4: Phenotypic clines for the mean (A) magenta and (B) yellow flower colour scores across 999 demes in the hybrid zone. The maximum likelihood estimate of cline width (w) and cline centre (c) is shown for each fitted cline. Panel B plots 1- yellow colour score to compare the magenta and yellow clines. The open circles denotes the mean colour scores in each deme and the coloured curve represents the best fit cline from Metropolis-Hastings algorithm. The phenotypic cline shape is stepped for both magenta and yellow colouration. The mean magenta and yellow colour scores of demes range from 0 to 0.88 and 1 respectively. The cline width is narrower for the yellow colour score while the cline centres are similar and differ by only 200m. The magenta colour scores are more symmetric in the right and left tails but highly asymmetric for the yellow colour scores.

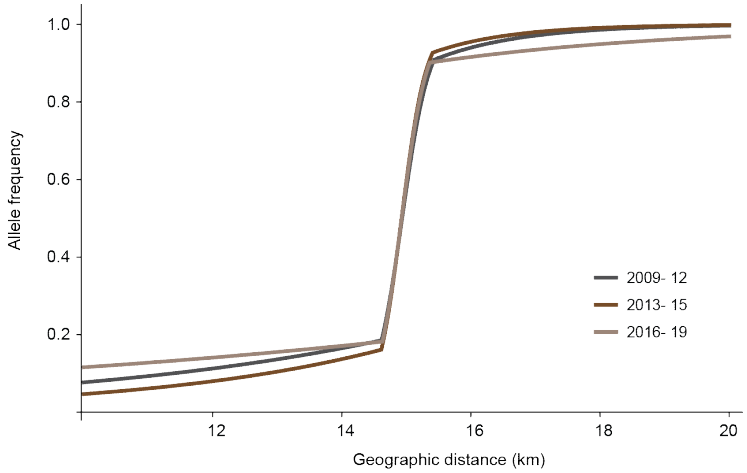


Figure C.5: Temporal clines in Rosea: clines are fitted for Rosea across 3 time points 2009 to 2012, 2013 to 2015 and 2016 to 2019 shown as allele frequency plotted across a smaller range from 10km to 20 km. The central cline shape is consistent across different time points with similar cline centres and cline widths (see table C.3). However, the cline shape in the right and left tails differ due to differences in the polymorphism across years.

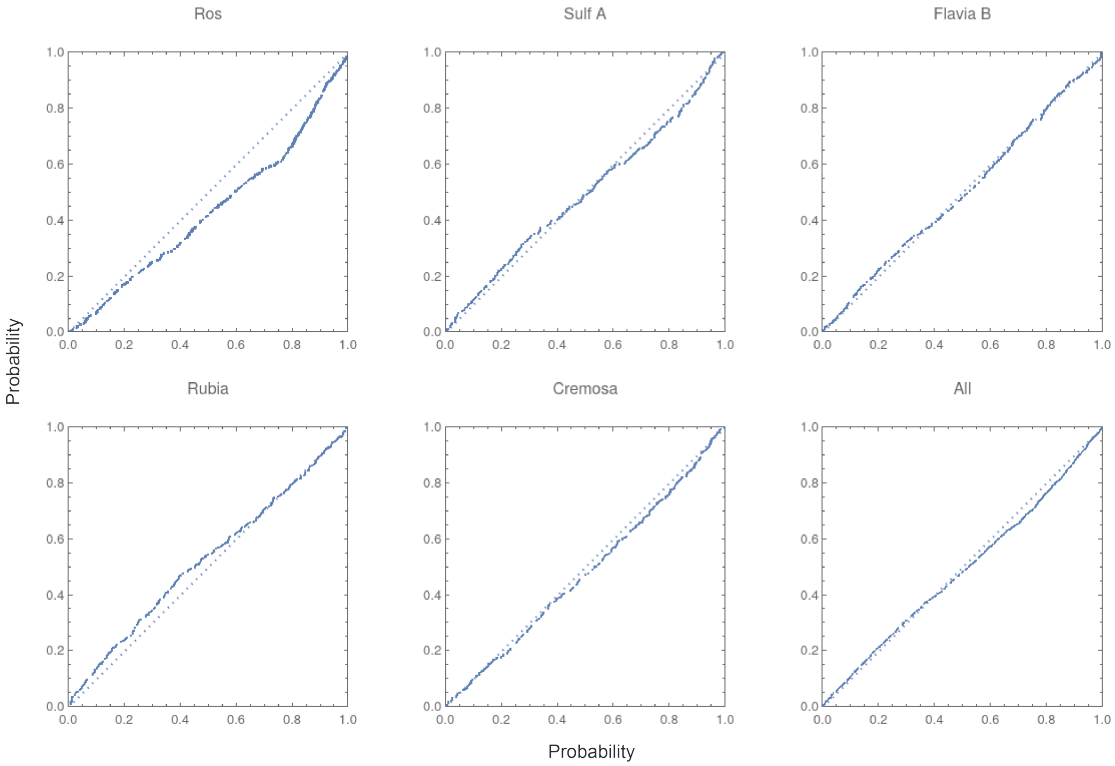


Figure C.6: Q-Q plot showing observed and expected F_{IS} for each loci and mean F_{IS} from all loci. Each plot compares the observed distribution of F_{IS} from each deme to the null distribution obtained by shuffling within each deme.

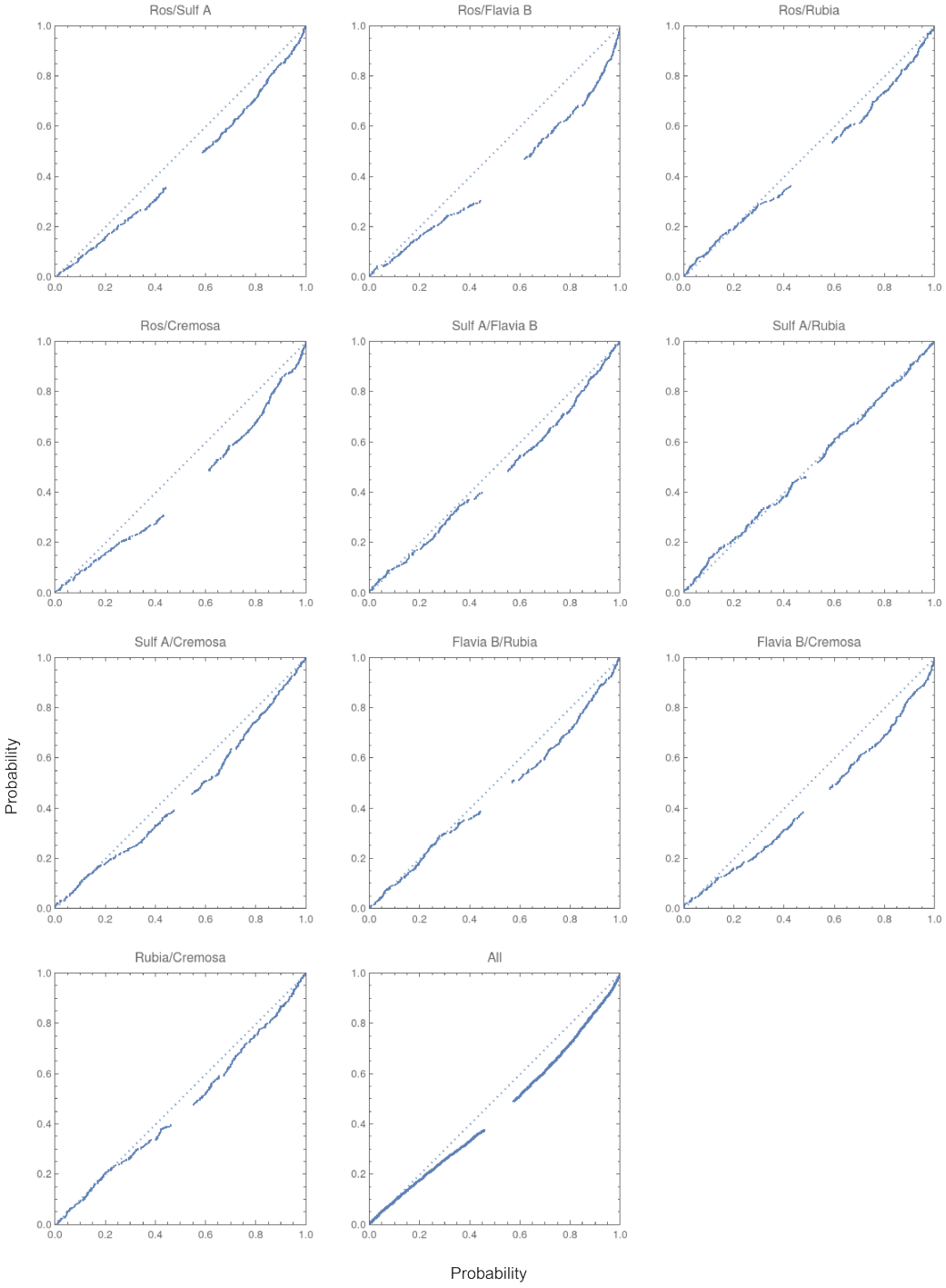


Figure C.7: Q-Q plot showing observed and expected $\hat{R}$ for each locus pair and mean $\hat{R}$ from all pairs of loci. Each plot compares the observed distribution of $\hat{R}$ from each deme to the null distribution obtained by shuffling within each deme.

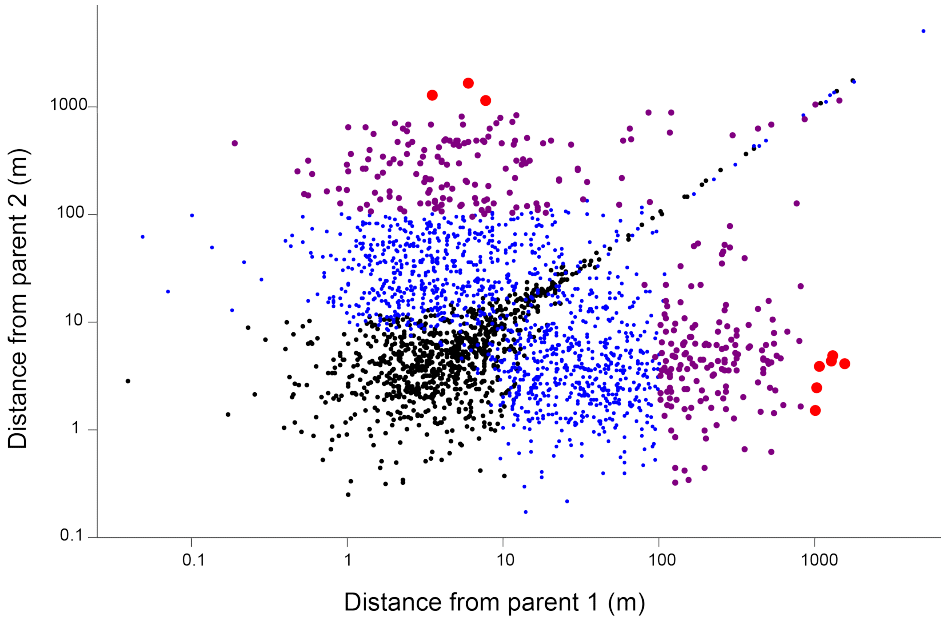


Figure C.8: Distance between an offspring and each of its parents (in X and Y axes respectively) for each of 2342 pedigree trios in log scale. Black dots represent trios where parents are within 20m, and thus mostly lie along the diagonal. Blue dots denote trios where parents are between 20 to 100m. Purple and red dots represent trios where parents are within 100-1000m and greater than 1km respectively. In both these scenarios, one of the parent-offspring distances is often much higher than the other.

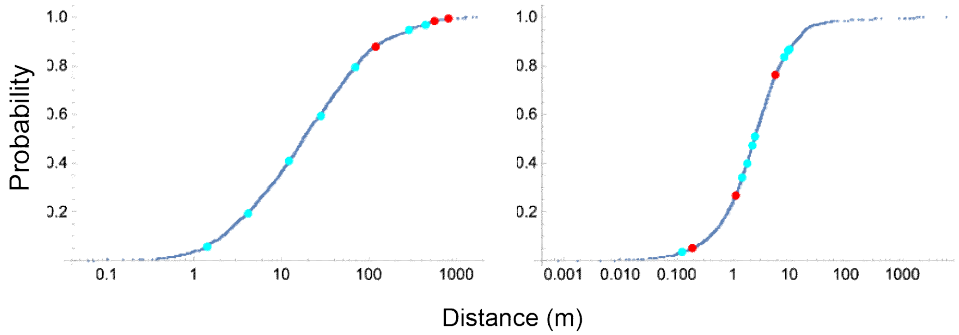


Figure C.9: The CDF of pollen dispersal distance, for all the trios (blue), for the 8 trios with 3 errors (cyan), and the 3 trios with 4 errors (red). The former are uniformly distributed, but the latter are in in the upper tail (P=88.1%, 98.5%, 99.5%), suggesting that these are mis-assigned. For seed dispersal (estimated as distance to the nearest parent; right) there is no suggestion of excess distance for individuals with more Mendelian incompatibilities.

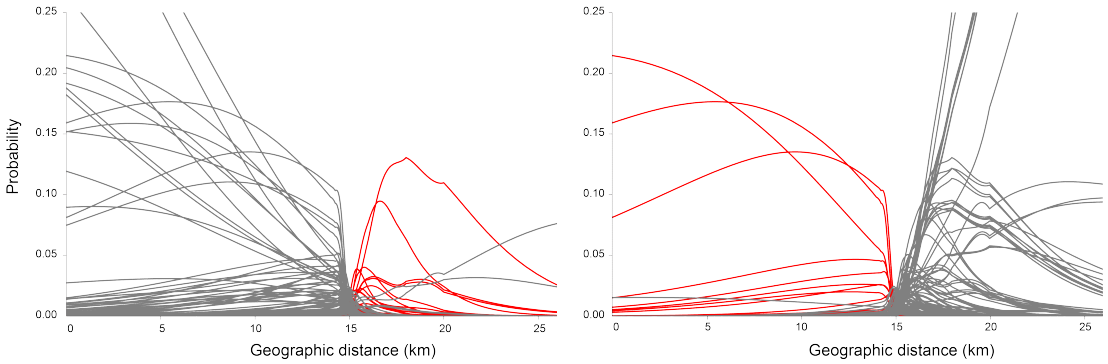


Figure C.10: Marginal probability of finding the mother of an individual at a given distance plotted along the transect. Each curve represents an individual in the yellow flank (left panel) and magenta flank (right panel) respectively. Grey curves denote individuals in the flanks likely to come from a mother on the same side. Red curves denote individuals in the flanks likely to come from a mother on the opposite flank.

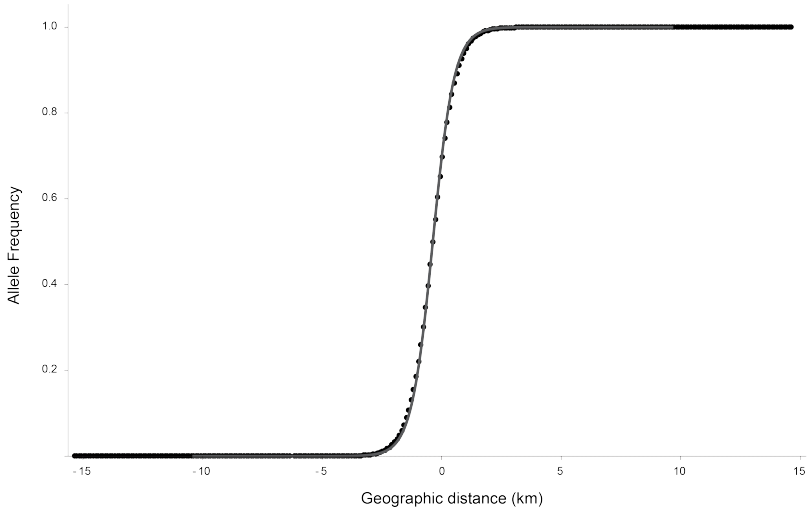


Figure C.11: Allele frequencies versus geographic distance (km) from single locus cline simulations considering Gaussian dispersal with σ= 200m and heterozygote disadvantage with $s = 0.1$. The black dots denote the simulated allele frequencies in the demes and the curve represents the allele frequencies from the expected sigmoid cline of width $\sqrt{8}\sigma/\sqrt{s}$.

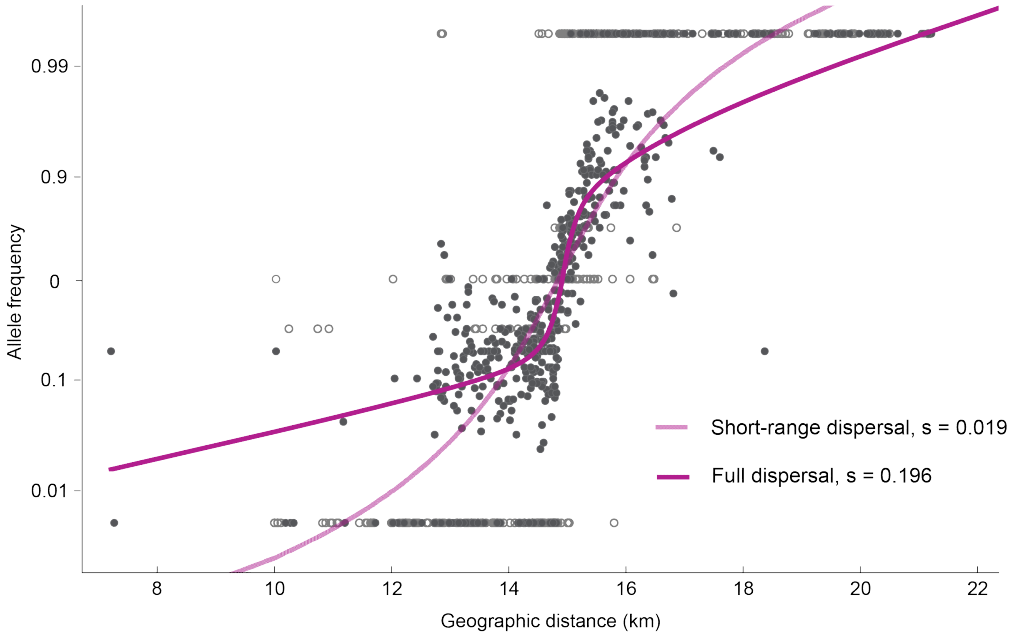


Figure C.12: Simulated clines at *Rosea* plotted on logit transformation (i.e allele frequency p is transformed into $p/(1-p)$ and shown on a log scale). ML estimates of selection are inferred from two simulations, one considering short-range dispersal (i.e inferred only from the trios) where $s = 0.019$ (in light magenta) and with the full dispersal kernel where $s = 0.196$ (in dark magenta). The former fails to produce the stepped cline as observed. Simulated cline with the inferred dispersal generates a stepped shape and captures patterns of allele frequency in the tails.

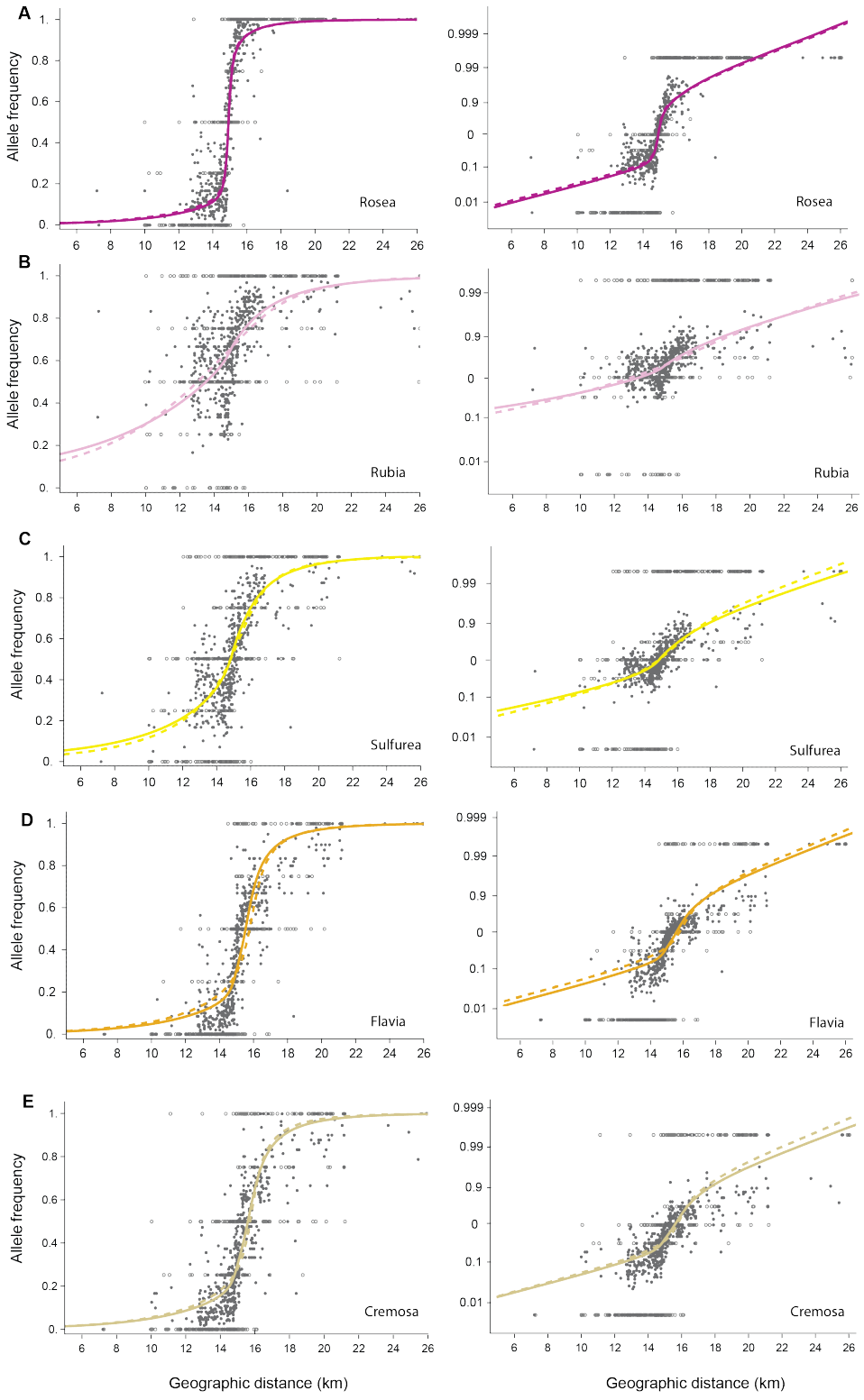


Figure C.13: Best fit cline shapes from single locus simulations (dashed) and multilocus simulations with asymmetric selection (solid) at each of the 5 unlinked loci. The left column plots it in a normal scale while the right shows the same in the logit transformed scale. The corresponding selection estimates are shown in Table 5.3.

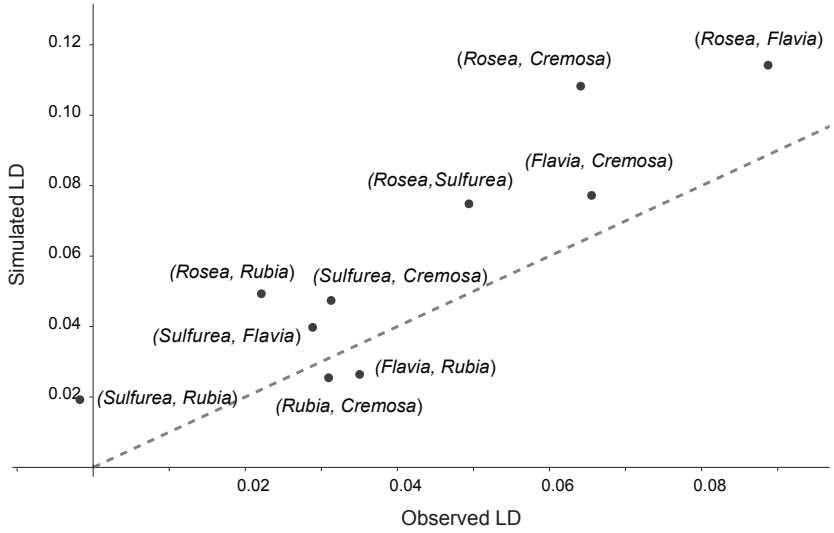


Figure C.14: Comparison of LD between 10 pairs of unlinked loci from multilocus simulations with the inferred dispersal vs that observed from the data.

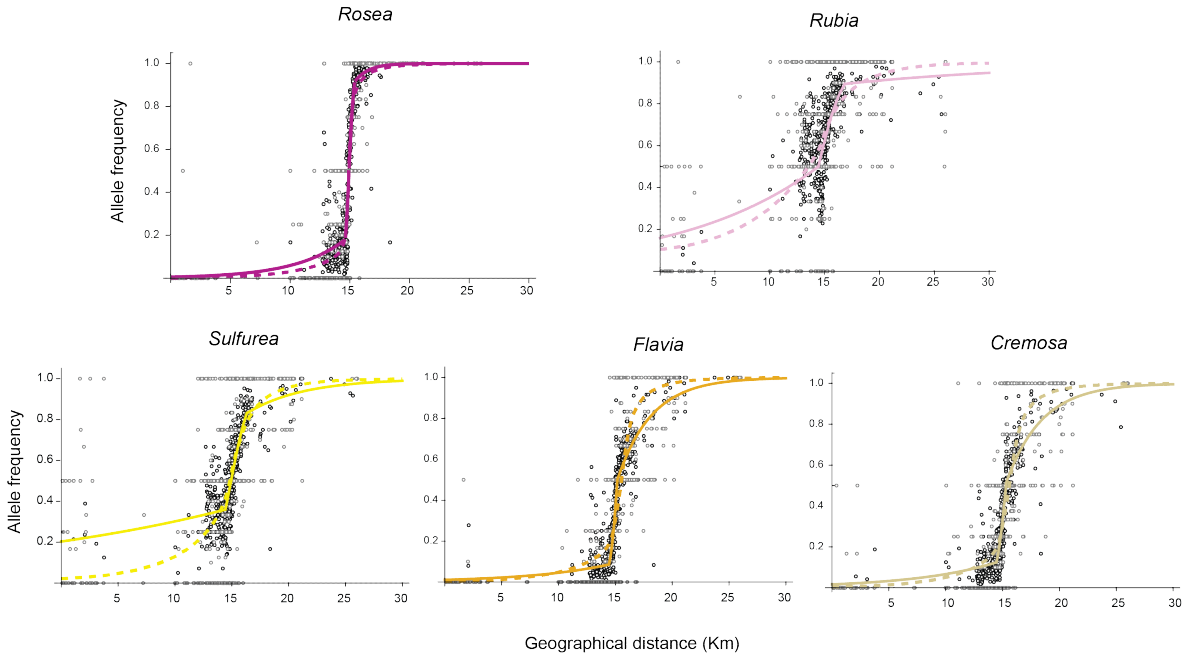


Figure C.15: Best fit cline shapes from multilocus simulations with asymmetric selection (solid) vs from descriptive cline fitting (dashed) at each of the 5 unlinked loci. The corresponding selection estimates are shown in Table 5.3.

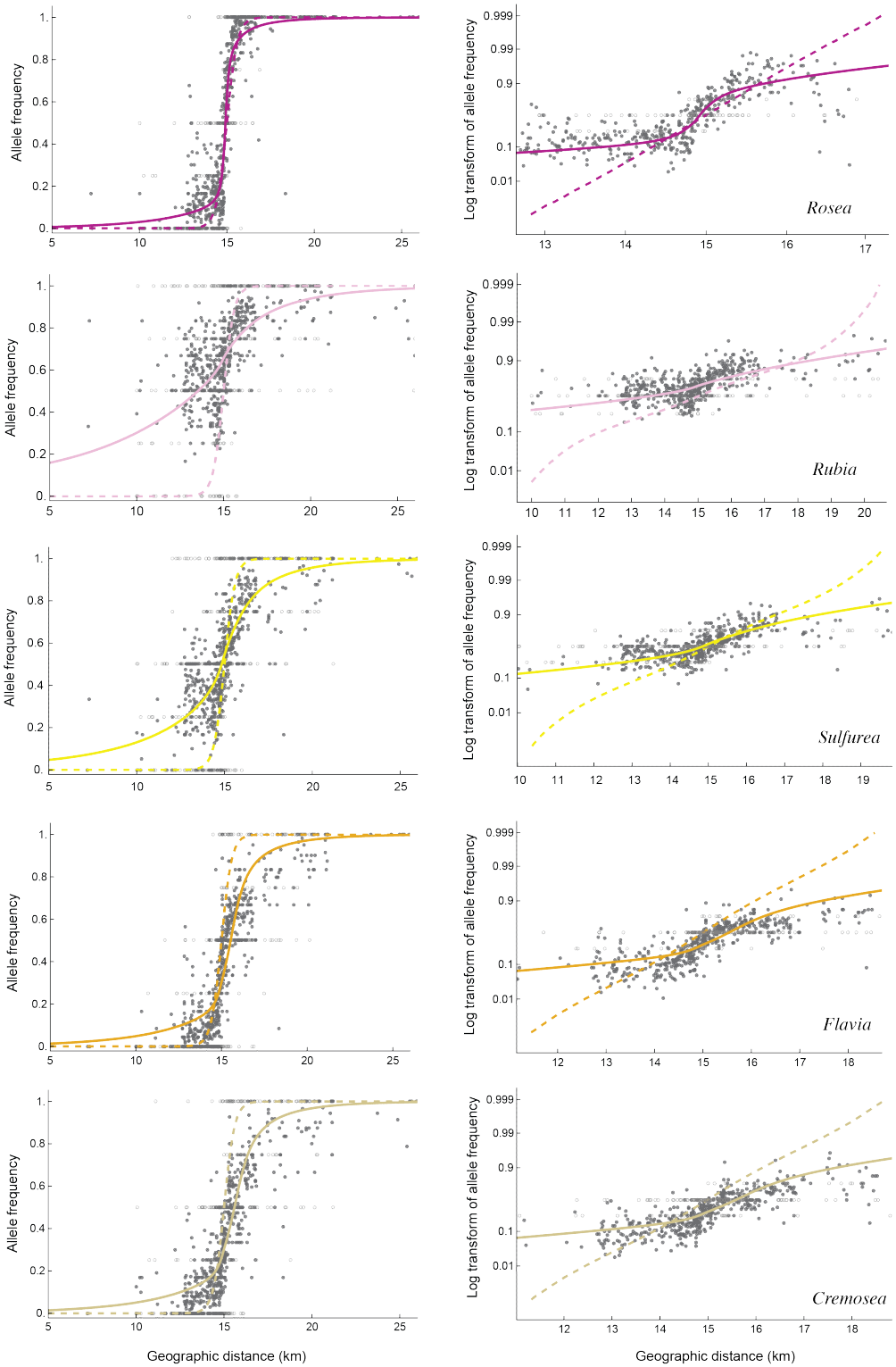


Figure C.16: Best fit cline shapes from multilocus simulations with inferred long range dispersal (solid) vs Gaussian dispersal with σ =161m (dashed) at each of the 5 unlinked loci. The left column plots it in a normal scale while the right shows the same in the logit transformed scale. The latter shows that multilocus simulation with Gaussian dispersal does not produce stepped clines, suggesting weak LD at the cline centre.

index	genotype	R / Y / V	Z	$\log_{10}(P_0)$	$\log_{10}(P_{Bx1})$	$\log_{10}(P_{F1})$	$\log_{10}(P_{opp})$
"S6684"	{2, 2, 1, 2, 0}	{0.5, 2.5, 1}	1646.62	-8.56024	-5.01054	-4.38194	**-1.83974**
"P0474"	{*1, 1, 1, 2, 1*}	{1.5, 2.5, *}	10 722.9	-3.99329	-1.9702	**-1.03414**	-2.77608
"M3104"	{*1, 1, 2, 1, 1*}	*	12 046.8	-4.65039	-2.82286	**-2.02388**	-3.21678
"J1675"	{*1, 2, 2, 1, 1*}	{0.5, 1, *}	12 812.8	-5.02378	-3.18803	**-2.33861**	-2.75338
"M1198"	{2, 0, 1, 2, 0}	{4, 1.5, *}	12 817.8	-3.92564	**-2.97617**	-3.15966	-3.36882
"M0381"	**{2, 2, 1, 1, 2}**	{5, 1, *}	12 820.2	-6.02382	-3.94979	-3.04573	**-1.49201**
"M0373"	{2, 1, 2, 0, 2}	{4, 1, *}	12 830.8	-7.05989	-5.14305	-4.76543	**-2.99839**
"P2906"	{*1, 1, 2, 1, 0*}	{4, 1, *}	12 832.1	-3.76651	-2.68385	**-2.47669**	-3.96958
"M0395"	**{2, 2, 2, 1, 1}**	{4, 1, *}	12 832.4	-6.19668	-4.10984	-3.21749	**-1.52764**
"M0399"	**{2, 2, 1, 2, 1}**	{4, 1, *}	12 834.1	-5.16352	-3.17915	-2.21766	**-1.08694**
"L1318"	{2, 1, 0, 2, 1}	{4, 2, *}	12 834.1	-3.73422	-2.61868	-2.31636	**-2.26751**
"S4464"	{2, 1, 1, 2, 0}	{4, 0.5, 3}	12 835.8	-3.91184	-2.67664	-2.35636	**-2.30314**
"M0391"	{2, 2, 2, 1, 0}	{4, 1, *}	12 836.3	-5.53804	-4.0289	-3.6852	**-2.28044**
"M0393"	{2, 1, 1, 2, 0}	{4, 1, *}	12 836.5	-3.91161	-2.67653	-2.35627	**-2.30314**
"P2969"	{2, 1, 2, 1, 0}	{4, 1, *}	12 840.6	-4.94267	-3.60618	-3.35527	**-2.74384**
"M0354"	{1, 0, 2, 2, 0}	{2, 1, *}	12 843.6	-4.17692	**-3.23211**	-3.42623	-4.47963
"M1611"	{*2, 2, 1, 1, 1*}	{4, 1, *}	12 855.3	-4.75078	-2.92378	-2.06655	**-1.64257**
"S4446"	{*2, 1, 1, 1, 1*}	{3.5, 1, 3}	12 880.	-4.149	-2.49915	**-1.73591**	-2.10597
"L1316"	**{1, 2, 2, 1, 2}**	{1.5, 1, *}	12 887.4	-6.25639	-4.19772	-3.3071	**-2.60282**
"J0775"	**{2, 2, 2, 2, 1}**	{3.5, 1, *}	12 888.4	-6.5743	-4.34767	-3.35467	**-0.972009**
"J0766"	{2, 2, 2, 1, 0}	{3.5, 1, *}	12 889.6	-5.5187	-4.01895	-3.67684	**-2.28044**
"J0768"	{2, 1, 2, 0, 2}	{3.5, 1, *}	12 904.6	-7.02967	-5.12815	-4.75111	**-2.99839**
"V6027"	{1, 2, 0, 1, 2}	{2, 1.5, 3.5}	12 938.3	-3.9783	-2.87661	**-2.58817**	-3.43492
"P2971"	{*1, 1, 1, 1, 2*}	{1.5, 0.5, *}	12 981.9	-4.20295	-2.59056	**-1.83225**	-3.18115
"P0888"	{*2, 1, 1, 1, 2*}	{4, 1, *}	12 983.	-5.3637	-3.49766	-2.69403	**-1.95541**
"P2974"	**{1, 1, 2, 2, 2}**	{2, 0.5, *}	12 983.8	-6.02567	-4.00816	-3.11192	**-2.51059**
"M0398"	{0, 0, 2, 2, 2}	{0.5, 1.5, *}	12 985.8	-5.47859	**-4.43498**	-5.44234	-5.40259

index	genotype	R / Y / V	Z	$\log_{10}(P_0)$	$\log_{10}(P_{Bx1})$	$\log_{10}(P_{F1})$	$\log_{10}(P_{opp})$
"V6899"	{0, 0, 0, 2, 0}	{0.5, 3, 2.5}	18 375.8	-9.30623	-6.11531	-5.31199	**-1.26318**
"V6898"	**{0, 0, 0, 1, 1}**	{0.5, 2.5, 2}	18 375.7	-9.02308	-5.62276	-4.52542	**-1.6172**
"V6310"	**{1, 0, 0, 0, 0}**	{0.5, 3, 1}	18 375.5	-8.81046	-5.21527	-4.00491	**-1.65773**
"V6886"	{0, 0, 0, 2, 0}	{0.5, 2.5, 1.5}	18 375.5	-9.30577	-6.11504	-5.31174	**-1.26318**
"V6896"	{*1, 1, 1, 2, 1*}	{2, 2, 1.5}	18 375.4	-3.74179	-1.48487	**-0.861017**	-3.42887
"S8697"	**{0, 1, 0, 1, 0}**	*	18 374.1	-8.76883	-5.36898	-4.25791	**-0.933579**
"S0611"	{*, 1, 0, 2, 0}	{1.5, 2.5, 2}	18 205.5	-3.71579	-2.85534	-2.27578	**-1.47056**
"M1369"	{0, 2, 1, 1, 1}	{0.5, 1, *}	16 797.5	-4.79714	-3.21285	**-2.8562**	-3.12103
"KTPla8"	**{0, 1, 0, 0, 0}**	*	16 770.	-7.73874	-5.17588	-4.15807	**-1.14172**
"P0057"	**{0, 0, 0, 1, 0}**	{0.5, 2.5, *}	16 679.3	-7.41817	-4.93055	-3.92265	**-0.869115**
"M0815"	{0, 1, 1, 2, 2}	{2, 1, *}	16 412.9	-4.31431	**-3.26738**	-3.5834	-4.19875
"M0818"	{0, 1, 0, 2, 1}	{0.5, 1, *}	16 407.4	-4.89528	-3.33361	-2.83463	**-2.07574**
"T4892"	**{1, 0, 0, 0, 1}**	{0.5, 2.5, 1}	16 358.9	-5.60791	-3.67708	-2.77922	**-2.40582**
"T3919"	{0, 2, 0, 2, 1}	{0.5, 1, 1.5}	16 320.5	-4.41297	-3.27854	-3.20304	**-2.74243**
"M0460"	{1, 0, 2, 2, 0}	{2, 1, *}	16 302.8	-4.07101	**-2.99048**	-3.19363	-3.99123
"Z2879"	**{1, 0, 0, *, 0}**	{1.5, 3, 1.5}	16 213.4	-4.63698	-3.06893	-2.31586	**-1.59192**
"S5374"	{*0, 1, 1, 1, 1*}	{0.5, 1, 1}	16 061.3	-4.27146	-2.88912	**-2.29752**	-2.45434

Figure C.17: (A)Improbable individuals on the yellow flank; there are 27/729=3.7% individuals with $P < 2 * 10^{-4}$ in the yellow flank (defined as <13Km). Probabilities are calculated from the allele frequencies in either flank, and the most likely assignment is shown in bold. Genotypes are roughly classified by eye (13 opposing homozygotes (red); 8 F1-like (italics); 6 foreign homozygotes (bold)). This classification approximately corresponds to the probabilities of coming from the local population, F1, or being a direct seed disperser. However, these probabilities do not show that chance of coming from the centre of the hybrid zone, which is the most likely explanation for the opposing homozygotes. Most individuals come from the immediate flank (between 2.5 and 1.5Km from the centre), but two come from further away. (B) The same for the magenta flank, where there are 17/1596=1.06% individuals with $P < 210^{-4}$ (defined as >16Km). Genotypes are roughly classified by eye (8 opposing homozygotes (red); 2 F1-like (italics); 7 foreign homozygotes (bold)). There is a cluster of yellow individuals 3.8Km out.

APPENDIX D

SUPPLEMENTARY INFORMATION FOR EFFECTS OF FINE-SCALE POPULATION STRUCTURE ON THE DISTRIBUTION OF HETEROZYGOSITY IN A LONG-TERM STUDY OF *ANTIRRHINUM MAJUS*

D.1 SNP panel

For each individual, DNA was extracted from leaf material collected from the field site, and was genotyped for the SNP panel by LGC Genomics (Middlesex, UK) using the KASP genotyping platform. Due to repeated sampling of the same individuals across years, the error rate of this method could be calculated, and was found to be low (mean error rate < 0.1% per locus).

Candidate loci were identified using a draft *A. majus* reference genome (~ 630 Mb across eight linkage groups; courtesy of Yongbiao Xue, Beijing Institute of Genomics); see ref Li et al., 2019. In this study, SNPs were chosen to have overall mean frequency between 0.1 and 0.9; 90% had frequency between 0.25 and 0.75. SNPs that showed excessive geographic differentiation were eliminated by requiring a linear regression gradient of allele frequency on east-west distance to be less than 0.09 km^{-1}; 90% of chosen SNPs had a gradient < 0.03 km^{-1}. Furthermore, we required that $F_{ST}<0.1$; F_{ST} was calculated by dividing the region into 200m squares, yielding 164 non-empty demes. Finally, the overall heterozygote deficit, F_{IS}, was required to be between -0.1 and 0.2; 90% of chosen SNPs had $-0.04 < F_{IS} < 0.1$. SNPs with heterozygote deficit $F_{IS} > 0.2$ also showed high F_{ST} and/or clinal gradient, whilst those with $F_{IS} < -0.1$ were likely due to genotyping artefacts (e.g., primers binding to more than one site in the genome). After applying these filters, 170 SNPs remained. Finally, we chose to work with the 91 SNPs that were assayed for at least 60% of the Planoles sample (i.e., at least 13,411 individuals).

D.2 Variation in inbreeding

The identity disequilibrium that we find is due partly to associations between linked SNP, and partly to associations between unlinked SNP (73% vs. 27%, respectively). Table D.2.1 shows that

correlations in h between SNP within linkage groups are consistently positive, averaging Pearson's $r = 0.01126$ - much higher than the average correlation of 0.00240 between all pairs of SNP, which are mostly unlinked. For the 155 individuals with H<0.3, the mean correlation within linkage groups, 0.0354, is much higher, reflecting the shared inheritance of large blocks of genome for close relatives. Correlations are higher between adjacent SNP, and yet higher in highly inbred individuals.

Table D.1: Correlations in heterozygosity within the 8 linkage groups (LG). The third and fourth columns give the mean correlation in H between loci within each linkage group, for all 22,353 individuals versus for the 155 individuals with H<0.3. The next two columns give the mean correlations between adjacent SNPs. The last two columns give g_2 values within each linkage group. Note that 1 of the 91 SNP was not assigned to a linkage group.

		H within LG		adjacent SNP		g_2 within LG	
LG	No. SNPs	all inds	H<0.3	all inds	H<0.3	all inds	H<0.3
1	13	0.01828	0.04423	0.08998	0.15209	0.02266	0.10093
2	15	0.02929	0.05021	0.12139	0.13723	0.03229	0.13187
3	10	0.00459	0.03298	0.00958	0.06359	0.01012	0.07800
4	12	0.00292	-0.00370	0.00294	-0.00530	0.00327	0.00512
5	12	0.02407	0.07854	0.02654	0.10365	0.02834	0.20210
6	15	0.00402	0.03128	0.00982	0.06743	0.00522	0.08369
7	6	0.00388	0.03862	-0.00159	0.03416	0.00571	0.11591
8	7	0.00304	0.01089	0.00334	0.00805	0.00297	0.06129
Mean	90	0.01126	0.03538	0.03275	0.07011	0.01382	0.09736

D.3 Effects of pollen dispersal on heterozygosity

The distribution of heterozygosity of offspring depends on distance between parents. We show this by simulating offspring, using all field-sampled individuals as mothers (Mathematica notebook in electronic supplementary material). We chose fathers close to a given distance away, by choosing 12 points evenly spaced on a circle, and taking the nearest individual to any of those points. The genotype of the offspring was determined by Mendelian inheritance based on parental genotypes. The mean heterozygosity of offspring from two parents is linearly related to their pairwise identity; thus, the increase in mean identity with distance (Fig. D.1 A, C) is a precise reflection of the decay in pairwise relatedness. The variance in heterozygosity decreases with distance, as individuals become less related (Fig. D.1 B, D). Both mean and variance of H change sharply over scales of a few metres, and are hardly affected by linkage (compare gray and black lines in Fig. D.1). The observed values from the field data (horizontal lines in Fig. D.1) are consistent with pollination from fathers ~10m away, but are of course the product of a broad distribution of distances.

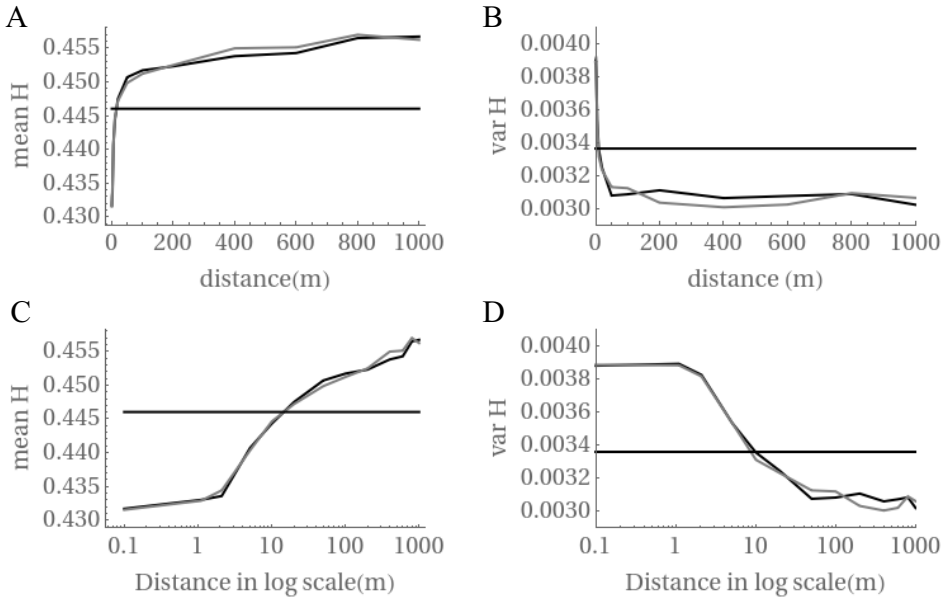


Figure D.1: Mean (A, C) and variance (B,D) of H as a function of the distance between parents. Offspring are generated with no linkage (black) or with linkage (gray); observed values from the field data are shown as a horizontal line. Plots C and D show distance in a log scale.

Table D.2: Mean and variance of multilocus heterozygosity (H) and identity disequilibrium (g_2) from field data and offspring simulated from three possible patterns of pollen dispersal (a leptokurtic kernel, a Gaussian kernel, and pollen from the nearest neighbour).

	H mean	H variance	g_2	g_2 CI
Field data	0.4460	0.0034	0.0029	0.0026 - 0.0033
Leptokurtic offspring	0.4458	0.0034	0.0020	0.0016 - 0.0024
Gaussian offspring	0.4323	0.0039	0.0053	0.0049 - 0.0057
Neighbour offspring	0.4314	0.0039	0.0056	0.0051 - 0.0060

Table D.3: Test statistic and *p*-value from *t*-test, *F*-test and Kolmogorov-Smirnov (KS) test for each pairwise comparison between heterozygosity calculated from field data and offspring simulated from leptokurtic, Gaussian, and nearest neighbour matings.

		t-test		*F* test		KS test	
Dataset	Dataset	*t*	*p*	*F*	*p*	*D*	*p*
Field	Leptokurtic	1.08	0.281	-0.55	0.579	0.015	0.015
Field	Gaussian	24.06	<0.001	-10.02	<0.001	0.094	<0.00001
Field	Neighbour	25.17	<0.001	-9.78	<0.001	0.103	<0.00001
Leptokurtic	Gaussian	23.07	<0.001	-9.80	<0.001	0.086	<0.00001
Leptokurtic	Neighbour	24.19	<0.001	-9.36	<0.001	0.090	<0.00001
Gaussian	Neighbour	1.06	0.29	0.31	0.75	0.009	0.36

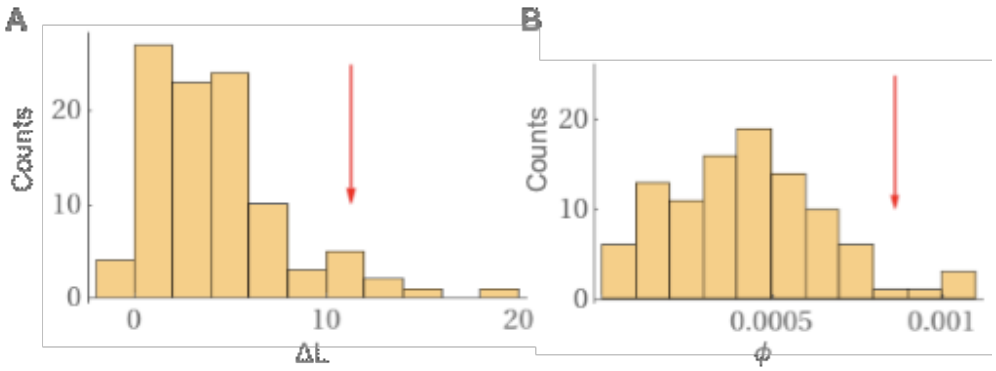


Figure D.2: The distribution of increase in the likelihood (ΔL) (A) and selfing rate, ϕ (B) between the single and mixed Gaussian distributions from 100 replicates of simulated matings from leptokurtic dispersal distribution. The red arrow points to the value observed in the field data.

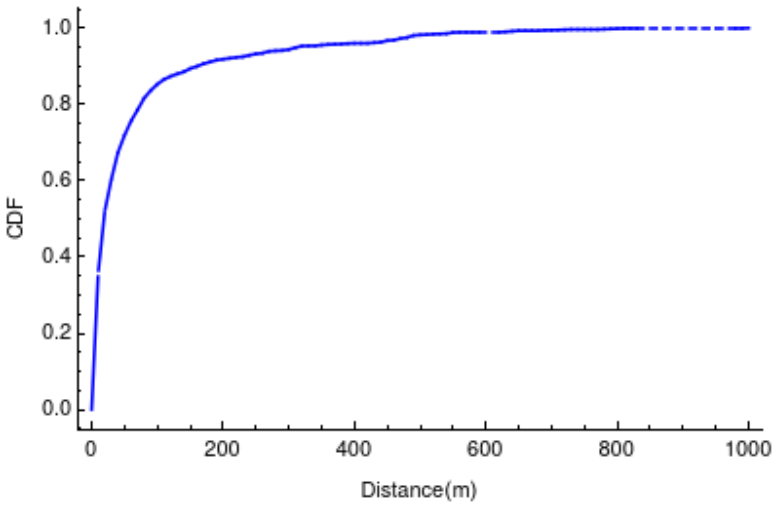


Figure D.3: CDF of the empirically measured pollen dispersal distribution. Here, 50% of matings occur within 20m and 75% of matings occur within 60m.

D.4 Heterozygosity in a simulated spatial pedigree

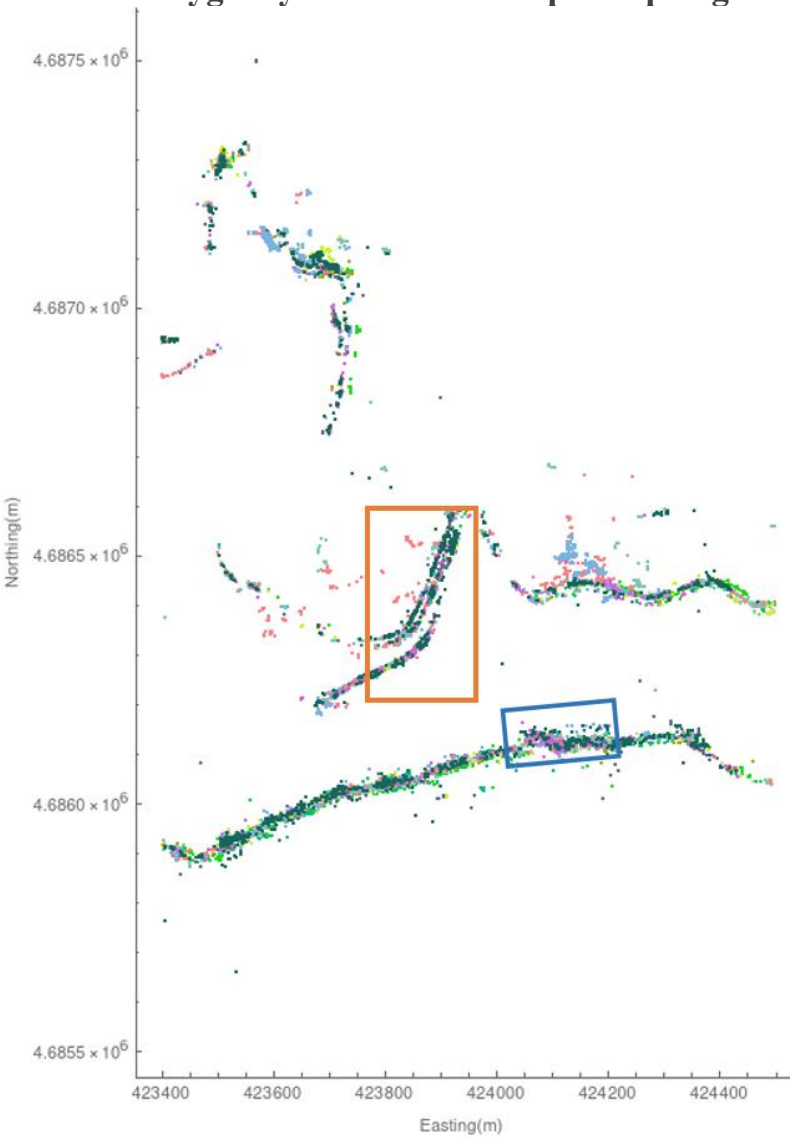
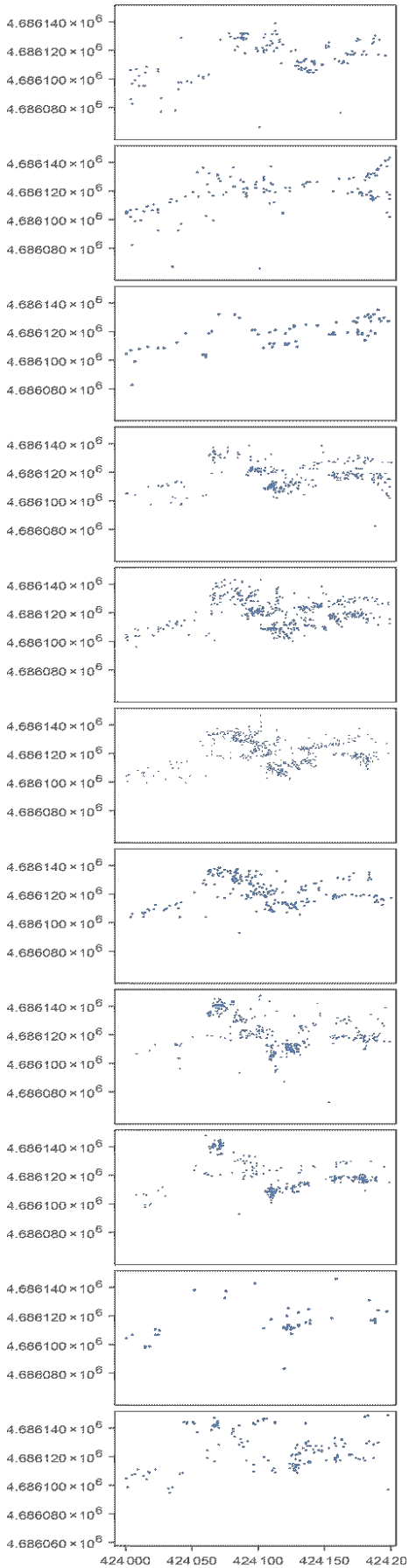


Figure D.4: Individual plant locations in the simulated region of the field site. Each colour represents a different year. See Fig. D.5 for time series of boxed areas.


4.686140 × 10^6
4.686120 × 10^6
4.686100 × 10^6
4.686080 × 10^6
4.686060 × 10^6
424 000
424 050
424 100
424 150
424 20

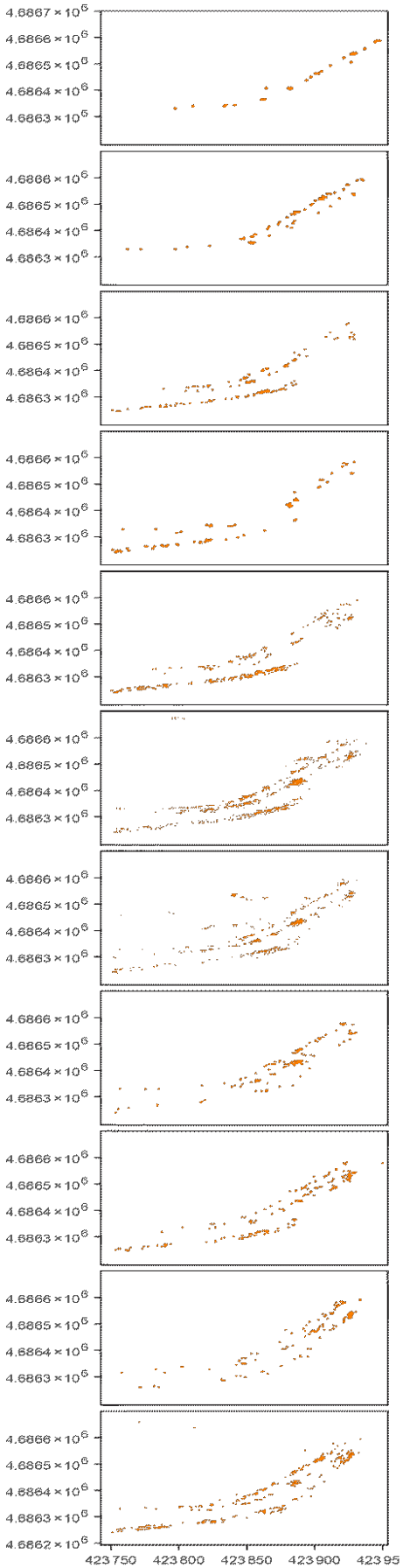

4.6867 × 10^6
4.6866 × 10^6
4.6865 × 10^6
4.6864 × 10^6
4.6863 × 10^6
4.6862 × 10^6
423 750
423 800
423 850
423 900
423 95

Figure D.5: Close-ups of sections of the lower road (left) and upper road (right) from the field data, showing changes in patchiness over time from 2009 (top) to 2019 (bottom). Sections are denoted as blue and orange in Fig. D.4.

Detailed Methods and Validations for Simulated Spatial Pedigree

Since there is no analytical result for probability of identities for a heterogeneously distributed population with leptokurtic dispersal, we validated the simulation by comparing pairwise relatedness calculated from the genotypes using 10 replicate genotypes (using the method described in 'Heterozygosity in a simulated spatial pedigree') against that directly calculated from the simulated pedigree (F). F can be considered as a N* N matrix with each element F_{ij} (corresponding to row i and column j) giving the probability of identity between individuals i and j, where N is the population size. If we start with a population of unrelated individuals, the probability of identity matrix at generation 0, F_0, would contain only 0's. The probability of identity of two distinct genes from a pair of distinct individuals i and j in generation g+1 is $F_{ij,g+1} = \sum_{k,l} M_{ik} F^*_{kl,g} M_{lj}$, where M_{xy}= 1/2 if y is a parent of x (with no selfing) and 0 otherwise, and $F^*_{kl,g} = F_{kl,g}$ if $k \neq l$ and $F^*_{kl,g} = \frac{1}{2} + \frac{1}{2} F_{kk,g}$ if $k = l$. $F^*_{kl,g}$ denotes the probability of identity of individuals k and l in the previous generation g (Charlesworth et al., 2010). F_{ij} as a function of distance was found from the genotypes and pedigree. Note that we use a smaller N=1000 for this calculation due to computational constraints of calculating F from the pedigree. We see that isolation by distance from the pedigree matches the average from 10 replicate genotypes (Fig. D.6). Furthermore, we verified our algorithm by comparing the proposed and realized seed and pollen dispersal distributions from the pedigree (Fig. D.7). Together, these two checks validate the estimation of F from the simulated pedigree, and the algorithm for choosing parents.

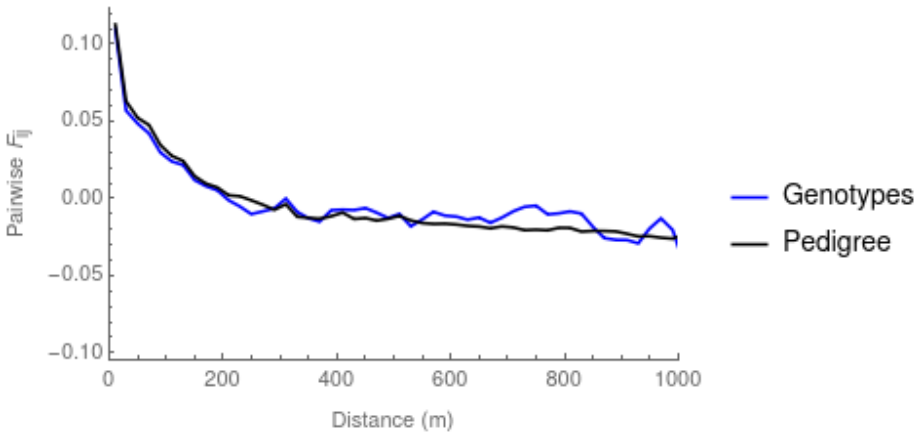


Figure D.6: Isolation by distance calculated for a simulated heterogeneous population of 1000 individuals calculated directly from the pedigree (black) and the average from 10 replicate genotypes (in blue).

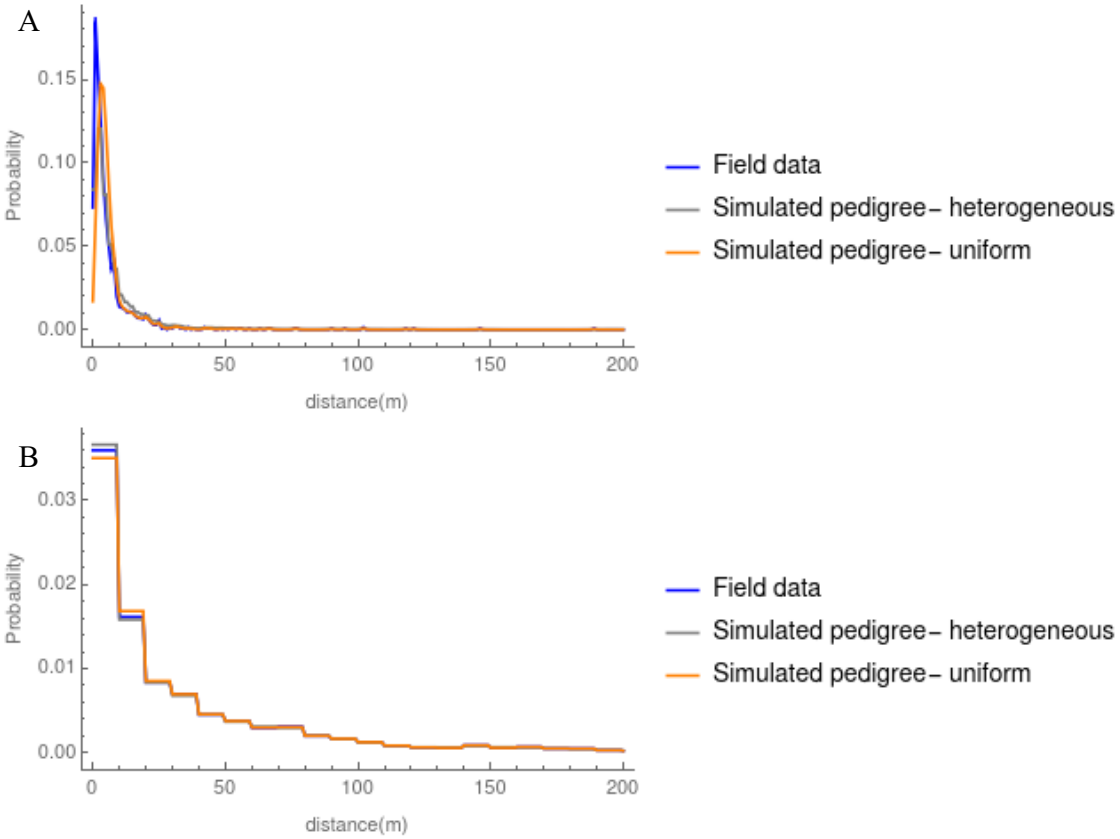


Figure D.7: Realized (gray and orange) and proposed (blue) seed (A) and pollen (B) dispersal distribution for the simulated pedigrees with heterogeneous and uniform population structure. Due to computational constraints, these are calculated from the last 300 generations for the simulated pedigree with uniform density. The pedigree with F_{ST} closest to that of the field data is shown for the heterogeneous case (also in Fig. D.8B, D.9).

Table D.4: Mean and standard deviation (SD) of the proposed and realized seed and pollen dispersal distributions for the simulated pedigrees with uniform and heterogeneous spatial structure.

	Seed dispersal		Pollen dispersal	
	Mean	SD	Mean	SD
Proposed	9.52468	38.2242	62.684	119.327
Heterogeneous pedigree 1	12.797	40.1884	62.5955	118.765
Heterogeneous pedigree 2	12.7785	40.0344	62.6493	118.834
Heterogeneous pedigree 3	12.7724	40.0098	62.5716	118.682
Heterogeneous pedigree 4	12.7863	40.0749	62.6127	118.778
Heterogeneous pedigree 5	12.7858	40.0773	62.5902	118.774
Heterogeneous- average	12.784	40.0770	62.6039	118.767
Uniform pedigree	10.4706	38.0741	63.1388	119.122

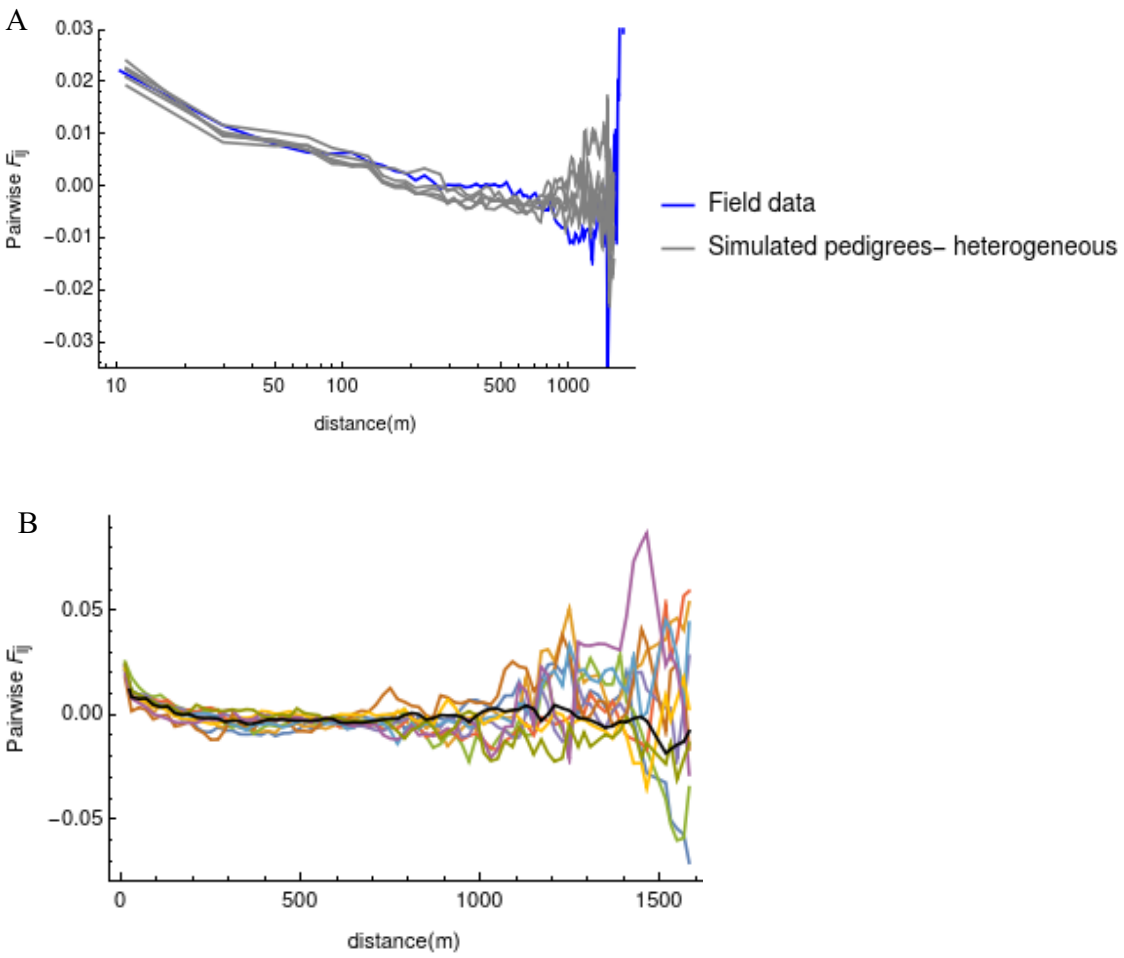


Figure D.8: (A) Isolation by distance for the field data (blue) and five simulated population pedigrees (gray) plotted on a log scale. (B) Isolation by distance from ten replicates of a single pedigree along with their average shown in black.

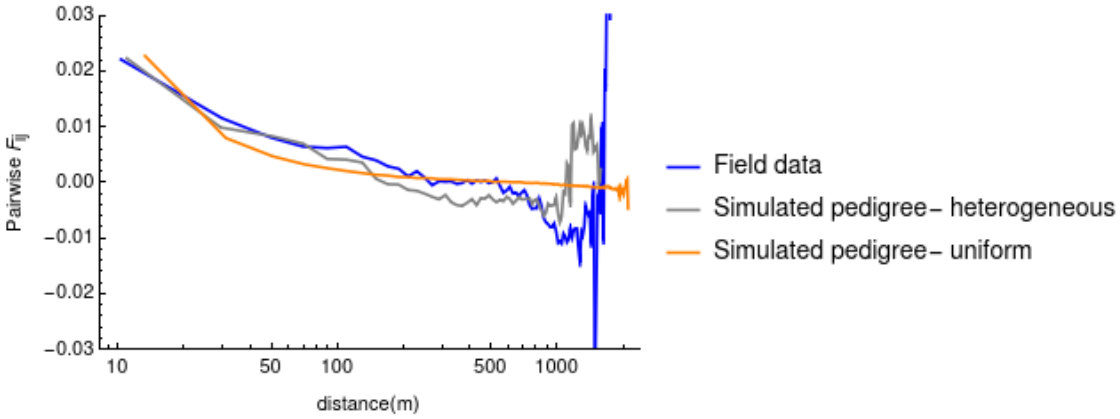


Figure D.9: Isolation by distance for the field data (blue), simulated pedigree with realistic spatial structure (gray) and uniform density (orange) plotted on a log scale.

Table D.5: F_{ST}, F_{IS} and g_2 values from the field data and from simulated (sim.) pedigrees with heterogeneous and uniform density. For the simulations, mean ± standard deviation of ten replicate sets of genotypes are shown for each pedigree, across the five pedigree means, and across all 50 replicates (ten replicates for five pedigrees).

		F_{ST}		F_{IS}		g_2	
Field data		0.022		0.0211		0.00262	
Sim. heterogeneous	Pedigree 1	0.0192	± 0.00383	0.0216	± 0.00236	0.00274	± 0.000723
	Pedigree 2	0.0226	± 0.00348	0.0254	± 0.00179	0.00258	± 0.000560
	Pedigree 3	0.0222	± 0.00221	0.0247	± 0.00149	0.00312	± 0.001120
	Pedigree 4	0.0239	± 0.00143	0.0244	± 0.00132	0.00240	± 0.000854
	Pedigree 5	0.0208	± 0.00346	0.0259	± 0.00160	0.00235	± 0.000448
	Across pedigree means	0.0217	± 0.0029	0.0244	± 0.0017	0.00264	± 0.000741
	Across all 50 replicates	0.0217	± 0.0033	0.0244	± 0.0023	0.00264	± 0.000797
Sim. uniform	Pedigree 1	0.0226	± 0.0009	0.0203	± 0.00059	0.00171	± 0.000083

References

M. Li *et al.*, "Genome structure and evolution of *Antirrhinum majus* L," *Nat. Plants*, vol. 5, no. 2, pp. 174–183, 2019, doi: 10.1038/s41477-018-0349-9.

B. Charlesworth and D. Charlesworth, *Elements of evolutionary genetics*. W. H. Freeman, 2010.