The Phylogeographic History and Contemporary Evolution of the Invasive Species *Avena barbata* Pott ex Link in California

by

Kate Crosby

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
October 2014

To my full-sibs, Jane & Erin,
and to my uncle Andrew (a.k.a. "MMQB")

Table of contents

List of Tables

List of Figures

ABSTRACT

Understanding the factors that precipitate a successful colonization of a new geographic region is a major goal in both ecology and evolution. In my doctoral thesis, I examined evolutionary patterns (Ch. 2), ecological patterns (Ch. 3), and eco-evolutionary processes (Ch. 4 & 5) associated with the invasion of a highly-selfing, annual grass, *Avena barbata* Pott ex Link into California from the Mediterranean Basin.

Because colonizing populations may experience severe genetic bottlenecks, multiple introductions of different genetic variants may be an important source of variation facilitating adaptation to the novel habitat. In the second chapter of my thesis, I show that there have been at least three introductions of *A. barbata* into California, and that substantial spatial mixing has occurred between haplotypes within populations in California. There was also evidence of population structure due to a north-south cline.

Although niche overlap between the endemic and novel habitat may facilitate biological invasions, niche shifts may also occur during invasion. In the third chapter, a species distribution model (SDM) was employed to characterize and compare the niches in the Iberian Peninsula (home range of *A. barbata*) to niches in California (invaded range). Niche separation was observed between the two geographic regions, suggesting that *A. barbata* occupies a different environment in California and may be evolving.

Traits with a known genetic basis may prove useful in tracking contemporary evolutionary change and provide insights on adaptive versus neutral processes. In the fourth chapter, I compared the contemporary frequency of two binary, heritable characters: lemma color and leaf sheath pubescence to that reported in past studies from the 1970s. Due to natural selection, I found that light lemma color and leaf sheath pubescence had increased in frequency.

Although intuition suggests that large plants should be at a selective advantage within populations, recent discussion in the literature suggests the opposite. The idea that small plants with lower total fecundity may be selectively favoured contradicts basic Fisherian principles. In the fifth chapter, using a simulation and field data from *A. barbata* recombinant inbred lines (RILs), I show that individuals with large body size are always favored.

List of Abbreviations and Symbols Used

AET	Actual Evapotranspiration
AFLP	Amplified Fragment Length Polymorphism
bp	Basepair
β	Directional selection gradient
CIM	Composite Interval Mapping
cpDNA	Chloroplast DNA
GBS	Genotype-by-sequence
GBIF	Global Biodiversity Information Facility
GRF	Gaussian Random Field
LG	Linkage Group
LOD	Log$_{10}$ of odds of linkage
LSC	Large single copy region
MIM	Multiple Interval Mapping
PET	Potential Evapotranspiration
pg	Picogram
QTL	Quantitative Trait Loci
RILs	Recombinant Inbred Lines
SDM	Species distribution model
SNP	Single nucleotide polymorphism
SSCP	Single strand conformational polymorphism

Acknowledgments:

First, and foremost, I would like to thank my advisor, Dr. Robert G. Latta (aka "Dr. Bob"). Bob has taught me how to be a conscientious scientist, has been a source of support, and has been one of the most interesting characters that I have ever had the pleasure of knowing. I will forever be grateful to him for instilling in me the importance of: "Go[ing] back to first principles." This statement will ring in my ears forevermore; hopefully, I will listen. Bob was great about allowing me to explore many topics, and encouraged me to participate in collaborative projects within the department. I will never forget the time we spent driving through California collecting wild oats. Our visit to Del Loma in the pouring rain ("not xeric!"), Calistoga Dump Road and Lick observatory "fence-jumping", and the situational irony of a lumberjack restaurant in Redding, whose lunchtime soundtrack included "Smalltown Boy" by the Bronski Beat.

My advisory committee, consisting of Drs. Jeremy T. Lundholm and Mark O. Johnston, has been "absolutely fabulous", and they were also very patient with me. They openly invited me into their home labs, provided advice when needed, and taught me how to ask more directed and interesting questions. Thank you also to Dr. Steve Franks who very kindly agreed to be the external examiner for this thesis.

I am especially grateful to Carman Mills and his assistants Mackenzie Bartlett (RGL lab alumnus) and Greg Britten for help in the greenhouse. Ian Paterson of the Marine Gene Probe Lab was also very helpful addressing our lab's needs in the molecular sense. Paul Kron and Dr. Brian Husband at the University of Guelph for assisting us with flow cytometry. Dr. Harold Bockelman at the USDA's National Small Grains Collection and Dr. Dallas Kessler at Plant Genome Resources Canada (PGRC) were instrumental in supplying international oat accessions. I am also grateful of the NSERC Discovery Grant awarded to Bob for fieldwork and lab work. Additionally, I am thankful for financial aid received from the Lett Fund, which allowed me to continue my studies for an additional year.

Thank you to the staff in the Biology Department office – in particular Julie Walker and Carolyn Young who were always helpful in dealing with administrative issues surrounding exams, room bookings, and pay schedules.

Being the only graduate student in a lab was difficult at times, and I could not have persisted without strong support from my sister lab (MOJ lab) and her members: Dr. Chris Kozela, Hannah Zatzman, Zoë Migicovsky. But, especially thanks to Dr. Maggie Bartkowska, who continues remotely to encourage, question, and support me from the far lands of Toronto, and Mike McElroy (and Tammy Streeter) for providing support in the form of day-to-day coffee/tea, advice, and a roof over my head.

Past and present departmental colleagues were especially helpful in getting me to think critically, write papers, challenge my comfort zone, and combat the loneliness of the "gurgle/drilling/dripping room". Thank you, Drs. David Roy Smith, Derek Tittensor, Denis Roy, Camilo Mora, and Alex Smith. I am also grateful for the superb banjo/guitar playing, bluenose class sailing, and green algal maestro in the form of Dr. R.W. "Bob" Lee. He provided much needed wisdom and practical common-sense perspective, which is often lost in the minutiae of day-to-day life in academia.

I am also incredibly appreciative of the support I received from (in no particular order): Michelle Lloyd, Dr. Njal Rollinson, Jackie Porter, Marina Milligan, Kristi

O'Brien, Jon Lapalme, Dr. Steph Boudreau, Dr. Friso Palstra, Dr. Bob Farmer, Sean Anderson, Adrian Doolittle, Greg Britten, Laurie Baker, Dr. Dan Noble, Dr. Paul Debes, Dr. Kyle Gardner, Taylor Stokes, Dr. David Mensink, Tomomi Suwa, Dr. Hal Whitehead, Dr. Hugh Broders, Marge Sullivan, Adrienne Malloy, and Pat Sullivan.

My immediate family – my parents (Don & Eileen), my two sisters Jane and Erin, and their partners Jody Rushton and George Solomon – were always there to push me on, provide wooded Cape Breton interludes, Bras D'or seascape "vistas", good food, jokes, and many other things. Special thanks to my nephew Sam, for generally knowing when it was time to "go play outside." Thank you!

Finally, I would be amiss if I did not thank my partner-in-crime, Myriam Lacharité, for all her years of encouragement and support.

Chapter 1 Introduction

1.1 Introduction

Studying biological invasions provides unique insight into processes that contribute to contemporary evolution in the wild. Identifying the factors that underlie successful colonization and spread is challenging because it requires understanding simultaneously the ecological and genetic factors that facilitate or inhibit adaptation and/or population persistence. First, investigators must grapple with whether the newly invaded range is also a novel environment. A novel environment may exert different selective pressures (both abiotic and biotic) than those in the native environment, which may result in a niche shift for the invasive species (Broennimann *et al.* 2007). Second, if species are invading a new geographic range and/or environment, they may be constrained by the amount of adaptive genetic variation at their disposal due to genetic bottlenecks founding population size and genetic bottlenecks (Dlugosch & Parker 2008; Prentis *et al.* 2008).

Recombination between multiple genotypes of the same species (Chun *et al.* 2009) or hybridization with native species (Daehler & Strong 1997) may allow invaders to generate novel genetic combinations that may combat potential founder effects and may ultimately contribute to invasiveness (Ellstrand & Schierenbeck 2000). Adaptive evolutionary change is likely to occur in a novel geographic range, if that novel range is also a novel environment.

A species colonizing a new geographic range may be more successful at persisting in the invaded range if the invaded environment overlaps substantially with the native environment. Similar environments between ranges indicate niche overlap or niche conservatism (Guisan *et al.* 2014). For invasive plants, niche conservatism (at least with respect to climate variables) is regularly observed (Peterson 2011; Petitpierre *et al.* 2012) and may be a one of the reasons for the apparent success of certain invasive species. Similarity in ecological niches between native and introduced ranges may also reduce the necessity of generating novel genetic variants that would facilitate and speed adaptation. Speculatively, niche overlap may be of greater importance than adaptive genetic variation in determining the success of invasive self-fertilizing species, whose effective recombination rate is much reduced relative to outcrossing species (Charlesworth & Wright 2001).

1.2 Historical background of the study system

Californian populations of *Avena barbata* Pott ex Link are a storied system of molecular ecology having first been brought to attention by the working group of R.W. Allard at UC Davis in the 1960s-1980s. *Avena barbata* populations in California were among the first wild organisms to have their genetic variation screened with "state of the art" genetic markers at the time - allozymes. Large geographic surveys of the state using five allozyme loci revealed the occurrence of largely two monomorphic genotypes of *A. barbata*, with few recombinants detected. Further, these two genotypes were termed 'mesic' and 'xeric' - 'ecotypes' because of their purported association with moist and dry environments at both macro- (Clegg & Allard 1972) and micro- (Hamrick & Allard 1972) geographic scales. Under the assumption that the populations were at migration-selection equilibrium, the system was regarded as an example of local adaptation and divergent selection on co-adapted genotypes throughout the 1960s-1980s (Allard et al. 1972, Clegg & Allard).

The suspected origin and colonization history of the ecotypes to California is sketchy, at best. It is thought that the oats were introduced during the Spanish Missionary period in the late 1700s and early 1800s (Minnich 2008), with possibly a second introduction having expanded in the 1890s (Blumler 2000). Intense surveys of individuals in the Mediterranean Basin (namely from Spanish populations in the Iberian Peninsula) revealed that the two ecotypes were each composed of a reduced subset of alleles than those found in Spain (Garcia et al. 1989; Pérez de la Vega et al. 1993), and Clegg & Allard (1972) speculate that the ecotypes in California are the result of selection against several introductions from the 'rich flora' of the Mediterranean Basin.

The original conclusion of local adaptation from the 1970s may have been premature, and some experiments from Allard's group suggest the distribution of genotypes may have been indicative only of population structure (Hutchinson 1982). An association of genotypes with particular environments may simply indicate population structure due to neutral evolutionary processes (not necessarily mutually exclusive ones either), i.e. different timing of multiple introductions, asymmetric migration, inbreeding and genetic drift. At least one past study in the 1970s indicates that mesic genotypes had higher fitness (although not significantly different) than the xeric genotypes in a common garden experiment (Hamrick & Allard 1975). Another attempt at colonization and a common garden experiment attempted to clarify the role of selection of the two genotypes to their respective environments, but showed little evidence of adaptive forces shaping the distribution of the monomorphic genotypes (Jain & Rai 1980).

More recent experimental work using reciprocal transplant common gardens of the original allozyme genotypes indicates that the original conclusion of local adaptation was erroneous (Latta 2009), and that the mesic genotype was consistently and significantly more fit than the xeric across years and environments (Latta 2009). Thus, the results from this experiment predict that the mesic genotype ought to be favoured by natural selection in the field (Latta 2009).

Conveniently, mesic and xeric genotypes are largely associated with two morphological true-breeding heritable characters: leaf sheath pubescence/glabrousness and light or dark lemma color (Miller 1977). The mesic genotype is associated with pubescence and light lemma; the xeric genotype is associated with glabrous leaf sheath and dark lemma. The amount of genotypic frequency data on these morphological markers collected statewide at over 100 populations in the form of dissertations in the 1970s makes this system a rare opportunity to study adaptive evolutionary change in the wild, with a clear prediction from a multi-year common garden reciprocal transplant experiment (Latta 2009).

1.3 Thesis objectives

The main goals of my thesis have been to isolate the number of introductions of the Californian invasive, annual, tetraploid, grass *Avena barbata*, characterize the climatic environmental differences between the native and home range, and evaluate whether there has been evolutionary change in the two heritable characters (mentioned above) since the 1970s.

Specifically, in my thesis I sought to: ascertain the number of introductions to California and genomic/ploidy changes that may have occurred since arriving in that range (Chapter 2), evaluate the degree of environmental/niche overlap vs. separation between the native and introduced range (Chapter 3), assess whether there has been contemporary evolution and recombination in two heritable traits in California over the past 40 years (Chapter 4), and assess selection for genotypes of large size in the field (Chapter 5).

Chapter 2: Evolving California genotypes of Avena barbata are derived from multiple introductions

Genetic bottlenecks occurring following colonization or at the edges of a range expansion may inhibit the persistence of a species in new habitat (Colautti *et al.* 2010). Assessing the number of different genetic variants in a new geographic range is a crucial first step to evaluating the amount of genetic variation available for evolutionary change. To this end, I quantified the number of maternal introductions of *A. barbata* to California using cpDNA, and phylogenetically placed these introductions with context within the larger Old World *Avena* species complex. Ploidy and genome size changes were investigated using flow cytometry. I also evaluated clinal population structure of cpDNA Californian haplotypes. T.O. Stokes (an undergraduate student in our lab) performed single strand conformational polymorphism (SSCP) analysis and R.G. Latta assisted in the manuscript preparation. This work has been accepted for publication in PeerJ (forthcoming).

Chapter 3: Gaussian Random Fields: Using a novel SDM technique that incorporates uncertainty to describe a niche shift from Iberia to California

My third chapter attempted to ascertain whether there was evidence to support the hypothesis that California (as a newly colonized geographic range) is also a novel niche, and whether evolution in this new niche has occurred (Chapter 4 –see below). Most studies attempting to isolate niche shifts have tended to focus on climatic variables (because these data are readily and publicly available), and many climate variables are convergent across continents, e.g. Mediterranean climate types (Ackerly 2009). Niche conservatism may be observed due the current limitations with respect to occurrence data and statistical techniques (Peterson 2011). Niche conservatism vs. niche shifts in *A. barbata* were evaluated in Iberia and California using data from GBIF, plus a number of other databases, and from our own geographic re-survey in 2010. Niche shifts were evaluated using a new technique for species distribution model (SDM) – Gaussian random fields (GRF) (Golding 2013) that incorporates uncertainty and sequentially orders importance of environmental variables.

Chapter 4: Contemporary evolution over 40 generations in an invasive annual grass

It is rare to have access to historical data with information on genotype frequency across a large geographic range. In my fourth chapter, I coupled data collected from geographic surveys of *A. barbata* in the 1970s to a survey conducted by our lab in 2010 to evaluate how the frequency of two heritable characters have changed over the past 40 years across over 100 populations throughout California. I focused specifically on the change in two heritable morphological characters, light lemma color and leaf sheath pubescence, which were predicted to be increasing in frequency from past observations made from a previous common garden experiment (Latta 2009). The underlying model of inheritance of these characters was investigated using a genetic map constructed from combined marker data of genotype-by-sequence (GBS) and amplified fragment length polymorphic (AFLP) markers. These markers were used to ascertain QTL in previously developed recombinant inbred lines (RILs) of *A. barbata*. I also assessed whether occasional recombination and drift as opposed to selection might be responsible for contemporary evolutionary change in California. Latta and Gardner (unpub) recorded data for one of the characters – leaf sheath pubescence from previous experiments. AFLP marker data is from (Gardner & Latta 2006), and GBS markers were assayed by Latta (unpub). The goal of this chapter was to assess whether adaptive evolutionary change had occurred in California.

Chapter 5: A refutation of the reproductive economy hypothesis.

It has been recently suggested in the literature that more offspring in total come from parents of small size as opposed to large size (Aarssen 2007; Neytcheva & Aarssen 2008; Chambers & Aarssen 2008). However, this reasoning is counter-intuitive because of strong positive correlations between fitness and large body size in plants. Using data from recombinant inbred lines (RILs) from a common garden field experiment conducted by R.G. Latta, and simulations I wrote and carried out R, I found that large plants consistently contribute more offspring per capita and are thus favoured by selection. This study was published in *Evolutionary Ecology* (Crosby & Latta 2013).

1.3 References

Aarssen L (2007) Death without sex—the "problem of the small" and selection for reproductive economy in flowering plants. *Evolutionary Ecology*, **22**, 279–298.

Ackerly DD (2009) Evolution, origin and age of lineages in the Californian and Mediterranean floras. *Journal of Biogeography*, **36**, 1221–1233.

Allard RW, Babbel GR, Clegg MT, Kahler AL (1972) Evidence for coadaptation in Avena barbata. *Proceedings of the National Academy of Sciences of the United States of America*, **69**, 3043–3048.

Broennimann O, Treier UA, Müller-Schärer H *et al.* (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10**, 701–709.

Chambers J, Aarssen L (2008) Offspring for the next generation: most are produced by small plants within herbaceous populations. *Evolutionary Ecology*, **23**, 737–751.

Charlesworth D, Wright SI (2001) Breeding systems and genome evolution. *Current Opinion in Genetics & Development*, **11**, 685–690.

Chun YJ, Nason JD, Moloney KA (2009) Comparison of quantitative and molecular genetic variation of native vs. invasive populations of purple loosestrife (Lythrum salicariaL., Lythraceae). *Molecular Ecology*, **18**, 3020–3035.

Clegg M, Allard R (1972) Patterns of genetic differentiation in the slender wild oat species Avena barbata. *Proceedings of the National Academy of Sciences*, **69**, 1820–1824.

Colautti RI, Eckert CG, Barrett SCH (2010) Evolutionary constraints on adaptive evolution during range expansion in an invasive plant. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1799–1806.

Crosby K, Latta RG (2013) A test of the reproductive economy hypothesis in plants: more offspring per capita come from large (not small) parents in Avena barbata. *Evolutionary Ecology*, **27**, 193–203.

Daehler C, Strong D (1997) Hybridization between introduced smooth cordgrass (Spartina alterniflora; Poaceae) and native California cordgrass (S. foliosa) in San Francisco Bay, California, USA. *American Journal of Botany*, **84**, 607–611.

Dlugosch K, Parker M (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology*, **17**, 431–449.

Ellstrand NC, Schierenbeck K (2000) Hybridization as a stimulus for the evolution of invasiveness in plants? *Proceedings of the National Academy of Sciences*, **97**, 7043–7050.

Gardner KM, Latta RG (2006) Identifying loci under selection across contrasting environments in Avena barbata using quantitative trait locus mapping. *Molecular Ecology*, **15**, 1321–1333.

Golding N (2013) Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications. 1–247. University of Oxford, UK. Doctoral Dissertation.

Guisan A, Petitpierre B, Broennimann O, Daehler C, Kueffer C (2014) Unifying niche shift studies: insights from biological invasions. *Trends in Ecology & Evolution*, **29**, 260–269.

Hamrick JL, Allard RW (1972) Microgeographical Variation in Allozyme Frequencies in Avena barbata. *Proceedings of the National Academy of Sciences of the United States of America*, **69**, 2100–2104.

Hamrick JL, Allard RW (1975) Correlations between quantitative characters and enzyme genotypes in Avena barbata. *Evolution*, **29**, 438–442.

Jain SK, Rai KN (1980) Population biology of Avena. VIII. Colonization experiment as a test of the role of natural selection in population divergence. *American Journal of Botany*, **67**, 1342–1346.

Latta RG (2009) Testing for local adaptation in Avena barbata: a classic example of ecotypic divergence. *Molecular Ecology*, **18**, 3781–3791.

Neytcheva M, Aarssen L (2008) More plant biomass results in more offspring production in annuals, or does it? *Oikos*, **117,** 1298-1307.

Peterson AT (2011) Ecological niche conservatism: a time-structured review of evidence. *Journal of Biogeography*, **38**, 817–827.

Petitpierre B, Kueffer C, Broennimann O *et al.* (2012) Climatic niche shifts are rare among terrestrial plant invaders. *Science*, **335**, 1344–1348.

Prentis PJ, Wilson JRU, Dormontt EE, Richardson DM, Lowe AJ (2008) Adaptive evolution in invasive species. *Trends in Plant Science*, **13**, 288–294.

Chapter 2 Evolving California genotypes of *Avena barbata* are derived from multiple introductions but still maintain substantial population structure

2.1 Abstract

Multiple introductions are thought to enhance the chance of successful colonization, in part because recombination may generate adaptive variation to a new environment. *Avena barbata* (slender wild oat) is a successful colonist in California, past noted for striking genetic divergence into two multilocus genotypes, but is still undergoing adaptive change. I sought to understand whether multiple introductions might be contributing to this change. I used cpDNA phylogeography of *A. barbata* within its home range and in its invaded range in California to determine the minimum number of separate introductions, and the spatial distribution of these introduced lineages. Our lab collected from sites throughout the state of California, where it is an invasive species. Accessions from a representative portion of *A. barbata*'s full native range were obtained from germplasm repositories. I sequenced seven intergenic chloroplast DNA loci for both *A. barbata* individuals in California (novel geographic range) and its ancestral range. 204 individuals were assayed for chloroplast haplotype within California using single strand conformational polymorphism SSCPs. Genome size was determined by flow cytometry. Californian accessions are tetraploid as expected, but their genome sizes were smaller than the Old World accessions. There were three haplotypes present in California that were identical to haplotypes in the native range. Within California, the presence of multiple haplotypes at a site was observed primarily in Northern and Central populations. Between populations there was still substantial structure with $F_{ST} \sim 0.33$, due to a shallow latitudinal cline caused by a preponderance of xeric haplotypes in Southern California. There was a minimum of three seed introductions to California. Recombination is thus likely to occur, and contribute to adaptation in new range in this highly-selfing, invader.

2.2 Introduction

Introduced and invasive species are likely to have to adapt to novel conditions (Dlugosch & Parker 2008; Prentis *et al.* 2008), and the first requirement of adaptation is access to a pool of genetic variation. The pool of adaptive genetic variation can be increased through introductions from multiple sources. Typically, adaptive genetic variation is much reduced in the new range compared to the home range, as many invasive species reproduce primarily via selfing or are facultatively asexual during colonization of the new range (Baker 1955; 1967; Price & Jain 1981; Barrett & Colautti 2008). During invasions, the main advantage of selfing is that individuals need not depend on pollinators or other conspecifics to pollinate with, for persistence or spread. Indeed, in selfing or clonal species, one or a few introductions could be enough to allow for successful establishment and persistence - even on a global scale (Le Roux *et al.* 2007). By contrast, outcrossing species generate new multilocus genotypes with each round of outcrossing, provided that enough individuals of different genotypes have been introduced that mates are available and to avoid inbreeding depression.

Self-fertilizing species typically suffer less from the effects of inbreeding depression than outcrossers (Husband & Schemske 1996), but they may still require adaptive genetic variation in order to respond to novel environmental conditions. It has been suggested that, multiple introductions of different genetic variants allow selfing species the opportunity for occasional outcrossing and recombination in new environments (Ellstrand & Schierenbeck 2000; Schierenbeck & Ellstrand 2009). Assessing the minimum number of introductions in a new environment is thus critical to evaluating how important the amount of genetic variation may be for self-fertilizing species in the new range, which will ultimately determine how much effective recombination is possible in the new range.

Californian populations of the highly-selfing (Marshall & Allard 1970), autotetraploid (Hutchinson et al. 1983), invasive annual grass, *Avena barbata* Pott ex Link are thought to have been introduced from the Iberian Peninsula roughly two centuries ago during Spanish colonization (Jain & Marshall 1967; Garcia *et al.* 1989). *Avena barbata* became widely known for a number of pioneering observations in the 1970s (Clegg & Allard 1972; Allard *et al.* 1972; Hamrick & Holden 1979) which found that there were predominantly two multilocus genotypes of *A. barbata* in California. Each genotype was monomorphic for a set of five allozyme loci, and few recombinants were found. One genotype was found in moist environments, while the other occurred in more arid environments, a pattern repeated at both large (Clegg & Allard 1972; Allard *et al.*

1972) and small geographic scale (Hamrick & Holden 1979), leading to the interpretation that these represented locally adapted ‘mesic’ and ‘xeric’ ecotypes. The allozyme combinations characteristic of the original Californian genotypes were not found in native Iberian populations (Garcia et al. 1989). This suggests that recombination among separate genotypes was part of the evolutionary history of *A. barbata* in California.

Past experimental work in which a novel environment was imposed in the greenhouse demonstrated that recombination between the two genotypes produced a few hybrid recombinants that were more fit than the parents (Johansen-Morris & Latta 2006; 2008). Additionally, a previous four-year field common garden, reciprocal transplant experiment of parental genotypes to wet and dry environments found that the mesic genotype and a few recombinants are consistently more fit than the xeric (Latta 2009). Thus, the mesic could be displacing the xeric genotype, or a recombinant could be displacing both as the prominent genotype throughout California. Collectively, this evidence does not support the idea that the two genotypes are locally adapted to moist and arid environments as Allard’s work suggested (though I retain the names ‘mesic’ and ‘xeric’ here), but rather that adaptive change is still occurring in California. This continuing evolution could be due to recombination between several different genetic variants as demonstrated by Johansen-Morris & Latta (2006; 2008) and/or the superior fitness of one lineage (likely the mesic) (Latta 2009).

Botanical records from the past cannot pinpoint the exact number of introductions, and there has been only conjecture that there may have been a second introduction of the mesic genotype (Blumler 2000; Minnich 2008) or several introductions to California (Clegg & Allard 1972). Thus, the first step in assessing the role of intermixing and recombination is to determine the minimum number of introductions to California and their spatial distribution. I constructed a cpDNA phylogeny of accessions from across *A. barbata*’s native range as well as its invaded range in California to determine the minimum number of introductions to California. In most plants, organelle genomes of both the chloroplast and mitochondrion are maternally inherited (Corriveau & Coleman 1988; Zhang *et al.* 2003), and provide a useful trace of seed introduction during colonization. I hypothesize that the mesic and xeric genotypes are the result of at least two separate seed (maternal) introductions to California from the native range. I sought to characterize the geographic distribution and the degree of spatial overlap of these introductions within California. As *A. barbata* is highly selfing, pollen movement is limited and recombination is restricted until a new genotype arrives in the population via seed (maternal) gene flow. Limited seed migration would tend to restrict levels of spatial intermixing, and I expected that for a selfing species most genetic diversity would occur amongst, rather than within populations (Hamrick & Godt 1996). A wider panel of individuals across California was assayed; these were sampled from both within and amongst populations to assess the validity of this expectation. Future studies using nuclear loci will examine the degree of recombination among genotypes in California. In order to rule out novel polyploid formation in an invader (often a result of hybridization (Novak *et al.* 1991; Daehler & Strong 1997)), I estimated the ploidy of a subset of accessions from both ranges using flow cytometry. As a cpDNA phylogeny of most of the species in the genus *Avena* is available (based on the trnL-F/trnF-R intergenic region - (Peng et al. 2010), I placed a root to our own cpDNA tree with a direct comparison of chloroplast DNA sequence data and ploidy estimates.

2.3 Methods

2.3.1 Source of material

Old World accessions of *A. barbata* were donated by Agriculture Canada's Plant Genome Resources of Canada (PGRC) germplasm station in Saskatoon, Saskatchewan, Canada, and the USDA National Small Grains Collection (NSGC) in Aberdeen, Idaho, USA. I attempted to obtain accessions from across *A. barbata's* range in the Old World, but with particular emphasis on the Iberian Peninsula (Spain and Portugal) as this is thought to be the origin of Californian *A. barbata* (Jain & Marshall 1967; Garcia *et al.* 1989; Minnich 2008). Official repository accession names, our own sample code names, GenBank accession numbers along with sampling locations and other information are given in supplementary material of the forthcoming publication.

In May 2010, our lab carried out an extensive geographic survey of *A. barbata* sites in California, USA. Our lab collected seeds from 95 sites matching locations described in three dissertations (Clegg 1972; Miller 1977; Hutchinson 1982) that studied allozyme variation in the 1970's. These sites represent the full range of growing conditions for *A. barbata*, and will be used for future studies comparing present day to past genotypic composition. I walked ~3-5 metres between each individual at a collection site to avoid sampling close relatives. In addition, I included in the analysis the mesic and xeric accessions used by Gardner and Latta (2006) to create a genetic mapping population of *A. barbata* – these seeds were kindly provided by P. Garcia from collections made in California during the 1980's, and had previously been genotyped using the original set of five allozyme loci (Latta *et al.* 2004). Seeds were germinated following our lab's standard protocol (Latta *et al.* 2004), and then planted in the greenhouse at Dalhousie University. Californian populations were grown in June 2010 and June 2011, while seeds from germplasm repositories were grown in June 2011 and June 2012.

2.3.2 Ploidy assessment

Forty-eight accessions from the Old World, and 24 Californian individuals were assayed for ploidy via flow cytometry. Flow cytometry was performed with a BD FACSCalibur flow cytometer at the University of Guelph, Ontario, Canada with CellQuest Pro software (BD Biosciences, San José, USA). All assays came from fresh leaf tissue harvested from young plants (10-20 cms in height) from the greenhouse at Dalhousie University and were shipped to Guelph in moist paper towels. Sample preparation was modified slightly from a previous protocol (Doležel et al. 2007). The DNA content standard used was *Vicia faba* (26.90 pg/2C) (Doležel et al. 1992). Approximately 0.5 cm^2 of *V. faba* and 1.2 cm^2 of *A. barbata* were chopped with a razor blade, and sat in cold extraction buffer for staining (100 μg/ml propidium iodide and 50 μg/ml RNAse – in this study LB01 buffer was used (Doležel et al. 2007)). The FL-2 peak analysis program was used to infer ploidy from 2C DNA content measurements of each individual against the DNA content standard. To examine whether there was any difference in DNA content (pg/2C) between California and the Old World, I performed a one-way two-sample randomization test based on 1000 Monte-Carlo re-samplings of the approximate distribution using the R-package 'coin' (Hothorn et al. 2006)

2.3.3 Chloroplast DNA variation

To create the phylogeny, I obtained cpDNA sequences from 49 Old world and 32 Californian accessions, which included two Mesic and two Xeric standards (the parents of the mapping population described in Latta et al. (2004)). I chose California sites that formed North-South and East-West transects by choosing at least one accession from sites previously sampled by Clegg (1972) and Hutchinson (1982). DNA extraction was carried out from leaf tissue following a slightly modified protocol of plant DNA extraction (Dellaporta et al. 1983); I also employed an expedient protocol optimized for seeds (Ivanova et al. 2008) for 10 samples.

I used seven previously described primer pairs (Ebert & Peakall 2009), Taberlet et al. (1991) for the large single copy (LSC) region of the chloroplast (Table 2.1). All PCR products were visualized on a 1.5% agarose TAE gel run at 60mA, 100V for approximately 1.5 hours. Sanger cycle-sequencing reactions for PCR products with single, clear bands were carried out by MacrogenUSA Inc. For each primer pair, I used only the forward primer for sequencing, with the exception of the trnT-F/trnL-R fragment, which was bi-directionally sequenced. To infer a root for our final phylogeny I screened one individual from each cpDNA haplotype using trnL-F/trnF-R (Taberlet et al. 1991), and compared this to Peng et al.'s (2010) phylogeny of the genus *Avena.*

Table 2.1 Intergenic locus region target, the primer pair (as given in reference), the number of SNPs for each locus, and the length of the fragment uploaded to GenBank. Sequencing primers are underlined. The eighth and last intergenic locus (trnL (UAA) 3'exon) is italicized because I only screened unique haplotypes among accessions from our dataset to ascertain the root of our tree see supplementary material (forthcoming publication).

Intergenic regions targeted	Primer Pair	Number of variable characters	Length of fragment in *A. barbata*(bp)	Reference
trnQ (UUG) – psbK	ANU11-L/ANU-12R	5	408	(Ebert & Peakall 2009)
atpB – rbcL*	ANU67-L/ANU68-R,	7	901	(Ebert & Peakall 2009)
psaI - ycf4*	ANU73-L/ANU74-R	3	356	(Ebert & Peakall 2009)
psaJ – rpl33	ANU83-L/ANU84-R	1	221	(Ebert & Peakall 2009)
rpl33 – rps18	ANU85-L/ANU86-R	4	595	(Ebert & Peakall 2009)
trnT (UGU)	trnT-F/trnL-R	4	510	(Taberlet et al. 1991)
trnL (UAA) 5'exon	trnT-F/trnL-R	4	259	(Taberlet et al. 1991)
trnL (UAA) 3'exon	*trnL-F/trnF-R*	*3*	*730*	(Taberlet et al. 1991)

*These loci were used in SSCP analysis

To assess the distribution of haplotypes in the introduced range, screening was expanded to more accessions within California, using single strand conformational polymorphism (SSCP). SSCPs are an efficient, and cost-effective method for screening many samples that isolate single nucleotide polymorphisms (SNPs) (Gasser et al. 2007). SSCPs were used for two chloroplast intergenic regions that displayed variation within California. Two Californian haplotypes were separated by a SNP at the intergenic region spanning *atpB* and *rbcL*. A third Californian haplotype differed from the first two at several loci of which an indel was screened for the intergenic region between *psaI* and *ycf4*. Double restriction digests of locus *atpB-rbcL* amplicons, with *HpaII* and *RsaI*, and of locus *psaI-ycf4* with *SspI* and *AluI*. These enzymes were chosen from the chloroplast sequence data to isolate the SNP and indel into smaller fragments conducive to SSCP assays. The fragments from double digests were denatured for 10 minutes at 95°C, snap frozen, and run on non-denaturing polyacrylamide gels for 17.5 hours at 502 V, 14 mA, and a constant wattage of 8W. I attempted to screen 5-10 accessions from sites that had at least one individual sequenced for chloroplast loci, and one accession at each of the remaining 71 sites. This two-level sampling scheme allowed us to evaluate the potential of admixture within sites, and also broad population structure between sites in California. In total, 204 samples were screened for SSCPs. All of the Californian accessions for which there was already chloroplast sequence data available were also assayed with SSCPs, and the accuracy of SSCP was confirmed for relevant loci (Supplementary Materials in publication 1 & 2).

2.3.4 Analysis

Trace files for each chloroplast locus were imported to Geneious v. 5.4.4, and aligned with each other using the MUSCLE algorithm within Geneious, default settings, and confirmed by eye. Traces with QV scores < 20 were discarded and not used in further analyses. Every character change (SNP or single indel) for each chloroplast locus was treated as an independent binary character. The only exception was the trnL-F (UAA) locus, which had several indels of multiple adjacent basepairs; these were treated as one multistate character. All polymorphic sites were then used to construct a maximum-likelihood phylogenetic tree using PhyML 3.0, iterated for 10,000 bootstraps. In order to doubly verify cpDNA tree topology, I also constructed a Bayesian phylogenetic tree using Mr. Bayes v. 3.2. For the construction of both trees, I used the HKY85 (Hasegawa, Kishino, and Yano) model of evolution. Other models were briefly explored, but there was so little cpDNA variation that each yielded identical tree topologies. Chloroplast trace files were blasted and annotated using the web server tool CpGAVAS, and uploaded to NCBI using the web server tool – BankIt.

For the Californian accessions assayed with SSCPs, I modeled the relative frequency of chloroplast haplotype at a site with respect to latitude using a generalized linear binomial model using the R package ‘lme4’ (Bates et al. 2013). For this approach, any samples in the same population were treated as non-independent observations. Within California I estimated the overall F_{ST} for individuals at geographic sites assayed with SSCPs (n = 204).

2.4 Results

2.4.1 Ploidy estimations

Flow cytometry gave 2C DNA content clustered at values of 8, 16 and 24 pg, (Table 2.2) implying variation in ploidy. *A. barbata* is tetraploid and 16pg was the most common 2C content, so accessions that had approximately 8pg and 24pg were inferred to be diploids and hexaploids, respectively. These genome size estimates are well within previously reported estimates of other *Avena* species (Bennett & Leitch 2005). All Californian accessions were tetraploid. However, I observed six diploid and two hexaploid accessions among the 48 old world accessions. The ranges of the genome size (pg/2C) of Californian and tetraploid Old World accessions overlapped (Table 2.2), but Californian accessions have approximately 1.5% smaller genomes on average than tetraploids from the old world (permutation test $Z = 3.9626$, $p = 0.00001$) (Fig. 2.1). Inferred ploidy is mapped onto our phylogeny (Fig. 2.2).

Table 2.2 Broad grouping of accessions evaluated with flow cytometry, number of plants assayed, mean 2C DNA content (pg/2C), the minimum and maximum range of the mean DNA content, inferred ploidy, and standard error (SE). See supplementary material (forthcoming publication) for detailed genome size information.

Group	Sample size	Mean 2C DNA content (pg/2C)	SE	Range of 2C DNA content (pg/2C)	Inferred ploidy
All Californian accessions	24	15.97	0.03	15.62 - 16.36	4x
Tetraploid old world accessions	40	16.21	0.03	15.68 -16.69	4x
Diploid (Clade 1 – Spain)	3	8.26	0.05	8.16 - 8.32	2x
Diploid (Clade 2 – Morocco, Greece, Spain)	3	8.65	0.10	8.48 - 8.83	2x
Babylon, Iraq and Giza, Egypt	2	24.99	0.56	24.43-25.55	6x

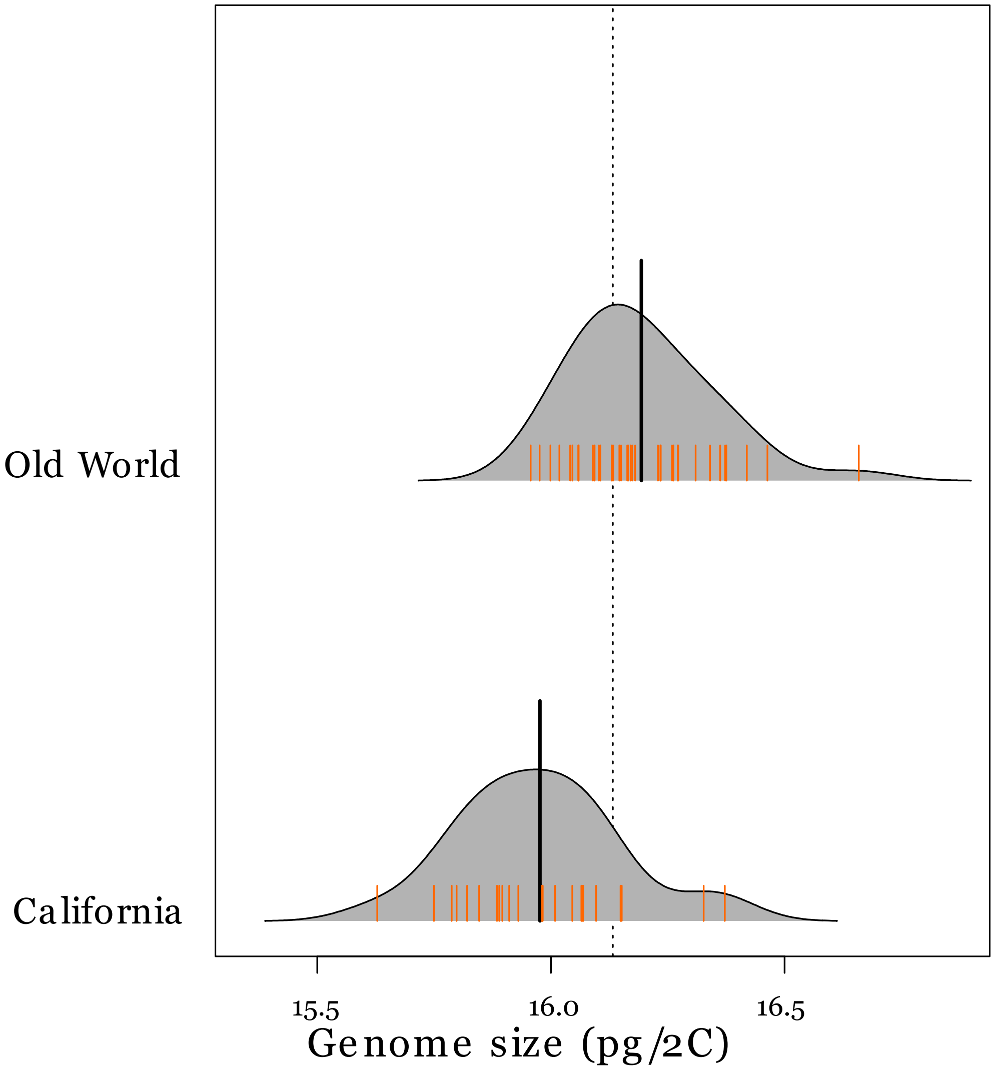


Figure 2.1 Probability density plot (grey-shaded region) and rug plot (orange hatch marks) of genome size values (pg/2C) for tetraploid *A. barbata* from the Old World (n = 40), and California (n = 24). The thick black line in the middle of each density plot is the median value for genome size. The dotted line is the overall median value.

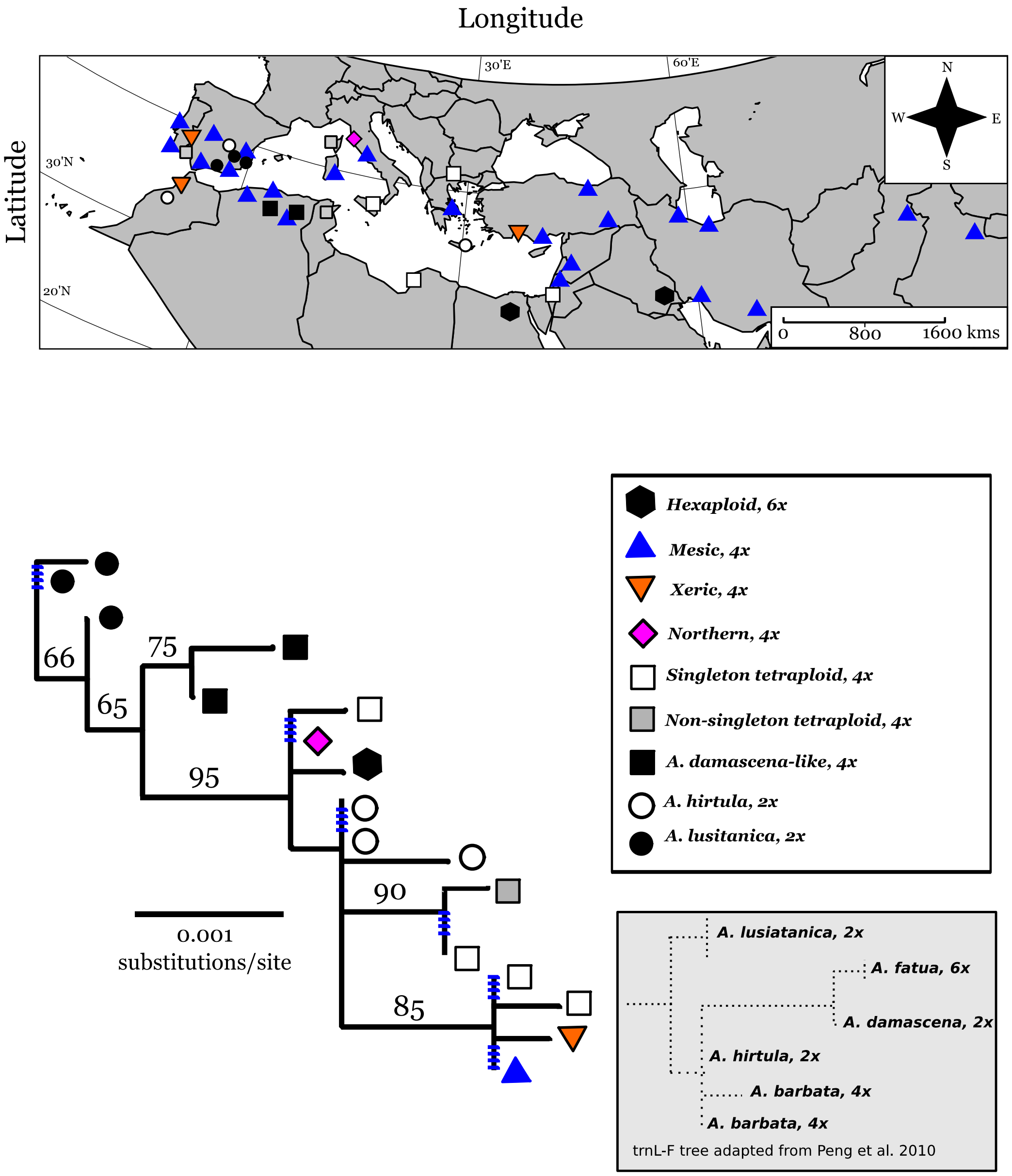


Figure 2.2 Map (Lambert azimuthal equal-area projection) of the Old World accessions. The haplotypes in colour on the map represent the accessions that occur in both the Old World and California. Maximum-likelihood phylogenetic tree based on 100,000 re-samplings. Bootstrap support is indicated at nodes. The tree was constructed using all informative chloroplast sites of seven loci. The ploidy for each haplotype is mapped onto the tree, not included as a character in the phylogeny. Inset tree is a drawn rough approximation of Peng et al.'s (2010) tree for context in explaining our phylogeny's hypothesized rooting. Blue hatched marks are collapsed branches, generally within a clade. Map of broad categories chloroplast haplotypes of Old World (European and Asian) wild oat accessions. "Other tetraploids" are not necessarily identical to each other, but all mesic and xeric haplotypes are identical to each other. See discussion for further elaboration on *A. damascena-like,* and *A. lusitanica* types.

2.4.2 Phylogeography in the Old World

From seven chloroplast loci, I obtained 3250 bp of chloroplast sequence from which I found a total of 18 different cpDNA haplotypes across *A. barbata*'s range. Overall, there was very little cpDNA variation, and total chloroplast sequence divergence was 0.86 % for all accessions in our phylogeny. Of the roughly six major clades in our phylogeny four were well supported (bootstrap support proportion of 75 or greater). However, the different haplotypes within each of these major clades were not well differentiated from each other with polytomies occurring in each clade (Fig 2.2). The most widely distributed haplotype in the Old World matched that of the mesic genotype. A haplotype matching that of the xeric genotype was very closely related to the mesic, being separated by only a single SNP. The xeric haplotype was present in three Mediterranean sites – for consistency, I refer to these haplotypes as "mesic" and "xeric". Most other Old World haplotypes were unique single point occurrences (in Fig. 2.2 I label these as "singleton tetraploids"), with the exception of the identical haplotypes isolated from accessions in Portugal (CN 25800), Corsica, France (PI 337963), and Tunisia (CN 19364).

Three diploid Spanish accessions belonged to one well-supported clade closely related to tetraploid Algerian accessions. The trnL-F/trnF-R sequences for these clades closely matched trnL-F/trnF-R sequences for the diploid *Avena lusitanica*, and the diploid *Avena damascena*, respectively (Peng et al. 2010). I therefore root our tree to these groups.

The other three diploids from Greece, Spain, and Morocco are basal to the clade containing the mesic and xeric haplotypes, along with other tetraploids (Fig. 2.2). The trnL-F/trnF-R sequence of these diploids matches those of *A. hirtula* and *A. barbata*, which had identical trnL/trnF sequences in Peng et al. (2010) *A. hirtula* is the inferred diploid A genome ancestor of the AB tetraploid *A. barbata* (Allard et al. 1993). I suggest the diploid accessions in this clade are likely *A. hirtula*.

The two hexaploids (found in Iraq and Egypt) had the same haplotype, which is also basal to the main clade. These two hexaploids do not match the trnL-F sequences of hexaploid *A. fatua* (or any other *Avena* species) from Peng et al.'s (2010) phylogeny.

Finally, all haplotypes in the main tetraploid clade containing the mesic and xeric haplotypes have trnL-F sequences matching those of Peng et al. (2010) *A. barbata* sequences.

2.4.3 Introductions and population structure in California

Three chloroplast haplotypes were identified in California. All three of these were also observed in the Old world (Figs. 2.2 & 2.3). The xeric and mesic haplotypes were of course present in California, given their association with the mesic and xeric allozyme genotypes of Allard et al. (1972). However, a third haplotype was found at sites in Northern California. One accession collected from Livorno, Italy had a chloroplast haplotype identical to that of this "Northern" haplotype (Fig. 2.2), and this haplotype was distantly related to the mesic and xeric haplotypes.

The SSCP approach, was able to differentiate between mesic and xeric haplotypes, and northern and mesic haplotypes (Supp. Material in forthcoming publication). The SSCPs revealed 106 accessions possessing the xeric allele at marker ANU 67-L/68-R, distributed mostly at southern latitudes (Fig. 2.3). I was then able to further differentiate 74 accessions that had the mesic allele from 24 Northern alleles at marker ANU 73-L/ANU 74-R in California (Fig. 2.3). Chloroplast sequence divergence = 0.28% for Californian haplotypes, in comparison with 0.86% sequence divergence of all accessions, which is more than a 3x reduction in sequence divergence for Californian haplotypes.

Of the 24 sites that were screened for more than one individual, 12 of these sites were polymorphic (Fig. 2.3). However, there is substantial spatial structure as to how this variation is distributed in California. I estimated $F_{ST} = 0.33$ based upon haplotype frequencies. The generalized linear binomial mixed model indicates a latitudinal cline in the distribution of haplotypes latitude ($\beta = -0.34$, $p < 0.0001$). The xeric haplotype is more likely to occur at sites in southern California, while northern locations show a higher frequency of the other two haplotypes.

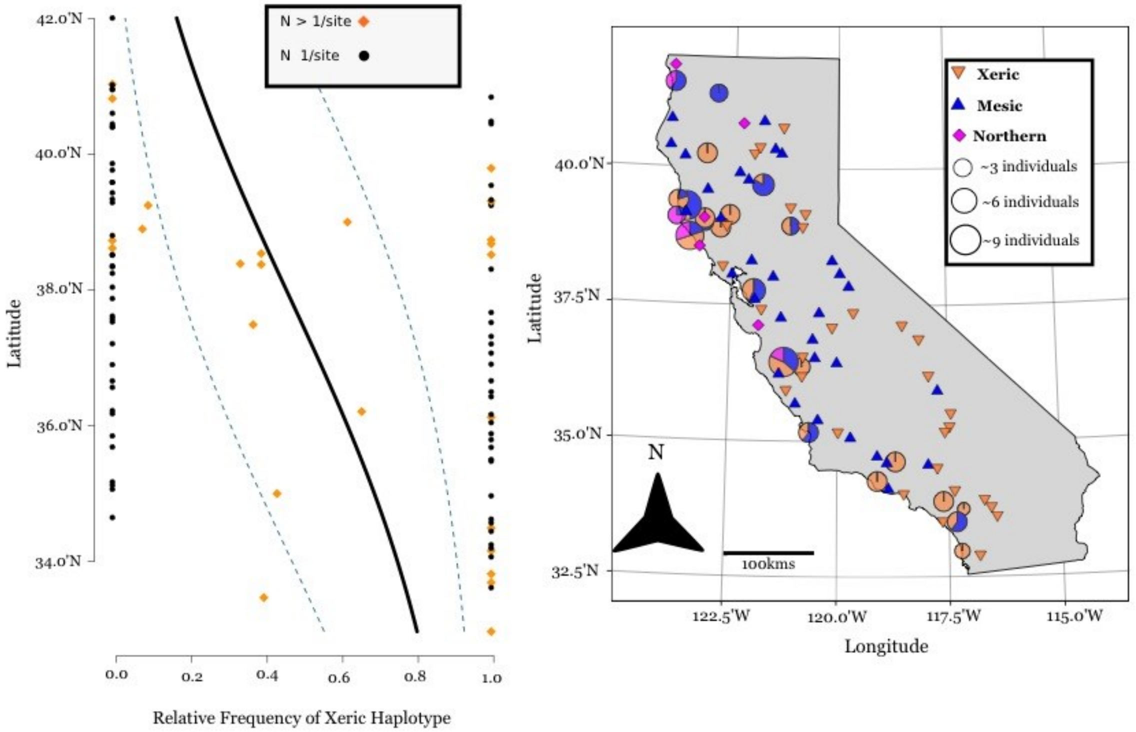


Figure 2.3 Left panel is the linear logistic regression of the relative abundance of the xeric cpDNA haplotype on latitude, the 95% confidence bands based on the logistic distribution. Note the latitude is on the y-axis for comparison with the map. Right panel shows the geographic distribution of haplotypes in California (Lambert azimuthal equal-area projection). Pie charts represent 24 sites where multiple individuals were assayed, and are sized relative to the number of individuals sampled at a site. Triangles represent 71 sites where one individual was sampled per site.

2.5 Discussion

Evidence from cpDNA sequences point to a minimum of three introductions from the Old World to California. I had originally expected at least two introductions because of the two previously described genotypes (Clegg & Allard 1972), and indeed, separate introductions of the mesic and xeric allozyme genotypes are seen. However I also discovered a third haplotype that was mostly confined to Northern California. All three haplotypes in California are also observed in the western Mediterranean, indicating that there was a minimum of three distinct lineages introduced to California. The presence of multiple cpDNA haplotypes is a necessary precondition for recombination to contribute to adaptation/colonization success in the new range (Ellstrand & Schierenbeck 2000).

The haplotypes within California show substantial large-scale among-population structure largely due to a statewide shallow North-South cline (Fig. 2.3). There are xeric haplotypes at Northern latitudes California, though their distribution is largely concentrated at southern latitudes. Mesic haplotypes are found mostly in the Northern region of California, but they are also found at many southern sites. This gradual cline of xeric haplotypes indicates that residual population structure remains from the 1970s, where during this time period xeric allozyme genotypes were predominant throughout California, and especially at southern latitudes (Clegg & Allard 1972).

But there is also clear spatial mixing within populations. The presence of different haplotypes within populations increases the probability that hybrid recombination could occur through occasional outcrossing between different selfing lineages of *A. barbata*. At the within-population level, spatial mixing of cpDNA lineages occurs most frequently in North-central populations close to San Francisco and the surrounding Bay Area. Interestingly, the Northern haplotype is quite distantly related to the other two Californian haplotypes (Fig. 2.2), so it is possible that it is introducing additional new nuclear alleles to the Californian populations of *A. barbata*, (not present in the mesic and xeric genotypes) from which additional new recombinant genotypes may be emerging. The occurrence of all three genetic lineages within populations; (e.g., Geyserville, Bodega Bay, Marshall, and San Ardo) allows for the possibility that recombination could greatly enhance genetic variation.

The three introductions to California may have occurred at different times or concurrently. With too little divergence time having passed since the presumed original *A. barbata* introduction to California (~200 years ago) (Jain & Marshall 1967; Minnich 2008), only small differences in our chosen cpDNA loci (i.e., no mutations have occurred in California, which would allow us to track movements within the invaded range), it is impossible to infer from the phylogeny which introduction came first. However, previous field studies of the allozyme genotypes and their recombinants suggest that mesic genotypes have higher fitness than the xeric type (Latta 2009). This leads to the prediction of the spread of genotypes derived from the mesic haplotype into areas formerly occupied by the xeric, especially in the northern region of California (Latta 2009). This would suggest that the mesic was introduced after the xeric.

Since neither multi-locus allozyme combination occurred in the old world (Garcia et al. 1989), the allozyme combinations described by Allard et al (1972) were thought to be recombinants of alleles present in Spanish populations. These findings support the idea that recombination is relevant to colonization and that it contributes to adaptation. However, if these recombinant allozyme combinations were separately introduced to California, as I speculate above, then hybridization may have occurred in *A. barbata* populations colonizing habitats further south in South or Central America which were subsequently transported to California (Blumler 2000). I cannot discern whether this initial recombination might have happened in California, or prior to arrival, because our geographic sampling is restricted to North America and the Mediterranean Basin. Further, as there were no new cpDNA haplotype mutants observed in California since leaving the Old World I would not be able to track the route of recombination or migration.

While flow cytometry data show that all Californian populations (accessions) are tetraploid, one conspicuous result was that the genome sizes (pg/2C values) of Californian tetraploid individuals were on average smaller (~1.5%) than those of the Old World tetraploid individuals. Lavergne et al. (2009) have demonstrated that a reduced genome size for plants is adaptive in novel or stressful environments (as would be experienced during an invasion) and is associated with a number of phenotypic traits, such as rapid cell division during stem elongation, that facilitate invasion in *Phalaris arundinacea* (reed canarygrass). Some have argued that intraspecific variation may be an artifact of measurement error (Greilhuber 1998; Dolezel & Bartos 2005). But our samples were grown under similar conditions and measured on the same machine, at the same time, with the same size standards. I therefore think that these highly statistically significant differences between Old World and Californian oats are genuine and worth further testing the hypothesis of reduced genome size in an invading population (Lavergne et al. 2009).

2.5.1 Old World

Our main purpose in examining the Old World was to determine the number and divergence of lineages introduced to California. While our Old World sampling was not intensive, it was more than sufficient to identify the main branches of the phylogeny (Nielsen & Slatkin 2013). The mesic haplotype was widespread and the most commonly found haplotype in the Old World - occurring as far east as India and as far west as coastal Portugal. Based on our hypothesized rooting of our cpDNA tree, with the diploid *A. lusitanica* haplotypes, the mesic haplotype also appears to be one of the most derived. This is the opposite of what is predicted under coalescent theory, with the baseline expectation being that the most abundant and/or widespread haplotype is usually basal (Templeton et al. 1995). This pattern suggests a recent and rapid spread of *A. barbata* individuals bearing the mesic cpDNA haplotype. Such a spread may have occurred if rapid expansion of the species range in the Old World had reduced genetic drift creating one widespread and abundant haplotype, as well as an excess of rare haplotypes (Excoffier et al. 2009), and our data appear to fit these expectations.

Although the large majority of our Old World accessions were tetraploid, a few were not. As every accession I obtained from germplasm repositories was originally identified and labeled as "*A. barbata*", I used Peng et al. (2010) trnL-F phylogeny of *Avena* species to infer a plausible root (*Avena lusitanica*) consistent with the larger *Avena* phylogeny. The *Avena* samples I investigated in the Old World seem to be part of a polyploidy complex or a series of repeated polyploidizations with multiple origins, which is not uncommon for plants (Soltis & Soltis 1999; Soltis *et al.* 2007; Husband *et al.* 2013). Near the hypothesized root of our tree, the Algerian samples from Djelfa and Batna are most closely related to the diploid *Avena damascena* sequences from Peng et al. (2010), yet our accessions are tetraploid, so I refer to these accessions as *Avena damascena-like* (Fig. 2.3). I hypothesize that these tetraploids could potentially be an intermediate between the diploid *Avena damascena* (2x) and the derivative hexaploid *Avena fatua* (6x), based the placement of these species in Peng et al. (2010) phylogenies. Our phylogeny contains no accessions that matched the trnL-F-trnF-R sequence of *A. fatua*. The only hexaploids from our phylogeny were found in Iraq and Egypt, and are closely related to the tetraploid Northern *A. barbata* haplotype and they are distantly related to hexaploid *A. fatua* from Peng et al. (2010).

The Northern haplotype is tetraploid, but quite divergent in cpDNA sequence from the mesic and xeric haplotypes. One could argue that the Northern haplotype could be another tetraploid oat species such as *A. abyssinica* or *A. vaviloviana,* which cluster closely with *A. barbata* and have identical trnL-F sequences (Peng et al. 2010). However, it is difficult to separate the whole *A. barbata/abyssinica/vaviloviana* group and some authors have proposed it should all be considered *A. barbata* (Rajhathy & Thomas 1974). These inferences, while worth revisiting to clarify phylogeographic hypotheses, are speculative, and were not the main focus of our study. Whether the Northern cpDNA haplotype is considered a separate species capable of hybridizing with A. barbata, or a distant lineage within one species complex does little to alter our main conclusion that there were three genetic lineages introduced to California. Within California there is almost certainly intermixing between the three introductions, especially in the North due to the presence of multiple haplotypes within populations. This sets up the possibility to test for recombination between the three different introductees, and the possible adaptive spread of a new recombinant genotype(s) to novel environments.

2.6 References

Allard RW, Babbel GR, Clegg MT, Kahler AL (1972) Evidence for coadaptation in Avena barbata. *Proceedings of the National Academy of Sciences of the United States of America*, **69**, 3043–3048.

Allard RW, Garcia P, Saenz-de-Miera LE, la Vega de MP (1993) Evolution of multilocus genetic structure in Avena hirtula and Avena barbata. *Genetics*, **135**, 1125–1139.

Baker HG (1955) Self-compatibility and establishment after“long-distance”dispersal. *Evolution*, **9**, 347–349.

Baker HG (1967) Support for Baker's law-as a rule. *Evolution*, **21**, 853–856.

Barrett S, Colautti RI (2008) Plant reproductive systems and evolution during biological invasion. *Molecular Ecology*, **17**, 373–383.

Bennett MD, Leitch IJ (2005) Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Annals of Botany* **95**:45–90.

Blumler MA (2000) Spatial analysis to settle an unresolved question in genetics, with both theoretical and applied implications. *Research in Contemporary and Applied Geography: A Discussion Series*, **24**, 1–42.

Clegg MT, Allard RW (1972) Patterns of genetic differentiation in the slender wild oat species Avena barbata. *Proceedings of the National Academy of Sciences*, **69**, 1820–1824.

Corriveau JL, Coleman AW (1988) Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *American Journal of Botany*, **75**, 1443–1458.

Daehler C, Strong D (1997) Hybridization between introduced smooth cordgrass (Spartina alterniflora; Poaceae) and native California cordgrass (S. foliosa) in San Francisco Bay, California, USA. *American Journal of Botany*, **84**, 607–611.

Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: version II. *Plant molecular biology reporter*, **1**, 19–21.

Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology*, **17**, 431–449.

Dolezel J, Bartos J (2005) Plant DNA flow cytometry and estimation of nuclear genome size. *Annals of Botany*, **95**, 99–110.

Doležel J, Cíhalíková J, Lucretti S (1992) A high-yield procedure for isolation of metaphase chromosomes from root tips of Vicia faba L. *Planta*, **188**, 93–98.

Doležel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, **2**, 2233–2244.

Ebert D, Peakall R (2009) A new set of universal de novosequencing primers for extensive coverage of noncoding chloroplast DNA: new opportunities for phylogenetic studies and cpSSR discovery. *Molecular Ecology Resources*, **9**, 777–783.

Ellstrand N, Schierenbeck K (2000) Hybridization as a stimulus for the evolution of invasiveness in plants? *Proceedings of the National Academy of Sciences,* **13,** 7043–7050.

Excoffier L, Foll M, Petit RJ (2009) Genetic Consequences of Range Expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

Garcia P, Vences F, Vega M (1989) Allelic and Genotypic Composition of Ancestral Spanish and Colonial Californian Gene Pools of Avena barbata: Evolutionary Implications. *Genetics*, **122**, 687–694.

Gasser RB, Hu M, Chilton NB *et al.* (2007) Single-strand conformation polymorphism (SSCP) for the analysis of genetic variation. *Nature Protocols*, **1**, 3121–3128.

Greilhuber J (1998) Intraspecific variation in genome size: a critical reassessment. *Annals of Botany*, **82**, 27–35.

Hamrick JL, Godt MJW (1996) Effects of Life History Traits on Genetic Diversity in Plant Species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **351**, 1291–1298.

Hamrick JL, Holden LR (1979) Influence of microhabitat heterogeneity on gene frequency distribution and gametic phase disequilibrium in Avena barbata. *Evolution*, **33**, 521–533.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006) A Lego System for Conditional Inference. *The American Statistician* **60**, 257-263.

Husband BC, Schemske DW (1996) Evolution of the magnitude and timing of inbreeding depression in plants. *Evolution*, **50**, 54–70.

Husband BC, Baldwin SJ, Suda J (2013) The Incidence of Polyploidy in Natural Plant Populations: Major Patterns and Evolutionary Processes. In: *Plant Genome Diversity Volume 2*, pp. 255–276. Springer Vienna, Vienna.

Hutchinson ES (1982) Genetic Markers and Ecotypic Differentiation of *Avena barbata* Pott ex Link. University of California, Davis. Doctoral Dissertation.

Hutchinson E, Price S, Kahier A, Allard R (1983) An experimental verification of segregation theory in a diploidized tetraplold: esterase loci in Avena barbata. *Journal of Heredity*, **74**, 381–383.

Ivanova NV, Fazekas AJ, Hebert PD (2008) Semi-automated, Membrane-Based Protocol for DNA Isolation from Plants. *Plant molecular biology reporter*, **26**, 186–198.

Jain SK, Marshall DR (1967) Population studies in predominantly self-pollinating species. X. Variation in natural populations of Avena fatua and A. barbata. *American Naturalist*, 19–33.

Johansen-Morris A, Latta R (2006) Fitness consequences of hybridization between ecotypes of Avena barbata: hybrid breakdown, hybrid vigor, and transgressive segregation. *Evolution*, **60**, 1585–1595.

Johansen-Morris AD, Latta RG (2008) Genotype by environment interactions for fitness in hybrid genotypes of Avena barbata. *Evolution*, **62**, 573–585.

Latta RG (2009) Testing for local adaptation in Avena barbata: a classic example of ecotypic divergence. *Molecular Ecology*, **18**, 3781–3791.

Latta R, MacKenzie J, Vats A, Schoen D (2004) Divergence and variation of quantitative traits between allozyme genotypes of Avena barbata from contrasting habitats. *Journal of Ecology*, **92**, 51–71.

Lavergne S, Muenke NJ, Molofsky J (2009) Genome size reduction can trigger rapid phenotypic evolution in invasive plants. *Annals of Botany*, **105**, 109–116.

Le Roux JJ, Wieczorek AM, Wright MG, Tran CT (2007) Super-Genotype: Global Monoclonality Defies the Odds of Nature (S-H Shiu, Ed,). *PLoS ONE*, **2**, e590.

Marshall DR, Allard RW (1970) Maintenance of Isozyme Polymorphisms in Natural Populations of Avena barbata. *Genetics*, **66**, 393.

Minnich R (2008) *California's Fading Wildflowers*. University of California Press, Berkeley and Los Angeles.

Nielsen R, Slatkin M (2013) *An Introduction to Population Genetics: Theory and Applications*. Sinauer Associates, Sunderland, MA.

Novak SJ, Soltis DE, Soltis PS (1991) Ownbey's Tragopogons: 40 Years Later. *American Journal of Botany*, **78**, 1586–1600.

Peng Y-Y, Wei Y-M, Baum BR *et al.* (2010) Phylogenetic investigation of Avena diploid species and the maternal genome donor of Avena polyploids. *Taxon*, **59**, 1472–1482.

Pérez de la Vega M, Garcia P, Allard RW (1991) Multilocus genetic structure of ancestral Spanish and colonial Californian populations of Avena barbata. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 1202–1206.

Prentis PJ, Wilson JRU, Dormontt EE, Richardson DM, Lowe AJ (2008) Adaptive evolution in invasive species. *Trends in Plant Science*, **13**, 288–294.

Price SC, Jain SK (1981) Are inbreeders better colonizers? *Oecologia*, **49**, 283–286.

Rajhathy T, Thomas H (1974) *Cytogenetics of oats (Avena L.)*. Ottawa : Genetics Society of Canada, 1974., Ottawa.

Schierenbeck KA, Ellstrand NC (2009) Hybridization and the evolution of invasiveness in plants and other organisms. *Biological Invasions*, **11**, 1093–1105.

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution*, **14**, 348–352.

Soltis DE, Soltis PS, Schemske D *et al.* (2007) Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon*, **56**, 13–30.

Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology*, **17**, 1105–1109.

Zhang Q, Liu Y, Sodmergen (2003) Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant and Cell Physiology*, **44**, 941–951.

Chapter 3 Gaussian Random Fields: Using a novel SDM technique that incorporates uncertainty to describe a niche shift from Iberia to California

3.1 Abstract

Assessing whether a niche shift has occurred during a biological invasion is important for understanding evolutionary dynamics, as different environments would tend to exert different selective pressures. Climatic niche shifts are generally thought to be rare among invaders. Using a novel species distribution model (SDM), Gaussian random fields (GRF) I investigated niche shifts and niche overlap in *Avena barbata*, an annual, invasive grass that has colonized all five Mediterranean climate regions. I constrained my investigation to two of these regions, the state California (where it is invasive), and the Iberian Peninsula (where Californian populations are thought to have originated). Occurrence data was obtained from the Global Biodiversity Information Facility (GBIF), and data from our own Californian re-survey in 2010. I examined between 11-12 climatic variables and one edaphic variable in modeling the niche of *A. barbata* using three different spatial sampling schemes to account for spatial bias in occurrence data. Different models were trained on subsets of data in both ranges and then used to predict probability of presence in both ranges, and isolate environmental variables that were important in predicting presence. I also used a two-factor MANOVA and linear discriminant functions to assess niche shifts and overlap in *A. barbata.* Overall, I found that models predicting to the opposite range had a large amount of uncertainty associated with them, MANOVAs showed significant difference between means of environmental variables in both ranges predicting presence and absence, and linear discriminant functions showed little overlap between geographic ranges. I interpreted this as evidence of a possible niche shift between Iberia and California, which may suggest that *A. barbata* is undergoing evolutionary change.

3.2 Introduction

The frequency of occurrence of climatic niche shifts during or throughout biological invasions is unclear. At least one review proposes that niche shifts are the exception as opposed to the rule (Guisan et al., 2014). Climatic niche overlap and conservatism may emerge as the most persistent findings because characterizations of the home and invaded niches focus mainly on climatic factors (precipitation and temperature). Generally, there can be fair amount of climatic overlap for different geographic regions, and particularly Mediterranean climates (Ackerly, 2009), and even more expansive geographic regions. For example, a meta-analysis of 50 Holarctic terrestrial invasive plants found that 80% of these have substantial niche overlap and thus niche stability when compared to their home range (Petitpierre et al., 2012; Guisan et al., 2014). Niche conservatism or niche stability occurs when a species in one geographic region moves (sometimes thousands of kilometers) to another geographic region, but continues to occupy the same niche or environmental conditions as it did in its geographic place of origin, in contrast, a complete niche shift is when a species either occupies the same or different geographic space, but shifts its environment or its niche (Guisan et al., 2014).

The main objective of species distribution models (SDMs) is the prediction of presence and absence of one or many species given a set of environmental conditions. In doing so, these models characterize - however imperfectly - the niche (Hutchinson, 1959). SDMs have been used to highlight niche changes or overlaps in species distributions/niches at different stages of colonization or invasion of new ranges (Václavík & Meentemeyer, 2011), predict new potential niches with respect to climate change (Pearman et al., 2008), and errors produced from SDMs may be informative as to whether the species has evolved or undergone a niche shift (Fitzpatrick et al., 2007).

Ecological studies using SDMs in the study of invasions or range expansions, may come to a crossroads with evolution, and either directly or indirectly address the question of local adaptation to a new niche, in that the invaders would experience low fitness back in their home range (Godsoe et al., 2009; Yoder et al., 2014). The process of local adaptation and divergent selection would inherently imply a species has undergone a niche shift, where certain genotypes of a species have higher or lower fitness in different environments. Niche shifts do not necessarily indicate adaptation, as a species may have always possessed the genetic resources or plasticity for a particular trait such as flowering time (Levin, 2009), but merely not had the opportunity for transplant to a new niche. There may also be eco-evolutionary dynamics, where release from ecological competitors (release from selection) affects the amount of genetic variation available in the population for future generations (Emery & Ackerly, 2014).

Invasive populations often suffer founder effects and the loss of *adaptive* (as opposed to neutral) genetic variation due to genetic bottlenecks relative to native populations. As such an invasion could result in four possibilities with respect to niche occupancy. First, the newly invaded niche may be home to a subset of genetic variation (niche contraction) that exists in the home range. Second, the newly invaded niche may be broader than the old niche (niche expansion). Third, there could also be some limited overlap between niches, but a niche shift occurs in the new habitat. Finally, there is also the possibility that there has been no niche shift, as environmental variables in either region show strong overlap with each other.

Avena barbata Pott ex Link is an annual grass that is invasive to California, and highly selfing. Following colonization in California, it was observed that there were few sets of monomorphic allozyme genotypes present and few recombinants (Clegg & Allard, 1972; Allard et al., 1972; Hamrick & Allard, 1975). It was further demonstrated that these allozyme genotypes were composed of a reduced subset of alleles of those found in Iberia – their suspected place of origin (Garcia et al., 1989). These observations suggest that introduced populations of *A. barbata* to California experienced a bottleneck and lost genetic variation during colonization. Further, it was observed that *A. barbata* individuals in California occupy populous annual stands unlike the small patches of ruderal stands it forms in its Iberian home range (Jackson, 1985), which may be indicative of a niche shift.

If adaptive genetic variation were lost during a or repeated bottleneck(s), then this leads to the prediction that individuals are restricted in occupying the full breadth of the niche occupied by their Spanish ancestors, i.e. SDMs trained in Iberia would over-predict the niche in California. But if, on the other hand, multiple introductions have occurred in California (Crosby et al. 2014 Ch.2), and these introductions have recombined (Ch.4) – the resultant new combinations might allow for niche overlap with Iberia, and possibly a niche shift if the invaders have evolved since being introduced to California. If this were the scenario, than models trained in Iberia would under-predict the niche.

In my study I sought to characterize the amount of niche overlap and niche shifts between invasive populations of *A. barbata* and native Iberian populations, in doing so, I also characterize the niche in both ranges. I apply a new method of SDM, using Gaussian random fields (GRF) (Golding, 2013), which is one of the only methods that deals well with "presence-only" data. I sought to characterize the niche in Spain, using a GRF model, and then predict the probability of presence in California using this model. I then repeated the process of modeling the niche in California and used this model to predict the probability of presence in Iberia.

3.3 Methods

3.3.1 Occurrence data: native range & invasive range

A. barbata populations are found on six continents with the exception being Antarctica, and show a preference for characteristic Mediterranean biomes. The Mediterranean Basin is thought to be its native range, but this may also extend well into the Middle East (Baum 1977). It has invaded Mediterranean biomes in Australia, North America (California), South America, and South Africa. As I was mostly interested in evaluating potential niche shifts and niche conservatism of *A. barbata* populations in California, I chose to define the "native range" as the Iberian Peninsula plus Northern Morocco, and the "invasive range" as the state of California. The Iberian Peninsula and Northern Morocco are thought to be the source of present day Californian populations, with initial colonization occurring in California during the Spanish missionary period on the Western side of the Americas (Garcia et al., 1989; Minnich, 2008). Additionally, both these regions were intensively surveyed, over almost the same period (1800s-2000s) most other regions were inconsistently or sparsely sampled. The only other region with comparatively intense sampling was the continent of Australia with over 3000 records of occurrences; I plan on revisiting this dataset at a later date.

Total *A. barbata* occurrence data was obtained from several sources: the public data portal for the Global Biodiversity Information Facility (GBIF) using the r-package 'rgbif' v. 0-5 (Chamberlain et al., 2014), the National Small Grains collection at the USDA, the California Consortium Herbaria, and our lab's own geographic survey (see Chapter Four) of the state of California in 2010. The total number of records of occurrences from this initial global search was 14554, with occurrence dates ranging from the early 1800s – 2014 in all environments. I screened the data as GBIF interfaces regularly with many databases and data portals, and this interfacing can result in duplicate records. I purged duplicates by keeping only one record for each unique pair of GPS coordinates. I also ensured that remaining occurrences were georeferenced, correctly – i.e. were in the correct country, and not in the ocean, by spatial query using the R-package 'raster', and by eye. The total number of occurrences for native and invasive ranges was 3916 with 998 occurrences in the invaded range, and 2908 in the native range.

3.3.2 Sampling bias and the generation of background (aka 'pseudo-absence') data

There are two main problems that can occur with SDMs sampling bias and the lack of "true" absence data in occurrence datasets. Sampling bias occurs because of over-sampling regions close to populated urban areas, roadways, and other easily accessible areas such as parks, and may create spurious associations between environment and occurrences when attempting to improve the predictability of an SDM. Sampling bias may also occur because a concentrated sampling effort has taken place due to various environmental, provincial, county, or region-wide biodiversity assessment programs.

To deal with sampling bias, Philips et al. (2009) suggest using a species that co-occurs with the species of interest, but may occupy a slightly shifted different distribution. I found this suggestion to be cumbersome to fulfill in California, given that *A. barbata* is invasive there and does not co-occur with a related species throughout both its ranges. Instead, I used three approaches to account for sampling bias. First, I subsampled both Iberia and California by overlaying a raster grid of varying sizes (20, 15, 10, 5-kms ^{2}) across the entire range and chose one individual in each grid square. By eye, I judged which of these grid sizes appeared to prune the data sufficiently in areas with many occurrences, but without reducing the total number of occurrences throughout the range. In both cases, a raster grid of 15-km^2 was chosen as a grid size. Second, I targeted my sub-sampling on only those regions in each range that may have been repeatedly sampled due to their proximity to populated areas or political boundaries. I overlaid a 15-km^2 raster grid over each partial range, and chose only one individual within each grid cell. Finally, because the data were obtained mostly from databases, I could not be certain as to whether these areas were sampled in a biased way or whether these represented actual densely occupied habitats. Thus, SDMs were also performed on datasets without systematically eliminating any of the occurrence data.

SDMs produce more accurate results if absence data recorded from a formal geographic survey is used, and unfortunately, no (Iberia) and/or little (California) were available for *A. barbata*. An alternative to using actual absence data to improve SDM accuracy is to include a large number of user-generated pseudo-absences (Barbet-Massin et al., 2012). I generated "pseudo-absences" as background data across each of the ranges. I generated an equal number of "absences" as there were "presences", and these were generated randomly across the range, in the following way: for each occurrence point, I drew a circle with a radius of 50-km^2 (or 4km) and one random point within this circle was designated an "absence".

3.3.3 SDM technique - GRaF

As there was both presence and pseudo-absence data points at my disposal, I used a new SDM technique to evaluate my selected environmental variables (see section below for selection process). Environmental variables were fit to SDMs based on Gaussian Random Fields (GRF) in the R-package 'GRaF' v. 0.1-12 (Golding, 2013). As using GRFs to model species distributions is a new method, I explain it here in a bit more detail.

GRF is a Bayesian method that provides the uncertainty of each estimate of probability of presence or absence associated with the SDM due to imperfect data. Most other SDM methods are constructed with imperfect data without accounting for it. The first step in constructing a GRF is generating an array of presences and pseudo-absences (coded as 1s and 0s) along with the environmental variables of interest at each coordinate; the independent variables are the environmental variables. In a GRF model, the mean is specified from the predictors and the response variable modeled in terms of errors from the mean. Multi-dimensional Euclidean distances are used in calculating the n-dimensional distance between locations; these distances were then converted into squared-exponential covariance functions. After computing the covariance function, GRaF provides another parameter "lengthscale" for each environmental covariate in the model. Lengthscale parameterizes how the correlation between presence-absence records decays with environmental distance (Euclidean distance between variables), and can be specified by the user or optimally estimated from the data by GRaF, I chose to optimally estimate the lengthscales. The variables with the shortest lengthscales produced from a GRF model can be interpreted as the main variables driving the relationship between niche and species occurrence. Further detailed information on GRaF is available in Golding (2013).

3.3.4 Initial selection of environmental variables

Another important consideration in generating an SDM with high predictive ability is choosing a representative suite of environmental variables thought to represent the niche of the species of interest. I used a combination of historical observations, literature specific to *A. barbata* (Pinero, 1982) and annual plants (Jackson, 1985), and past observations over four field seasons in California (Latta, pers. comm). While these choices are subjective to a certain degree, I chose variables based on reported life-history traits of *A. barbata*, climate data, and edaphic (soil) traits of each region. For each class of environmental variable I aimed to choose variables that would be relatively consistent for the time period of when the occurrences were recorded 1800s – present, so I could compare the initial colonization period in California (Blumler, 2000; Minnich, 2008), up to present day.

Because *A. barbata* is a winter annual plant, the following observations were used to select an appropriate suite of climatic variables: Nov-Feb precipitation in each region is thought to aid in initial sprouting and growth, along with warmer winter temperatures, a warm wet spring in March-April for initial growth and flowering, and indirectly, a hot, dry, summer is thought to aid in eliminating other vegetation that might otherwise compete with *A. barbata* for resources. Raster grids for mean monthly precipitation and temperature were obtained from the WorldClim database v. 1.4 at a resolution of approximately 1km at the equator or 30 arc seconds. In addition to these climate variables, I also obtained raster grids for mean monthly actual evapotranspiration (AET) and mean monthly potential evapotranspiration (PET) as the difference between AET and PET (i.e. water deficit) are relevant to plant survival.

Soil type as assessed by various classification systems (e.g. USDA and World Reference Base for soil types) may not adequately capture similarities between regions, instead emphasizing categorical differences. As my goal was to be able to compare native and invasive regions, I included one continuous variable, soil pH in a 1 X 1 km grid, which has been shown to greatly improve the predictability of plant SDMs (Dubuis et al., 2012), and is measured using the same method in both locations. Oats and other annual grasses are known to prefer slightly acidic soils, but can exist within a broad range of pH values ~ 3.5-7.

In total, this initial selection left us with thirty-three environmental variables (Table 3.1) to consider for further niche evaluation. All raster grids containing environmental data were clipped to the extent of the range of occurrences in both California and Iberia.

Table 3.1 Summary of raster datasets and environmental variables used in GRF SDMs.

Raster data sets	**Environmental layers contained within**	**Resolution**	**URL**	**Ref.**
Mean monthly Temperature	November – July (9)	~1km at the equator	http://www.worldclim.org/formats	(Hijmans et al., 2005)
Mean monthly Precipitation	November -March (5)	~1km at the equator	http://www.worldclim.org/formats	(Hijmans et al., 2005)
Mean monthly Actual Evapotranspiration (AET)	November – July (9)	~1km at the equator	http://www.cgiar.csi. org.	(Trabucco & Zomer 2010)
Mean monthly Potential Evapotranspiration (PET)	November – July (9)	~1km at the equator	http://www.csi.cgiar.org	(Trabucco & Zomer 2009)
Mean Soil pH	Soil pH (1)	~1km at the equator	http://soilgrids1km.isric.org	(ISRIC, 2014)

3.3.5 Final selection of environmental variables

To reduce the amount of complexity in explaining and interpreting the niche of *A. barbata* in its home range I sought to choose fewer variables that might best explain the niche of *A. barbata*. I examined the degree to which environmental variables were correlated to each other in the Iberian Peninsula and in California. To estimate the degree of correlation between variables, I used Spearman's correlation coefficient at cut-off values of: >|0.99|, >|0.98|, >|0.95|, and >|0.90|, removing one of the correlated variables from the dataset. In Iberia, this reduced the number from variables to thirty-three, twenty-five, twelve, and six, respectively. In California, this reduced the number variables from twenty-eight, twenty-one, eleven, and five, respectively. We iteratively fit GRaF models 100 times to different random samples of 30% of my occurrence data with each different set of variables. In each case, the optimal lengthscale option (as specified above) was turned on, and flat priors were used.

The deviance information criterion (DIC), the number of effective paramters (pD), and the remaining 70% of presence/pseudo-absences that were correctly classified were used to choose the optimal number of environmental variables. Like other information criterion a DIC value estimates the compromise between model fit and model complexity. Lower DIC values are generally preferred to higher DIC values. In Bayesian models, pD is akin to the number of parameters used to explain the data, for example a pD = 10.9 would be akin to a 10th degree polynomial model. Higher values of pD represent a more complex model, lower values a less complex model. I sought to minimize pD, while also maintaining a low DIC. To calculate proportion of correctly classified presences and pseudo-absences predicted using a GRaF model, I set at posterior threshold estimate of 0.5. Any posterior estimate of 0.5 was classified as present, and any estimate below 0.5 was classified as an absence. These classifications were then compared to the actual records of presences and absences to calculate proportion of records that were correctly classified with the GRaF model.

3.3.6 Model prediction and prediction diagnostics

For each of the sampling scenarios each trained Iberian GRaF model was used to predict the remaning (70%) of the records in Iberia, and then all records in California. I then took my Californian trained model and predicted the remaining occurrences in California, and then all of Iberia. The strength of each trained GRaF model was evaluated using three criteria: the area under (AUC) the receiver operating curve (ROC), number of occurrences and absences that were correctly classified, and mean model uncertainty. AUC values scale from 0 to 1 and evaluate a model's ability to correctly classify occurrence records; an AUC value of 1 indicates that the model correctly classifies all records of occurrence, a value of 0.5 is the same as random chance, AUC values < 0.5 perform worse than random chance. But, AUC has faced criticism in the past (Lobo et al., 2008), so I used number of occurrences that were correctly classified based on a threshold value that was the average between model sensitivity (true positive rate) and specificity (false positive rate). Mean model uncertainty was also calculated by subtracting each of the upper 95% credible intervals from the lower 95% credible intervals for each posterior estimate of probability of presence or absence.

3.3.7 Niche comparison

For ease of interpretation with respect to niche comparison and overlap, I narrowed my comparison by focusing on the two variables with the shortest lengthscales (those variables in a GRF model that are driving the relationship between probability of presence and the niche) from each of the training models in both ranges, i.e. four environmental variables in total. These variables were then used for a coarse comparison using a two-factor multivariate analysis of variance (MANOVA) for the factors of occurrence and geographic area, i.e. presence vs. absence and Iberia vs. California. I performed MANOVAs for each of the three spatial sampling schemes. In order assess the amount of niche overlap with respect to these two factors I performed linear discriminant analysis (LDA) for each of the spatial sampling schemes, and also on the complete dataset with all thirty-three environmental variables included for the purpose of niche comparison.

3.4 Results:

3.4.1 Sub-selection of environmental variables

For the initial GRaF analysis where I selected a subset of the original thrity-three environmental variables, all one-way ANOVAs were significantly different for each of the three parameters evaluated for choosing the optimal number of environmental variables (Fig. 3.1 & 3.2).

For both Iberia and California, the comparison of means using Tukey's post-hoc tests revealed that all means for the pD and the number of records correctly classified were significantly different from each other. In Iberia, the mean DIC differed for only the level of six environmental variables; the other levels did not differ from one another (Fig. 3.1). In California, the pairwise mean DIC differed for only the level of five variables (Fig. 3.2). To minimize model complexity (lower pD), but also maintain a good model fit (lower DIC) I chose twelve environmental variables to characterize the home range niche of *A. barbata*; in California, I chose to use eleven variables.

In Iberia, twelve variables correctly classified presences/absences on average 65.6%, a 3% improvement from six variables that classified 62.6%, while models using all thirty-three variables improved the classification by only 2.8%, mean 68.4% (Figs. 3.1 & 3.3-3.5). The chosen twelve environmental variables were: mean actual evapotranspiration (AET) for the months of November, December, May, and July; mean potential evapotranspiration (PET) for March, May, and July; mean temperature for the months of May and July; mean precipitation for the months of February and March; and soil pH (Figs. 3.3-3.5). In general, these variables were thought to be a suitable subset of uncorrelated descriptors of the Iberian Mediterranean environment. There was much more uncertainty around probability of presence posterior estimates in relation to environmental variables for the 15 km2 raster dataset (Fig. 3.4), than there was for the datasets with more occurrence data included (Fig. 3.3 & 3.5). The only common environmental variable with a short lengthscale between any of the spatial datasets was average warm May temperature between 10-15°C (Figs. 3.4 & 3.5).

In California, the eleven variables correctly classified records on average 68.5 %, a 4.7% improvement from five variables that correctly classified on average 63.8% of records, while models using twenty-eight variables improved the classification by only

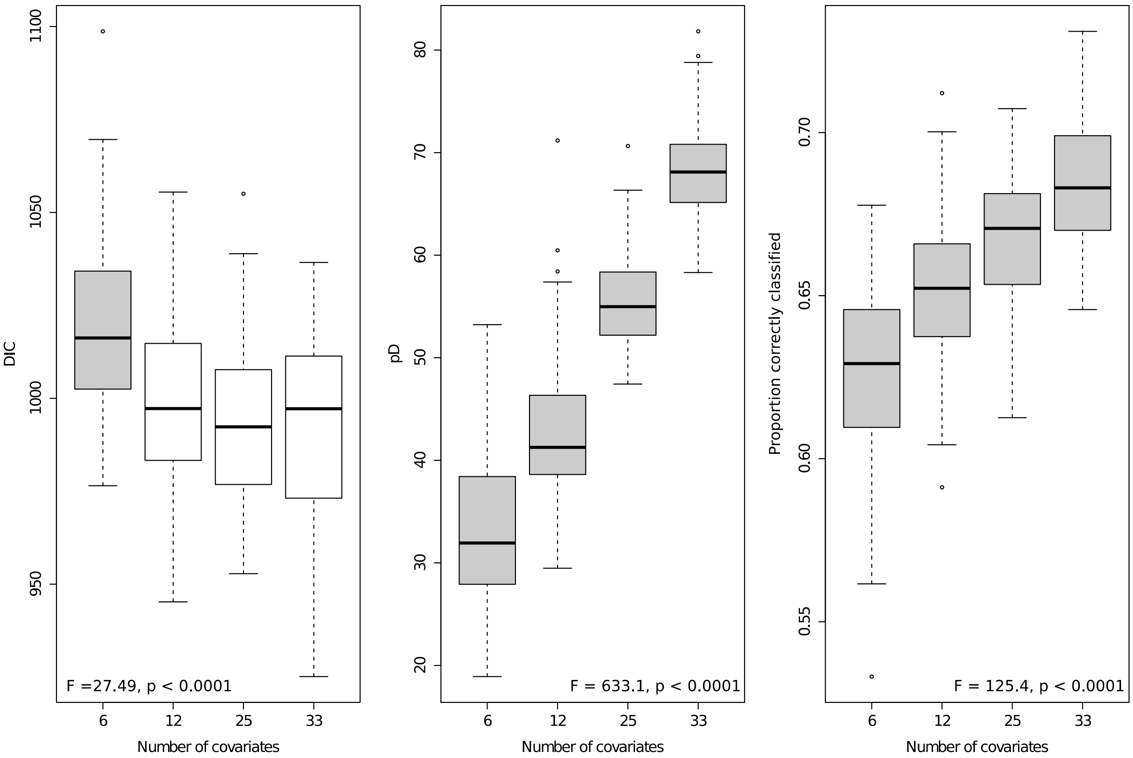


Figure 3.1 Iberian training model diagnostics for DIC, pD (effective number of model parameters) for each level of environmental variables characterizing the niche in Iberia, and the proportion of occurrences correctly classified as “present” or “absent” using a random selection of 25% of the Iberian dataset, iterated 100 times. One-way ANOVA results presented at bottom of each panel, shaded boxplot(s) in each panel indicate significance of Tukey’s post-hoc comparisons.

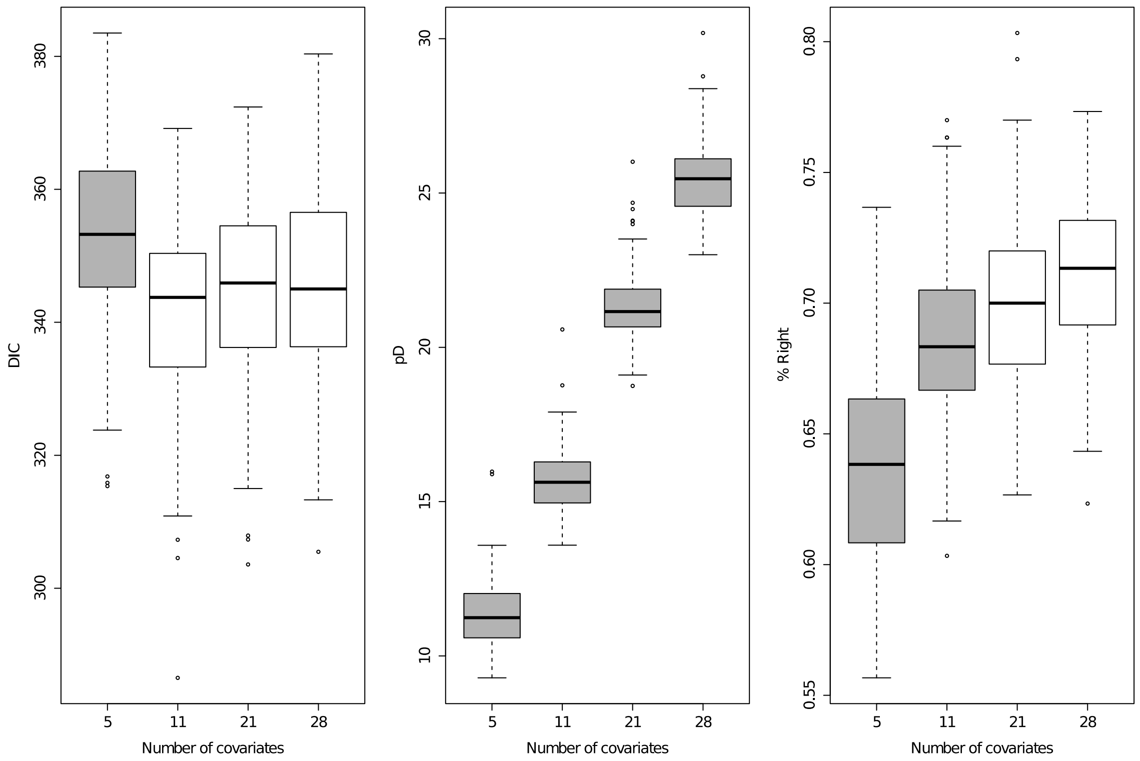


Figure 3.2 Californian training model diagnostics for DIC, pD (effective number of model parameters) for each level of environmental variables characterizing the niche in California, and the proportion of occurrences correctly classified as "present" or "absent" using a random selection of 25% of the Californian occurrence dataset, iterated 100 times. One-way ANOVA stats presented at bottom of each panel, shaded boxplot(s) in each panel reveal the significance of Tukey's post-hoc comparisons.

2.6%, mean 71.1% (Figs. 3.2 & 3.6-3.8). For California, the chosen eleven environmental variables used were: actual evapotranspiration (AET) for the month of January; potential evapotranspiration (PET) for January, February, March, April, and July; mean temperature for the months of December, April, and July; mean precipitation for March; and soil pH (Figs. 3.6-3.8). As in Iberia, there was much more uncertainty with the 15 km^2 raster dataset (Fig. 3.7) relative to the two other spatial datasets (Figs. 3.6 & 3.8), which contained more occurrence data. Mean December temperature had a short lengthscale (high predictive power) in all three Californian occurrence datasets, and predicted a reasonably high probability of presence between mean temperature of 5-10°C (Figs. 3.6-3.8).

GRaF models in both regions had four variables in common: mean monthly AET for November, mean monthly PET for March, mean July temperature, and soil pH (Figs. 3.3-3.8). High probability of presence occurred in both ranges for average July temperatures of ~ 25°C, and for slightly acidic soils ~ 5-5.5 pH (Figs. 3.3-3.8). However when examining the full dataset lengthscale plots, mean monthly PET in March was slightly higher for California than Iberia, as was mean monthly AET in November (Figs. 3.5 & 3.8). This suggests that *A. barbata* occurs in moister environments in California than it does in Iberia.

3.4.2 Testing model performance

The Iberian trained models were more complex than its Californian counterpart (Table 3.2) having a higher number of effective parameters (pD) in the model, and a higher DIC as a result. As expected, for all three spatial sampling schemes, the Iberian trained model correctly classified more occurrence data in Iberia than California, had a higher AUC in Iberia than in California, and had much lower mean model uncertainty in Iberia than in California (Table 3.3). Prediction maps generated using datasets with fewer occurrence records had higher uncertainty (grey points in Figs. 3.9 & 3.10) than datasets where all occurrence data was included. However, when all occurrence data was included in the models, most of the true presence predictions from the Iberian trained model were centered on heavily populated areas in both Iberia and California (Fig. 3.11).

The Californian trained model had a lower number of effective parameters (pD) in the model, and a lower DIC than the Iberian model. It correctly classified more occurrence data in California than Iberia for all three spatial sampling schemes. However, the Californian trained model performed better at predicting occurrences in Iberia the than the Iberian model performed in predicting presences in California (Table 3.3).

3.4.3 Niche comparison

Results of the two-factor MANOVA for all three spatial sampling schemes point to significant differences in the multivariate means between geographic areas, presences and absences, and the interaction between those two single factors. The factor that displayed the most variance between groups was geographic area (Table 3.4) for all three spatial sampling schemes. For every sampling scheme, and all of the data as well, ordinal plots of the first two-discriminant axes analyses revealed separation between geographic areas on the first axis (Fig. 3.12). This separation along the first discriminant axis decreased as fewer records were examined with respect to each of the most important environmental variables, with the spatial data set of one record per 15 km^2 showing the most amount of overlap (Fig. 3.12). The overlap as the amount of records decreases is perhaps a function of model uncertainty; with fewer records the 15-km2 dataset had the higher uncertainty and less resolution in correctly classifying presences vs. absences (Table 3.3). There was separation between presences and absences in the Californian along the second axis for most spatial categories examined (Fig. 3.12), but substantial overlap for Iberian presences and absences, suggesting that Iberian pseudo-absences occur in the same environments as Iberian presences.

Table 3.2 Summary of model diagnostics with training datasets.

Training model	Spatial sampling scheme	Size of training data set (number of records)	Number of variables in training model	Seed number used	Model DIC	pD
Iberian training model	Subsampling bias	1019	12	380714689	1191.22	48.49
Californian training model	Subsampling bias	359	11	3213	427.62	15.65
Iberian training model	15 km^2 raster grid	522	12	380714689	661.55	23.82
Californian training model	15 km^2 raster grid	118	11	3213	158.65	10.24
Iberian training model	All occurrence dataset	1554	12	380714689	1641.98	65.92
Californian training model	All occurrence dataset	520	11	3213	524.64	20.00

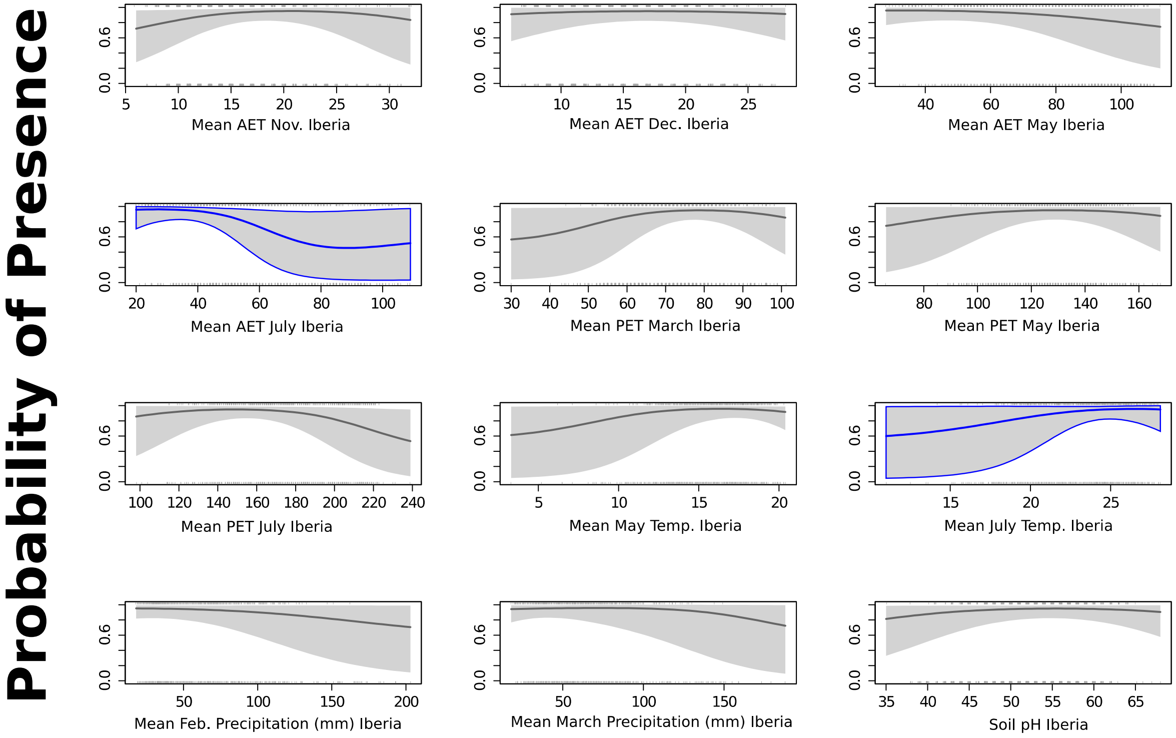


Figure 3.3 Trellis plot of twelve environmental variables in training the Iberian model for the dataset accounting for subsampling bias. The two shortest length scales (those variables that are driving the relationship between niche and occurrence) of the model outlined in blue, 95% Bayesian credible intervals shaded.

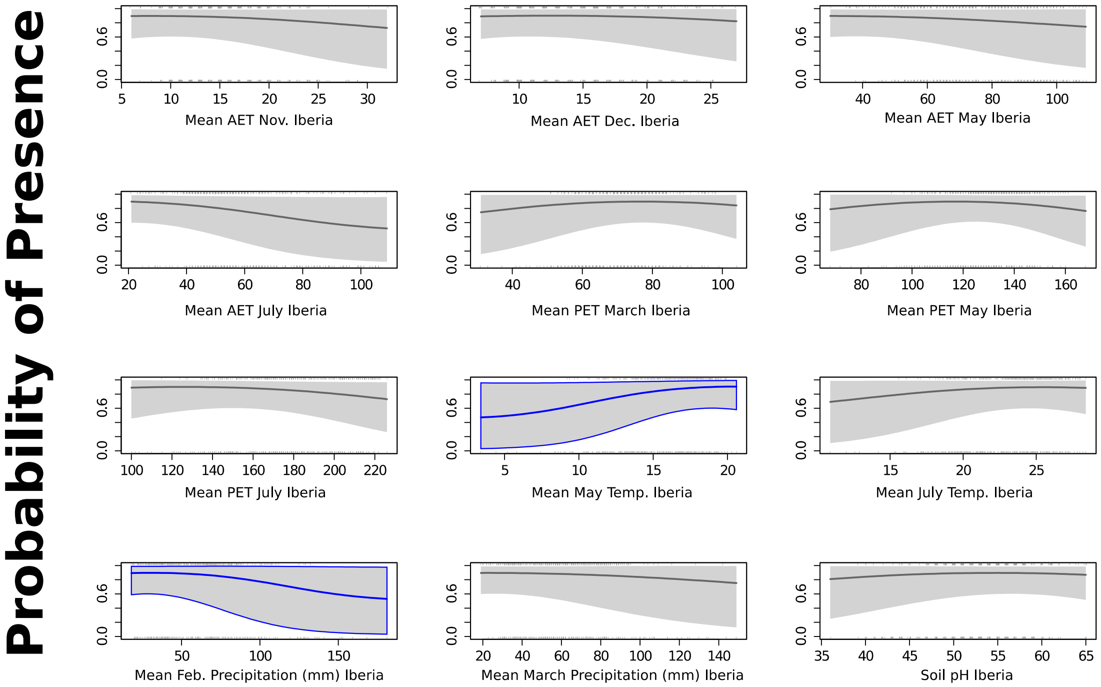


Figure 3.4 Trellis plot of twelve environmental variables in training the Iberian model for the 15 km^2 grid raster dataset. The two shortest length scales of the model outlined in blue, 95% Bayesian credible intervals shaded.

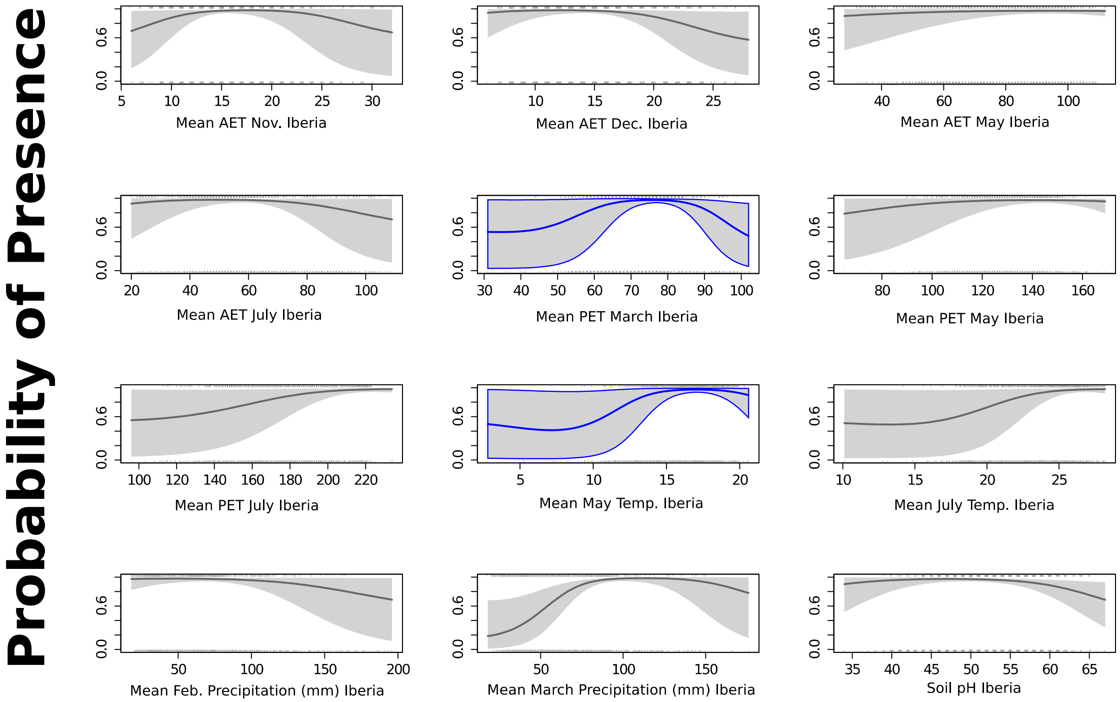


Figure 3.5 Trellis plot of twelve environmental variables in training the Iberian model for the full dataset. The two shortest length scales of the model outlined in blue, 95% Bayesian credible intervals shaded.

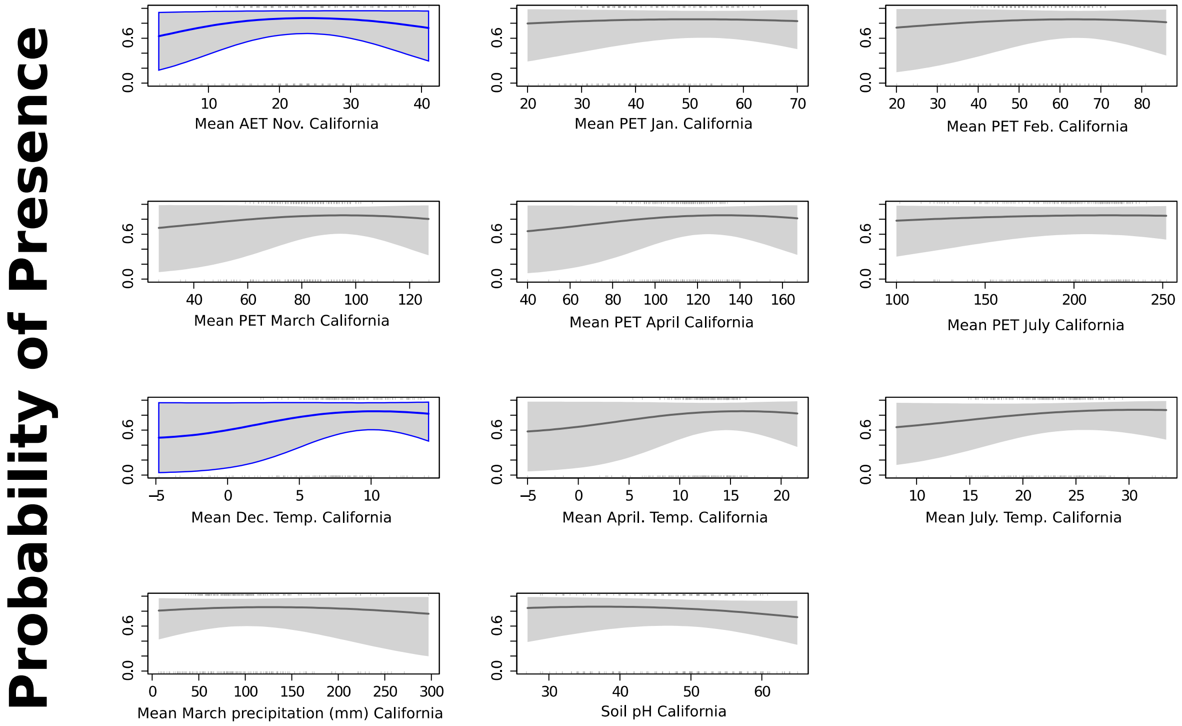


Figure 3.6 Trellis plot of eleven environmental variables in training the Californian model for the dataset accounting for subsampling bias. The two shortest length scales of the model outlined in blue, 95% Bayesian credible intervals shaded.

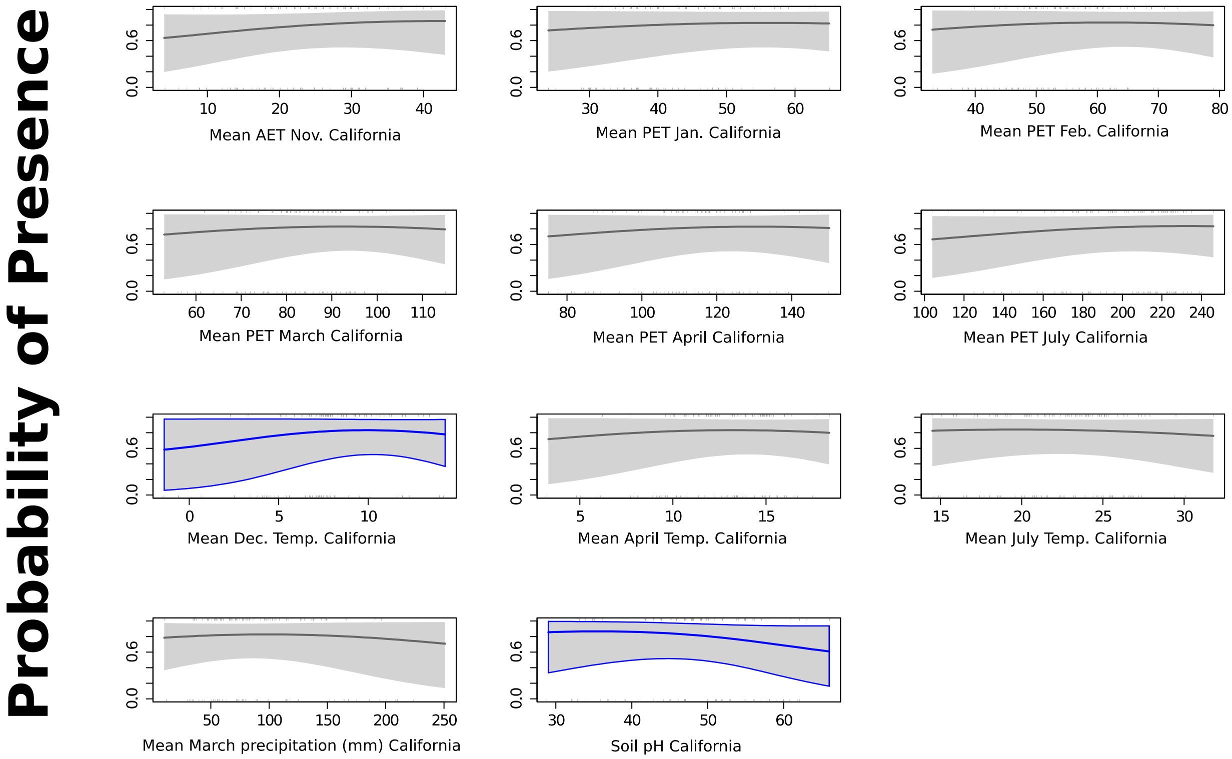


Figure 3.7 Trellis plot of eleven environmental variables in training the Californian model for the 15 km^2 grid raster dataset. The two shortest length scales of the model outlined in blue, 95% Bayesian credible intervals shaded.

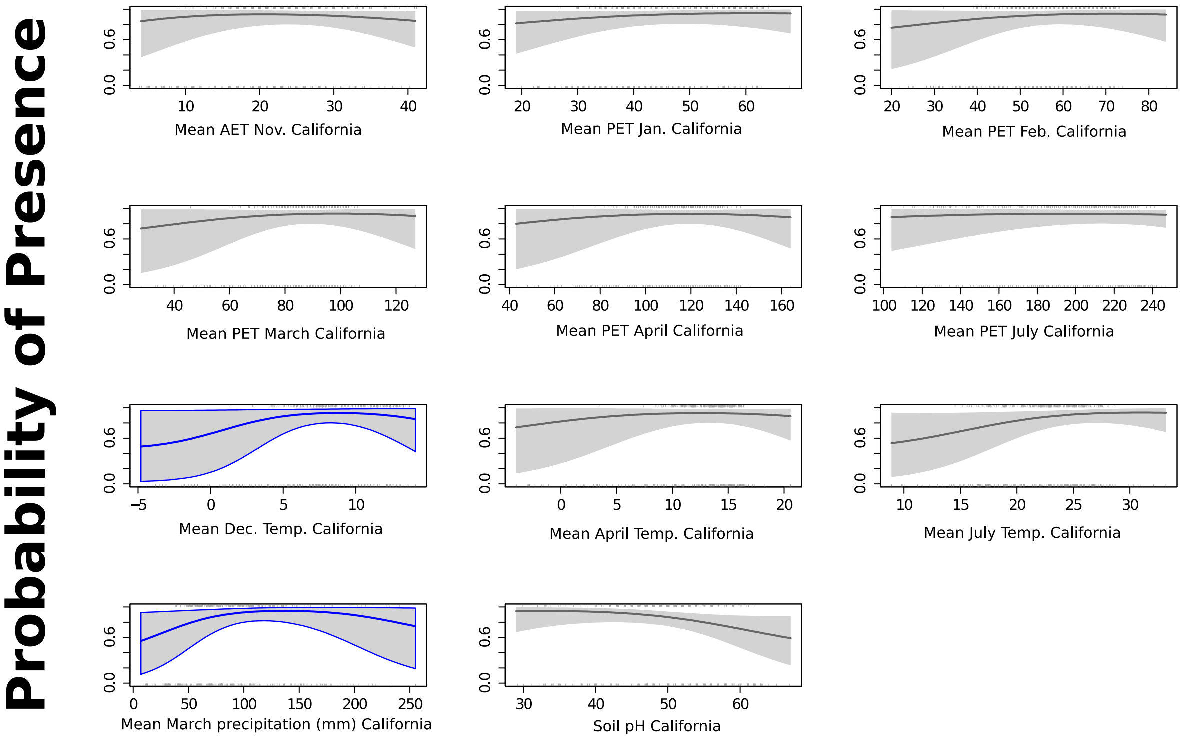


Figure 3.8 Trellis plot of eleven environmental variables in training the Californian model for the full dataset. The two shortest length scales of the model outlined in blue, 95% Bayesian credible intervals shaded.

Table 3.3 Summary of model predictions for each spatial sampling scheme.

Model predictions category	Spatial sampling scheme	#Records predicted	Mean model uncertainty	AUC	AUC STDEV	Specificity	Sensitivity	Threshold	Proportion correctly classified
Iberia to Iberia	Subsampling bias	2532	0.394	0.737	0.010	0.548	0.807	0.678	0.617
Iberia to Iberia	15 km^2 raster grid	1219	0.374	0.645	0.016	0.438	0.760	0.599	0.600
Iberia to Iberia	All occurrence data included	3625	0.332	0.827	0.007	0.720	0.772	0.746	0.671
California to California	Subsampling bias	900	0.354	0.729	0.017	0.492	0.831	0.661	0.679
California to California	15 km^2 raster grid	275	0.514	0.666	0.033	0.500	0.799	0.649	0.520
California to California	All occurrence data included	1215	0.320	0.823	0.012	0.611	0.841	0.726	0.682
Iberia to California	Subsampling bias	1200	0.842	0.619	0.016	0.520	0.650	0.585	0.547
Iberia to California	15 km^2 raster grid	679	0.796	0.624	0.022	0.546	0.663	0.605	0.550
Iberia to California	All occurrence data included	1735	0.817	0.595	0.013	0.376	0.794	0.585	0.521
California to Iberia	Subsampling bias	3376	0.548	0.624	0.010	0.441	0.760	0.601	0.594
California to Iberia	15 km^2 raster grid	1749	0.708	0.593	0.013	0.595	0.531	0.563	0.510
California to Iberia	All occurrence data included	5179	0.576	0.636	0.008	0.575	0.640	0.607	0.531

Table 3.4 Summary of MANOVA results from factors presence vs. absence and Iberia vs. California for three spatial sampling schemes, Pillai's trace indicates variance between groups.

Factor	Spatial sampling scheme	df	Pillai's trace	Approximate multivariate F-value	Probability (>F)
Occurrence	15 km^2 raster grid	1	0.046	28.891	0.0001
Geographic Range	15 km^2 raster grid	1	0.247	197.733	0.0001
Occurrence * Geographic Range	15 km^2 raster grid	1	0.042	26.180	0.0001
Residuals		2416			
Occurrence	Subsampling bias dataset	1	0.060	73.000	0.0001
Geographic Range	Subsampling bias dataset	1	0.596	1693.310	0.0001
Occurrence * Geographic Range	Subsampling bias dataset	1	0.022	26.110	0.0001
Residuals		4588			
Occurrence	All occurrences dataset	1	0.010	196.300	0.0001
Geographic Range	All occurrences dataset	1	0.503	1744.950	0.0001
Occurrence * Geographic Range	All occurrences dataset	1	0.030	54.220	0.0001
Residuals		6910			

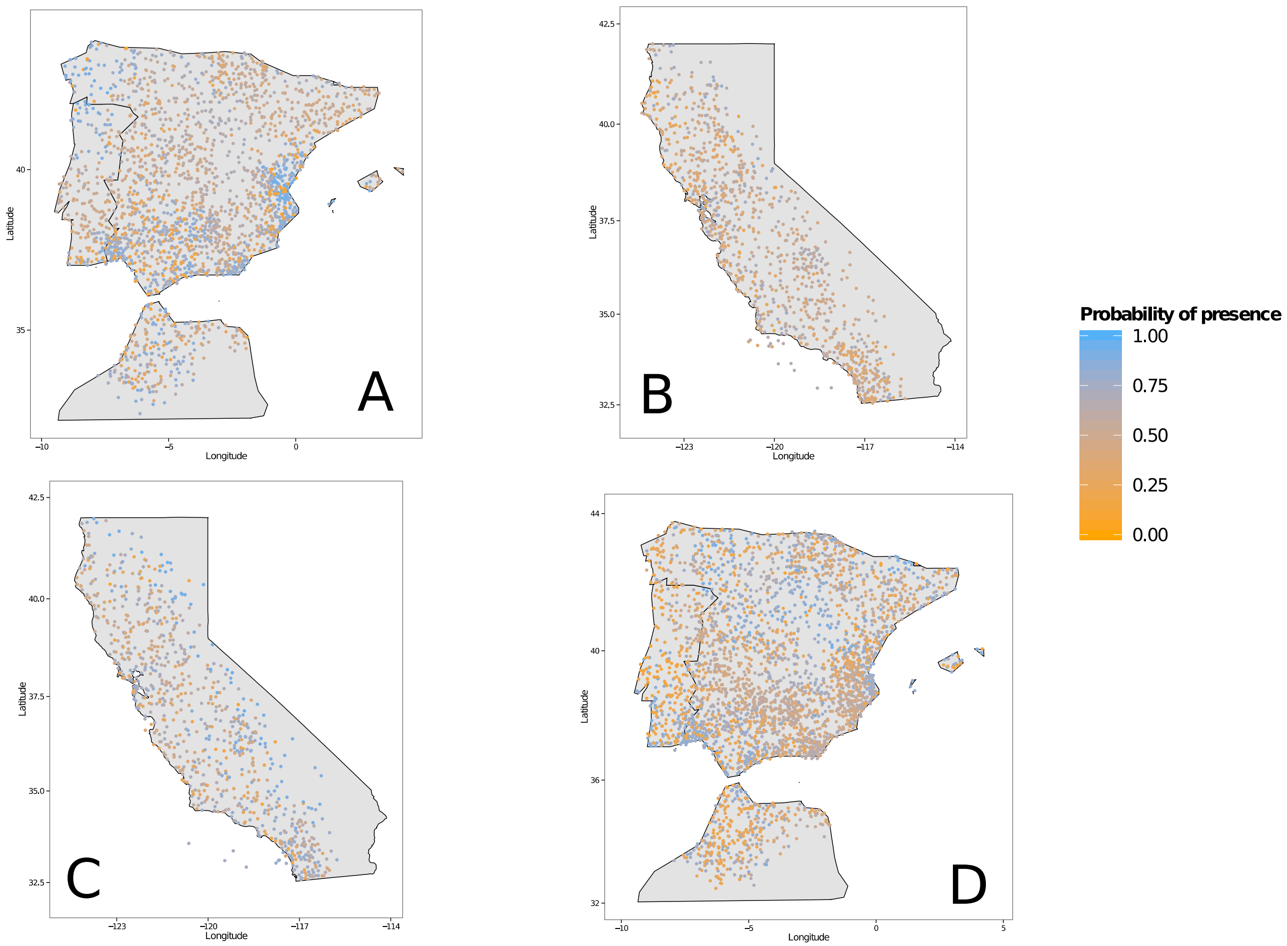


Figure 3.9 Prediction maps for probability of presence with occurrence data subsampled in most populated areas. Iberian model predicted to Iberia (panel A), and then California (panel B). Californian model predicted to California (panel C), and then Iberia (panel D).

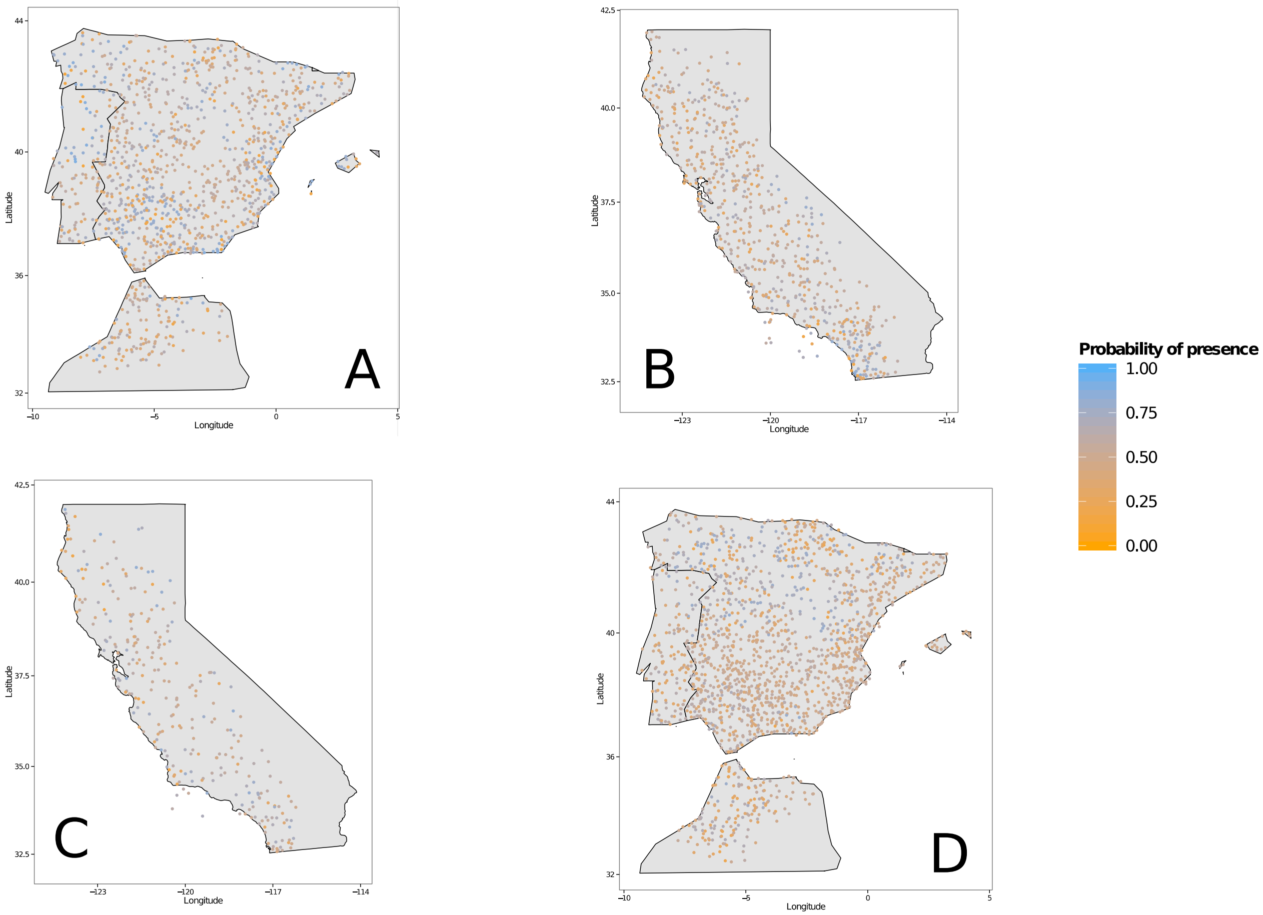


Figure 3.10 Prediction maps for probability of presence with occurrence data sampled using 15 km^2 raster grid. Iberian model predicted to Iberia (panel A), and then California (panel B). Californian model predicted to California (panel C), and then Iberia (panel D).

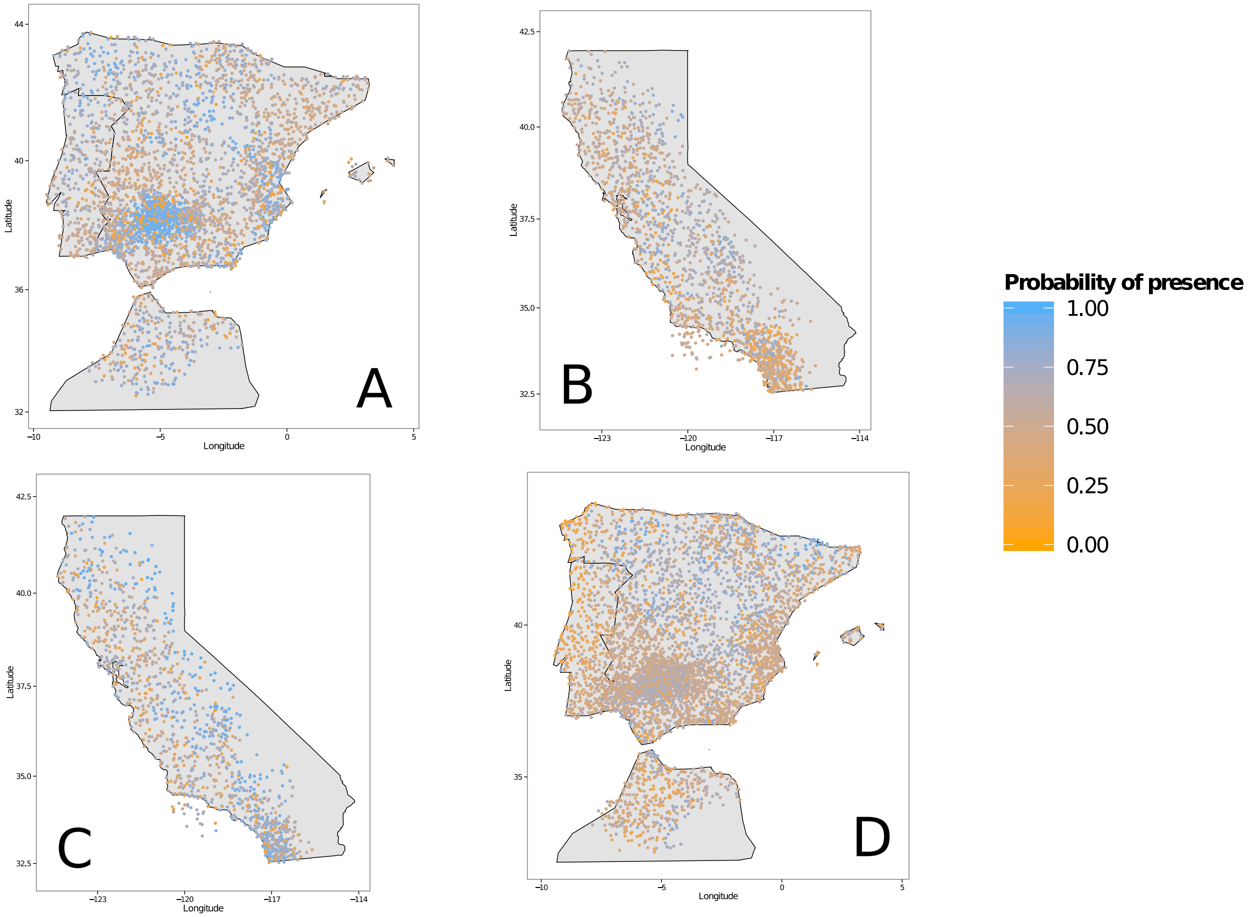


Figure 3.11 Prediction maps for probability of presence with all occurrence data included. Iberian model predicted to Iberia (panel A), and then California (panel B). Californian model predicted to California (panel C), and then Iberia (panel D).

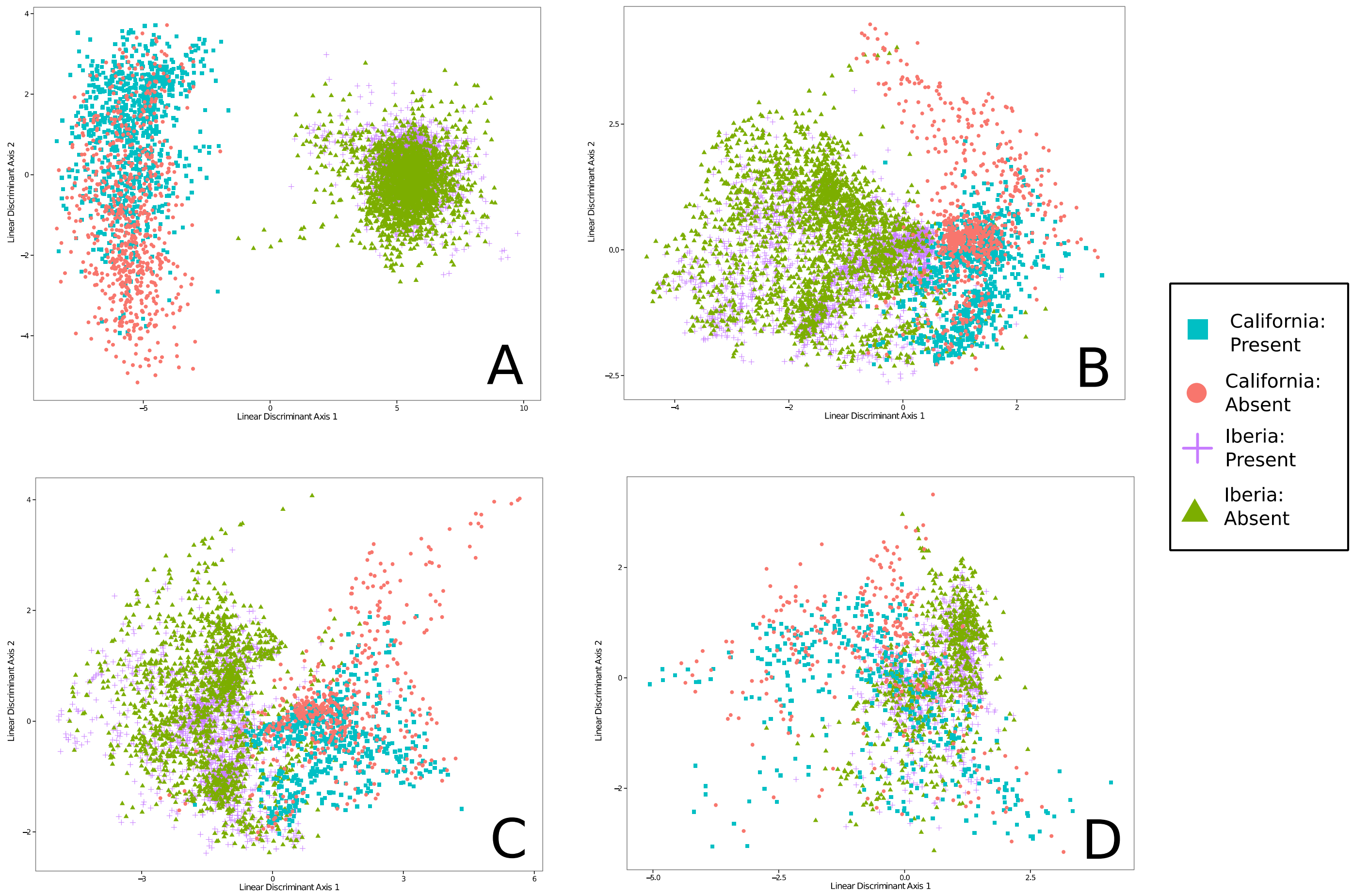


Figure 3.12 Multi-panel plot of linear discriminant functions illustrating niche overlap with respect to the shortest lengthscales in the models. Panel A is the linear discriminant plot of all thirty three environmental variables and occurrence data for comparison with panels B-D: all occurrence data (B), sub-sampling bias data set (C), 15 km^2 raster grid data set (D).

3.5 Discussion

In her general study of ecological origins of Mediterranean grasses in California, Jackson (1985) notes that most annual grasses in their native Mediterranean Basin form small, disjunct communities, not the dense annual stands they do in California. This is potentially indicative of either a climatic niche shift or the release from natural enemies of the introduced Mediterranean Basin grasses to California. *Avena barbata* occurs in all five Mediterranean climate regions, which are characterized by warm winters with substantial rainfall occurring until early spring, and hot, dry summers (Jackson, 1985). In looking at environmental variables of these two climate regions, the Mediterranean Basin (focusing on Iberia), and California, I attempted to characterize niche overlaps and/or niche shifts in *A. barbata* occurrence using a new type of SDM – GRF (Golding, 2013), which incorporates uncertainty in predictively modeling species occurrences.

Models trained in both ranges had three common climatic variables (mean AET Nov., mean PET March, mean temperature July), and one edaphic variable (soil pH) shared between them, and this demonstrates that there is some niche similarity between the ranges. High average July temperatures predicted a high probability of presence in the ranges, but average PET in March and AET in November was noticeably different for the ranges with a higher average PET in March and AET in November in California predicting a high probability of presence compared to Spain. However, the environmental variables with the shortest lengthscales in each of the models were generally different between the ranges, and lengthscale parameters and model complexity were sensitive to the spatial dataset under consideration.

As expected, in every case, each of the models trained in their own range performed better at predicting occurrences in their own range than they did in the opposite range. In general, Iberian trained models across all three spatial sampling scenarios poorly predicted Californian occurrences. Californian trained models did better at predicting presence in Iberia, but there was also a high degree of uncertainty associated with these predictions. I suggest that the high uncertainty associated with each of the trained models attempting to predict the opposite range suggests that the relative importance of certain environmental variables differs between both geographic ranges and points to a climatic niche shift. The MANOVA and linear discriminant function analysis of the most important lengthscales in GRaF models are also suggestive of niche dissimilarity. However, as spatial datasets were pruned down to a fewer number of occurrences, this separation lessened.

SDMs are very much dependent on both accurate occurrence and environmental data, and the spatial and timescales at which environmental data are collected. In general, these data were interpolated at a high resolution for a raster dataset (~1km at the equator), but the number of observation stations does differ between Iberia and California, with California having more observation stations than Iberia (Hijmans et al., 2005).

An additional consideration is that these raster datasets are not necessarily "high-resolution" for an annual plant, which may prefer specific microclimates that cannot be captured with the current scale of environmental data. Recently, it has been suggested that in order to accurately describe the macroclimate of a species, substantial measurement and characterization of the microclimate must occur first to inform large geographic scale SDMs (Gillingham et al., 2012; Harwood et al., 2014). My study was limited to the macroclimate, due to time and the spatial spread of occurrences in both these ranges.

This study also mainly considered climatic variables (with the exception of soil pH) in attempting to describe the niche of *A. barbata*. As important as climatic variables can be in describing the niche of a species, these variables might serve better to inform the range limits of the species as opposed to the distribution within each range. It is also probably important to consider other biotic variables, such as mutualistic or symbiotic relationships with other species (Godsoe et al., 2009). In the case of invasions, the release from ecological competitors ('the enemy-release hypothesis') in the new range may be an additional factor in predicting probability of presence (Colautti et al., 2004; Emery & Ackerly, 2014). I did not make an attempt to include biotic variables in the present study, as doing so in both geographic ranges would have proven a difficult task (not to diminish their potential importance in predicting *A. barbata* occurrences). Another potential factor that may have played a role in determining the distribution of *A. barbata* in the invasive range are fire cycles, where an invasive grass provides ample fuel for fires, which then increase in frequency destroying the native vegetation (reviewed in D'Antonio and Vitousek 1992).

Sampling effort recording presence and absence can also heavily bias which environmental variables are most important in predicting presence (Phillips et al., 2009), and to some extent this difference is observed between the three spatial sampling schemes used here. By selectively altering the number of occurrences by targeting only one individual in densely sampled areas based on a reasonable raster grid size or pruning the data across the total range, the relative importance of environmental variables changes in the GRF model. The one exception to this observation is a warm average mean temperature in December in California, suggesting that warm winters in California are associated with a high probability of presence for *A. barbata*. By pruning the dataset, the average uncertainty in each model shifts dramatically as well, with the sparser the dataset (in terms of number of occurrences) the greater the uncertainty in model predicitions.

When all evidence is considered, I think it is possible that *A. barbata* occupies a different climatic niche in California than it does in Iberia, and may preferentially occupies different environments than those it occupies in Iberia. First, subsets of different environmental variables were better at optimally predicting a high probability of presence in models trained in both ranges. This on its own may not necessarily indicate a shift because there could be substantial seasonal offsets between the two ranges, for example the rainiest month in Iberia may be November, and the rainiest month in California may be Dec. However, two of the variables shared between models trained in both ranges had different means – these were mean March PET and November AET. This is the first time (to my knowledge) GRF has been used to investigate niche shifts. I would tend to advance that high uncertainty associated with models predicting to the opposite range means that the subset variables that predicts presence in Iberia (for example) does not reliably predict presence in California. No matter the spatial category, there was always a fair amount of uncertainty with models predicting occurrence in the opposite range. Finally, both the two-way MANOVA and LDA seem to show that when the most important environmental variables are considered, there is little overlap between them when considered as multivariate wholes.

I note that this potential climatic niche shift is not indicative of evolutionary change or local adaptation in either range. In the present study, it cannot be ruled out that colonizing *A. barbata* individuals may have had substantial physiological and also phenotypic plasticity available to changing stresses like moisture gradients, which may have then lead to altering a trait such as flowering time (Sherrard et al., 2009). If evolutionary change were occurring in California, I would hypothesize it is because of other factors such as recombination (Chapter 4) possibly between the three maternal introductions (Chapter 2) from different parts of the native range.

3.6 References

Ackerly DD (2009) Evolution, origin and age of lineages in the Californian and Mediterranean floras. *Journal of Biogeography*, **36**, 1221–1233.

Allard RW, Babbel GR, Clegg MT, Kahler AL (1972) Evidence for coadaptation in Avena barbata. *PNAS*, **69**, 3043–3048.

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Blumler MA (2000) Spatial analysis to settle an unresolved question in genetics, with both theoretical and applied implications. *Research in Contemporary and Applied Geography: A Discussion Series*, **24**, 1–42.

Chamberlain S, Boettiger C, Ram K, Barve V, McGlinn D (2014) Package "rgbif" Interface to the Global Biodiversity Information Facility API. R package version 0.5.0. 1–67.

Clegg MT, Allard RW (1972) Patterns of genetic differentiation in the slender wild oat species *Avena barbata*. *PNAS*, **69,** 1820–1824.

Colautti RI, Ricciardi A, Grigorovich IA, MacIsaac HJ (2004) Is invasion success explained by the enemy release hypothesis? *Ecology Letters*, **7**, 721–733.

D'Antonio CM, Vitousek PM (1992) Biological invasions by exotic grasses, the grass/fire cycle, and global change. *Annual Review of Ecology and Systematics*, **23**, 63–87.

Dubuis A, Giovanettina S, Pellissier L, Pottier J, Vittoz P, Guisan A (2012) Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. *Journal of Vegetation Science*, **24**, 593–606.

Emery NC, Ackerly DD (2014) Ecological release exposes genetically based niche variation. *Ecology Letters*, **17**, 1149–1157.

Fitzpatrick MC, Weltzin JF, Sanders NJ, Dunn RR (2007) The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Global Ecology and Biogeography*, **16**, 24-33.

Garcia P, Vences F, Vega M (1989) Allelic and Genotypic Composition of Ancestral Spanish and Colonial Californian Gene Pools of Avena barbata: Evolutionary Implications. *Genetics*, **122**, 687–694.

Gillingham PK, Huntley B, Kunin WE, & Thomas CD (2012) The effect of spatial resolution on projected responses to climate warming. *Diversity and Distributions*, **18**, 990–1000.

Godsoe W, Strand E, Smith CI, Yoder JB, Esque TC, Pellmyr O (2009) Divergence in an obligate mutualism is not explained by divergent climatic factors. *New Phytologist*, **183**, 589–599.

Golding N (2013) Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications. 1–247. University of Oxford, UK. Doctoral Dissertation.

Guisan A, Petitpierre B, Broennimann O, Daehler C, & Kueffer C (2014) Unifying niche shift studies: insights from biological invasions. *Trends in Ecology & Evolution*, **29**, 260–269.

Hamrick JL, Allard RW (1975) Correlations between quantitative characters and enzyme genotypes in Avena barbata. *Evolution*, **29**, 438–442.

Harwood TD, Mokany K, Paini DR (2014) Microclimate is integral to the modeling of plant responses to macroclimate. *Proceedings of the National Academy of Sciences*, **111**, E1164–E1165.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hutchinson G (1959) Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist*, **93**, 145–159.

ISRIC – World Soil Information, 2013. SoilGrids: an automated system for global soil mapping. Available for download at http://soilgrids1km.isric.org.

Jackson LE (1985) Ecological origins of California's Mediterranean grasses. *Journal of Biogeography*, 349–361.

Levin DA (2009) Flowering-time plasticity facilitates niche shifts in adjacent populations. *New Phytologist*, **183**, 661–666.

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Minnich R (2008) *California's Fading Wildflowers.* University of California Press, Berkeley and Los Angeles.

Pearman PB, Randin CF, Broennimann O, Vittoz P, Knaap WOVD, Engler R, Lay GL, Zimmermann NE, Guisan A (2008) Prediction of plant species distributions across six millennia. *Ecology Letters*, **11**, 357–369.

Petitpierre B, Kueffer C, Broennimann O, Randin C, Daehler C, Guisan A (2012) Climatic niche shifts are rare among terrestrial plant invaders. *Science*, **335**, 1344–1348.

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

Pinero D (1982) *Correlations between enzyme phenotypes and physical environment in California populations of Avena barbata and Avena fatua.* University of California, Davis. PhD Dissertation.

Sherrard ME, Maherali H, Latta RG (2009) Water stress alters the genetic architecture of functional traits associated with drought adaptation in Avena barbata. *Evolution*, **63**, 702–715.

Václavík T, Meentemeyer RK (2011) Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. *Diversity and Distributions*, **18**, 73–83.

Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, & Tiffin P (2014) Genomic signature of adaptation to climate in Medicago truncatula. *Genetics*, **196**, 1263–1275.

Chapter 4 Contemporary evolution over 40 generations in an invasive annual grass

4.1 Abstract

Observing a unidirectional shift in heritable characters over time across a broad geographic range strongly suggests adaptive evolutionary change has occurred. The opportunity to examine adaptive contemporary evolutionary change over a large geographic region is decidedly rare, as studies quantifying this type of change typically have historical datasets from only a few sites. *Avena barbata* populations in California have heritable morphological trait data from over a hundred populations (statewide) in two traits (lemma color and leaf sheath hairiness) dating back to the 1970s. A previous common garden field experiment predicts that there should be an increase of the frequency of light lemma and leaf sheath pubescence. In 2010, these sites were re-surveyed, and the frequency of each of the heritable traits was calculated. Over 40 years, the frequency of both light lemma color and leaf sheath pubescence increased at all but the most southern populations in California. Each of these traits appears to have two underlying QTL on different linkage groups, and both appear to be epistatic in nature. A simulation strongly suggested that adaptive evolution, rather than drift and recombination, were responsible for the observed changes. This study demonstrates ongoing adaptive evolutionary change in California in two heritable characters over 40 years.

4.2 Introduction

A unidirectional shift in the frequency of heritable polymorphic traits over a short time period, in many populations, and throughout a large geographic range strongly suggests natural selection. Reviews of studies showing changes in the frequency of heritable characters through time demonstrate evolution over short periods of time (Reznick and Ghalambor 2001; Carroll et al. 2007). Nonetheless, the opportunity to observe contemporary evolutionary change by natural selection over several decades and in the wild is still relatively rare. It requires obtaining long-term studies or datasets of morphs with accompanying genetic data. Most of these evolutionary studies are well-known, and use a combination of breeding designs and/or field observations to first disentangle the genetic basis of trait and then survey phenotypes in the field and include: different favored morphs of *Drosophila pseudobscura* in between different regions in the southwestern United States (Dobzhansky and Levene 1948), the ongoing study of an insular population of Soay sheep, industrial melanism in peppermoth (*Biston betularia*) in Britain (Kettlewell 1956), Peter and Rosemary Grant's ongoing studies with Darwin's finches (Grant and Grant 1993; 2002). Very few (if any) studies have been able to isolate the signal of natural selection in polymorphic heritable characters throughout a large geographic area over several decades.

In the era of next-generation sequencing (NGS) analysis, obtaining massive amounts of molecular genetic data at increasingly low cost would seem to trump morphological markers with a known heritable basis (Davey et al. 2011). Yet, heritable morphological markers are still advantageous in that they allow for rapid and convenient screening of genotypes in large, geographic, field surveys. Populations of the annual grass, *Avena barbata* Pott ex Link (common name: slender wild oat) in California present a unique opportunity to study the dynamics of contemporary evolution in heritable traits for two reasons. First, because there exists a large amount of background information on past genotype frequencies based on morphological heritable markers over a large geographic area (statewide); and second, previous experimental common garden work has predicted the direction of evolutionary change by selection.

R.W. Allard and his working group originally pioneered Californian populations of *A. barbata* as a study system of local adaptation in the 1970s - 1980s. It was among some of the first species to be genetically characterized in the wild, by molecular markers nearly 40 years ago. Five allozyme loci were observed almost exclusively in two distinct combinations (out of a possible thirty-two) throughout the state of California, (Allard et al. 1972; Clegg and Allard 1972; Miller 1977). Each of the two combinations appeared to exist in divergent environments along a moisture gradient; and, assuming equilibrium between migration and selection, it was concluded that these two multi-locus genotype combinations represented two locally adapted ecotypes – termed 'mesic' and 'xeric' for the genotypes occurring moist and arid environments, respectively (Allard et al. 1972). Two morphological true-breeding traits – leaf sheath pubescence and lemma color – were associated with these two monomorphic allozyme combinations. Light lemma color and leaf sheath pubescence strongly correlates with the mesic ecotype, while dark lemma color and lack of pubescence (or glabrous leaf sheaths) strongly correlates with the xeric ecotype (Clegg and Allard 1972).

Past work from reciprocal transplant common garden multi-year field experiments in California (2003-2004 and 2006-2007), demonstrated that mesic genotypes were more fit than xeric genotypes in both arid and moist environments(Latta 2009), contradictory to the original conclusion of local adaptation on the geographic scale suggested by Allard et al. (1972). Because mesic genotypes are consistently more fit than xeric genotypes (Latta 2009), I predict that traits associated with the mesic genotype ought to have increased in frequency over the past 40 generations in California. A unidirectional shift in the frequency of these characters would provide strong evidence that selection has influenced the distribution of these traits in the state of California. Note, that I did not consider the adaptive significance of each of the morphological markers, nor do I have any reason to believe that lemma and leaf sheath traits are directly under selection. The goal was to use the morphological markers as a way of tracking evolutionary change of the mesic and xeric genotypes.

While the classic ecotypic trait combinations are those most likely to found throughout California in the present because of the high rate of self-fertilization and lack of pollen migration in *A. barbata*, recombination may also influence observations in the present time period. Additional work with recombinant inbred lines (RILs) created via crossing the two parental ecotypes (Latta et al. 2004), demonstrated transgressive segregation in QTL associated with fitness for late-generation recombinants in a novel (greenhouse environment) (Johansen-Morris and Latta 2006). This work suggests that a novel recombinant may be spreading throughout California.

A geographic resurvey of *A. barbata* was conducted in 2010 and assessed the frequency of lemma color and leaf sheath pubescence in populations that were previously screened in the 1970s, or spanning approximately 35-40 generations. This is a rare opportunity, as very few studies have this amount of historical genetic data from this large an area to compare to a present day survey, and further a clear prediction regarding the direction of selection. I used information about population sites from three doctoral theses of R.W. Allard's students (see methods), established the genetic basis of each trait, and performed simulations to characterize the distribution of trait values under drift and occasional recombination, as opposed to selection.

4.3 Methods:

4.3.1 Geographic resurvey

Over the course of three weeks in May 2010 a geographic resurvey of a subset of previously studied Californian slender wild oat populations was carried out – for a total of 105 collection sites. Geographic coordinates and site descriptions from three doctoral dissertations from Allard's students (Clegg 1972; Miller 1977; Hutchinson 1982) were used to locate *Avena barbata* collection sites surveyed in the 1970s. Each of these dissertations was chosen because each contained sites characteristic of the full range of growing conditions throughout the state of California. Clegg (1972) and Hutchinson (1982) sites were surveyed intensively in the past and as such collections of between 50-100 individuals were made. Miller (1977) sites were more geographically widespread than Clegg and Hutchinson sites, but fewer individuals were collected in the past. Between 10-20 individuals were collected during our re-survey of Miller's (1977) sites. Original site descriptions from this era provided map coordinates that were resolved to the nearest minute of arc, i.e. approximately 1 km. For greater than 90% of the sites, we were able to get within at least four kilometers (kms) radius (and in most cases 2 kms) of the original site as described in the 1970s. For each individual, mature seeds were targeted for collection for this study and other studies using molecular markers and measurement of phenotypic traits in the greenhouse. In general, the later along in flowering, the more seeds were collected in the field. Each site was widely surveyed by walking randomly between individuals and attempting to obtain a number of individuals in early and late stages of flowering. At each site, the trait pubescence of each individual was quickly established by visual examination of the leaf sheath. The aggregate count and overall frequency of pubescent (vs. glabrous individuals) was recorded, e.g. 12/18 individuals or 0.6667 at the Willits collection site.

Dark lemma and leaf sheath pubescence counts per site were not recorded for Clegg (1972), Miller (1977), and Hutchinson (1982) dissertations, so I inferred them using the smallest whole number consistent with the reported frequency.

Lemma color of the recent Californian collection was assessed after returning from the field for 2272 individuals collected from the 2010 re-survey, which is roughly comparable to our estimate of 2264 individuals surveyed in 1970s. Lemma color is often described as having a variety of "distinct" colors by domesticated oat breeders (red, black, gray, yellow, and white) (Coffman 1964). However most past publications stemming from Allard's laboratory categorize (by eye) the lemma color trait of *A. barbata* as binary: gray/light vs. black/dark. Admittedly, using a quantitative assay may have been more objective, but in order to keep methods consistent with Allard's group, I also judged lemma color as a binary trait after returning from the field, by eye. For our Californian collection, I used two extreme examples of lemma color from the field as our base for categorizing an individual's lemma colour: one individual from Oak Bottom as our standard for light lemma, and an individual from Malibu as the standard for dark lemma (Fig. 4.1).

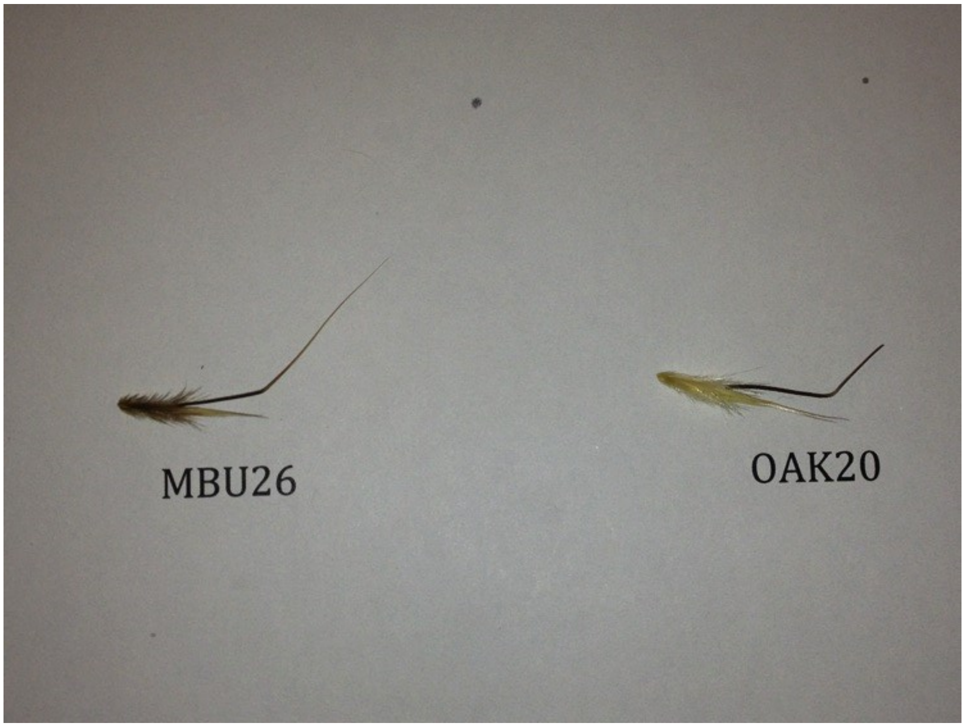


Figure 4.1 The two standards used to assign dark or light lemma color to Californian individuals.

4.3.2 QTL mapping of traits in the RILs

Recombinant inbred lines (RILs) were bred (Gardner and Latta 2006) from a cross between classic mesic and xeric parent genotypes, with propagation in later generations by single seed descent (minimizing selection) for past experimental work on the genetic basis of fitness (Latta et al. 2004). I used these to map QTL underlying the trait of lemma color. In the RILs, I used four sets of each of the mesic and xeric ecotypes, including the original xeric mother (X189) of the RILs as a baseline for color assessment of 180 RILs. RILs from individuals within the same line from the F6, F7, and F8 generations were compared to ensure that lemma color was a true-breeding character. As a general note, I also observed that lemma color did not vary among seeds within an individual RIL genotype, nor among seeds from an individual collected from the field.

502 genotype-by-sequence (GBS) markers were available (Latta, unpub.) and were generated by an established protocol (Poland et al. 2012) along with 129 AFLP markers (Latta et al. 2010) for *A. barbata* RILs. These 631 makers were used concurrently in establishing the heritable basis of the traits of lemma color and pubescence in 94 of the F6 RILs. These data are a smaller subset of the 180 RILs with complete genetic map (AFLP and GBS) data and phenotypic data. The ~ 90 remaining RILs have only AFLP data available for them, but are currently being screened with GBS markers. At present, this genetic map is provisional and linkage group numbers differ from the ones originally presented in Gardner & Latta (2006).

An observed 3:1 ratio of dark lemma:light lemma and pubescent leaf sheaths:glabrous leaf sheaths in the F6 generation of the RIL cross, and suggests an epistatic mode of inheritance in both traits. With heterozygotes being produced from a cross a 3:1 ratio would imply that one gene underlies each of the traits. However, with F6 RILs (which are homozygotes due to repeated selfing), an observed 3:1 ratio of two phenotypes (e.g. dark vs. light) would imply several loci, i.e. an epistastic or a threshold trait model act to produce each of the traits.

I performed composite interval mapping (CIM) and considered a single-QTL model as a preliminary ‘first pass’ in searching for a single locus associated with lemma color. Because of suspected underlying epistasis, a two-dimensional scan (Broman et al. 2003; Arends et al. 2010) across all markers in the genomic map was performed to explore any interactions between pairs of loci. Permutation tests (n =1000) as described by Churchill and Doerge (1994) were used to ascertain statistical significance of a QTL location for both the one and the two dimensional genome scans. I then considered a multiple-QTL model, performing multiple-interval-mapping (MIM) and searched for loci associated with lemma color and leaf sheath pubescence. In exploring the multiple QTL model, I dropped one locus at a time and compared (using maximum likelihood) the reduced model to the full model to assess the percent of phenotypic variation explained by each locus on its own, and the interaction between them. However, the p-values used in assessing significance for both traits in the MIM QTL model are based on pointwise tests on markers of interest, whereas the two-dimensional scan tests all pairs of markers across the genome. All QTL mapping analyses were carried out using the R package 'rqtl' v. 1.32-10 (Broman et al. 2003; Arends et al. 2010), which runs in the R core environment.

4.3.3 Statistical analysis of phenotypic data

As a coarse initial analysis, I examined whether both single trait frequency observations differed significantly from each other in both time periods using two-sample proportion χ^2 – tests on a statewide level, i.e. treating samples from sites as independent on a statewide level. Using the same test, I also determined whether multi-trait combinations (under the assumption of linkage equilibrium –see further section in methods) changed significantly over time.

Data of past studies of the geographic distribution of the ecotypes from Allard's lab and our own cpDNA work (see Chapter 2) suggested the existence of a North-South latitudinal cline in California. Additionally, leaf sheath pubescence data in 2010 revealed a North-South cline (Latta unpub.), with populations fixed for glabrous leaf sheaths and dark lemma color (xeric type), and chloroplast xeric haplotypes (see Chapter 2) being mostly confined to southern California. I fit generalized linear mixed models to the data to infer whether the latitudinal cline in the relative frequency of light lemma color had shifted between time periods of survey collection. All models were fit with R-package 'lme4' v. 1.1-7 (Bates et al. 2013). Individuals within a site were treated as non-independent observations since samples from the same site are not independently drawn from the entire state (i.e. not true replicates), and the intercept (site frequency of lemma color) varied freely between sites as a random effect. Latitude was then added to the models as a fixed effect (model statement with fixed effects below):

Frequency of Light Lemma ~ Latitude + Time Period + (1|Site)

I first considered datasets collected at the two different time periods (1970s vs 2010) separately, and then merged both datasets and added time period of survey as a fixed effect. As repeat measures of the same sites were made, time period is a within site variable. Past observations suggest that genotypes with mesic traits may represent a second introduction to California (Ch.2), and genotypes with mesic traits were rarely observed outside the San Francisco Bay Area (latitude 36-38 N) in California (Allard et al. 1972; Clegg 1972; Miller 1977).

4.3.4 Simulation of change under neutral expectations

Observed changes in trait frequencies over a short period of time, could also be the combined result of occasional recombination (as previously observed in *A. barbata*) and random genetic drift (the efficiency of which is increased in small populations). Consequently, I simulated the expected distribution of each character (starting with the observed frequency in the 1970s) under drift alone over 40 generations. Simulated distributions expected under drift starting from frequencies observed in 1970s were then compared to observations made from the re-survey in 2010.

To more accurately estimate the approximate starting frequency of each trait in the simulations, collection sites were grouped together according to one-degree bands of latitude. I chose this option of spatially grouping sites by latitude, as many of the past collections at certain sites (especially Miller's sites) were small, and it would have been difficult to accurately estimate the frequency of each trait. Starting frequencies of each trait were averaged within bands of latitude.

For each simulation, I set the outcrossing rate at 0.02, and kept the population size constant at 1000 individuals. I modeled the inheritance of each trait as if it were controlled by two epistatic loci (see results in QTL section for how each trait is inherited). Migration was not considered in our simulations, as *A. barbata* is highly selfing, and as such pollen flow between populations is limited, and while seed migration between populations may occur, I expect it to be relatively rare between distant populations over 40 generations. Additionally, I did not consider mutation, as mutations occurring within the 40 generations at the loci underlying the traits of interest would be rare. Thus, the only potential new source of genetic variation moving forward in time in the simulations was the shuffling of genotypes via occasional 2% outcrossing. I used Fisher's method (Fisher 1932) to combine calculated probabilities across independent bands of latitude, and evaluated the null hypothesis that average observed 2010 trait frequencies do not differ from the simulated distributions expected under drift and occasional recombination. All simulations were performed in the R v. 3.1.0 core environment, and the script to run a basic simulation is provided in Appendix 3.

4.3.5 Assessment of potential recombinants in California

As pubescence was only recorded as aggregated counts for a site, I could not identify multi-trait morphotypes directly. However, the frequencies of two phenotypic traits from both 2010 and the 1970s, lemma color and leaf sheath pubescence, allowed us to estimate morphotype combination frequencies of all sites at both time periods (Supp Table 1). Under the assumption of linkage equilibrium, estimates of morphotype frequencies were obtained by multiplying the frequency of pubescence or glabrousness recorded at a site by the frequency of observed dark or light lemma individuals. In general, most of our sites were fixed for at least one trait: in 2010 there were 82 sites, and in the 1970s there were 79 sites where at least one trait was fixed in the population, so the estimate is likely to be accurate. For example, if at a site there was a 0.5 frequency for dark lemma and a frequency of 1.0 for site pubescence, I multiplied 0.5 * 1.0 = 0.5, i.e. I estimated that half of the individuals at that site had dark lemmas and were also pubescent. However, the assumption of linkage equilibrium is likely not conservative enough for polymorphic populations for both traits, as I expected dark lemma color to be disproportionately associated with glabrous leaf sheath (the xeric morph). For example, Bell's Station in the 1970s was 0.15 glabrous and 0.74 individuals had a dark lemma (Supp table 1b). At most, 0.15 individuals were both glabrous and dark, (the xeric morphotype) with remaining 0.59 dark individuals having pubesecent leaf sheaths. Under linkage equilibrium (multiplying glabrous and dark frequencies) this would yield 0.11 dark and glabrous, and

0.63 dark and pubescent individuals. The true genotype frequency lies somewhere between these two estimates, but because the majority of sites had a least one trait fixed or not very much difference between the two extremes, I assumed linkage equilibrium. The direction of change was then estimated in the multi-trait frequencies, to determine whether individuals with classic trait combinations increased or decreased relative to recombinant morphotype combinations.

4.4 Results

4.4.1 QTL mapping of heritable traits

The initial one-dimensional scan detected two QTL associated with lemma color on linkage groups (LG) 2 and 10 (Fig. 4.2). For leaf sheath pubescence, the one-dimensional scan also detected two QTL, LG 1 and 8 (Fig. 4.3).

For QTL associated with lemma color, the two-dimensional genome scan across all pairs of markers for 94 individuals with complete phenotypic, GBS, and AFLP data revealed that the two inferred QTL on LG 2 and 10 acted additively (LOD significance threshold for additive model = 15.3, n = 1000 permutation tests). No experiment-wise significant LOD scores for epistatic interactions were detected for thresholds established with permutation tests (Fig. 4.4). However, when the pair of QTL at LG 2 and 10 was examined exclusively, I found that the full model explained 55% of variance in the phenotype (Table 4.1). The comparison of each locus and the interaction between the two loci to the full model revealed significant contribution to phenotypic variance for all model terms (Table 4.1) including the interaction term, which explains 4% of the observed phenotypic variance in lemma color. The significant interaction term lends support to an epistatic model of inheritance for the trait of lemma color. If an individual has all mesic alleles at markers close to both QTL, it will likely have the light lemma phenotype, but if it has even one xeric allele at either QTL it is more likely to have the dark lemma phenotype. Or more simply, aabb gives light lemma, the other three double homozygotes give dark lemma. One other interesting finding of note is that the QTL on LG 10 for lemma color occurs very close to a QTL significantly associated with mean fitness in the RILs in the field (Latta et al. 2010).

For pubescence, the two-dimensional genome scan across all pairs of marker locations for 86 individuals with phenotypic data (8 individuals were dropped due to missing data) also revealed two QTL that acted additively on LG 1 and LG 8 (LOD significance threshold for additive model = 18.4, n = 1000 permutation tests) in influencing leaf sheath pubescence (Fig. 4.5). As with lemma color, no significant LOD scores for epistatic interactions were found when all pairs of marker locations in the genome were considered. With the QTL model constructed with MIM - the full model explained 66% of variance in the phenotype of leaf sheath pubescence (Table 4.1). The comparison of each locus and the interaction between the two loci to the full model revealed significant additive contributions for each marker close to the QTL to explaining phenotypic variance in pubescence, but despite the observed 3:1 ratio of pubescence to glabrousness in the F6 RILs the interaction term was not significant (Table 4.1).

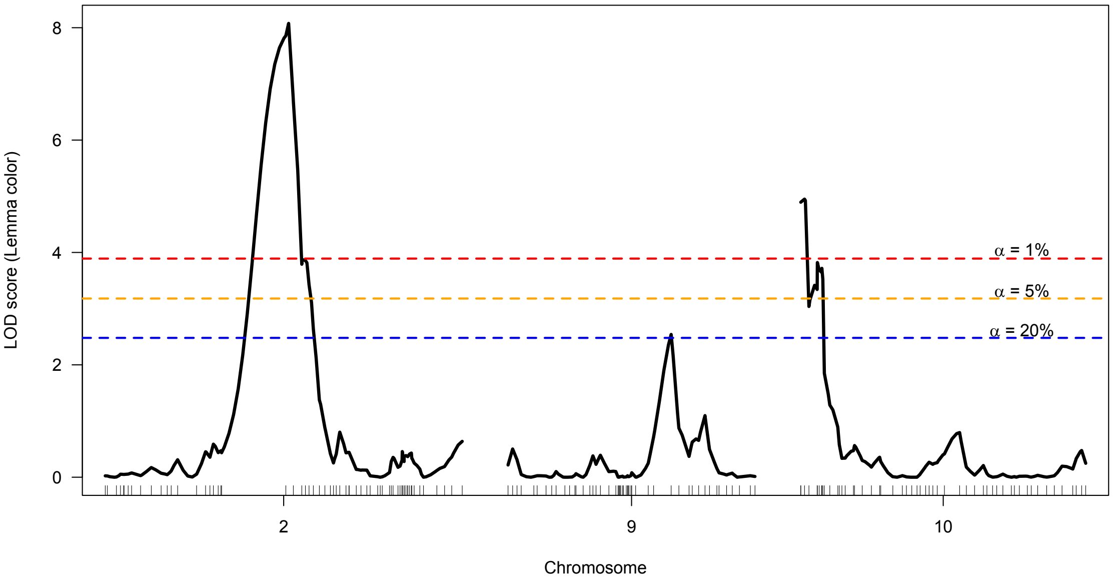


Figure 4.2 LOD plot of linkage group(s) associated with lemma color (LG 2 and 10) in the RILs carried out using CIM. Hatched blue, orange, and red lines are cut-offs for α-values = 0.2 (LOD = 2.48), 0.05 (LOD = 3.18), and 0.01 (LOD=3.89) thresholds based on 1000 permutation tests.

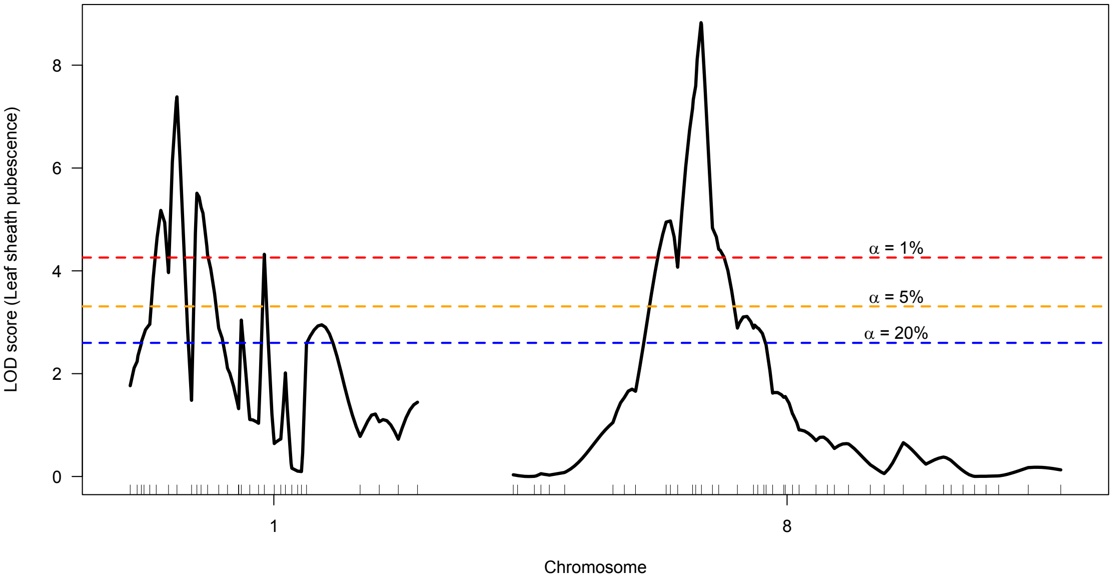


Figure 4.3 LOD plot of linkage group(s) associated with leaf sheath pubescence (LG 1 and 8) in the RILs carried out using CIM. Hatched blue, orange, and red lines are cut-offs for α-values = 0.2 (LOD = 2.60), 0.05 (LOD = 3.31), and 0.01 (LOD = 4.26) thresholds based on 1000 permutation tests.

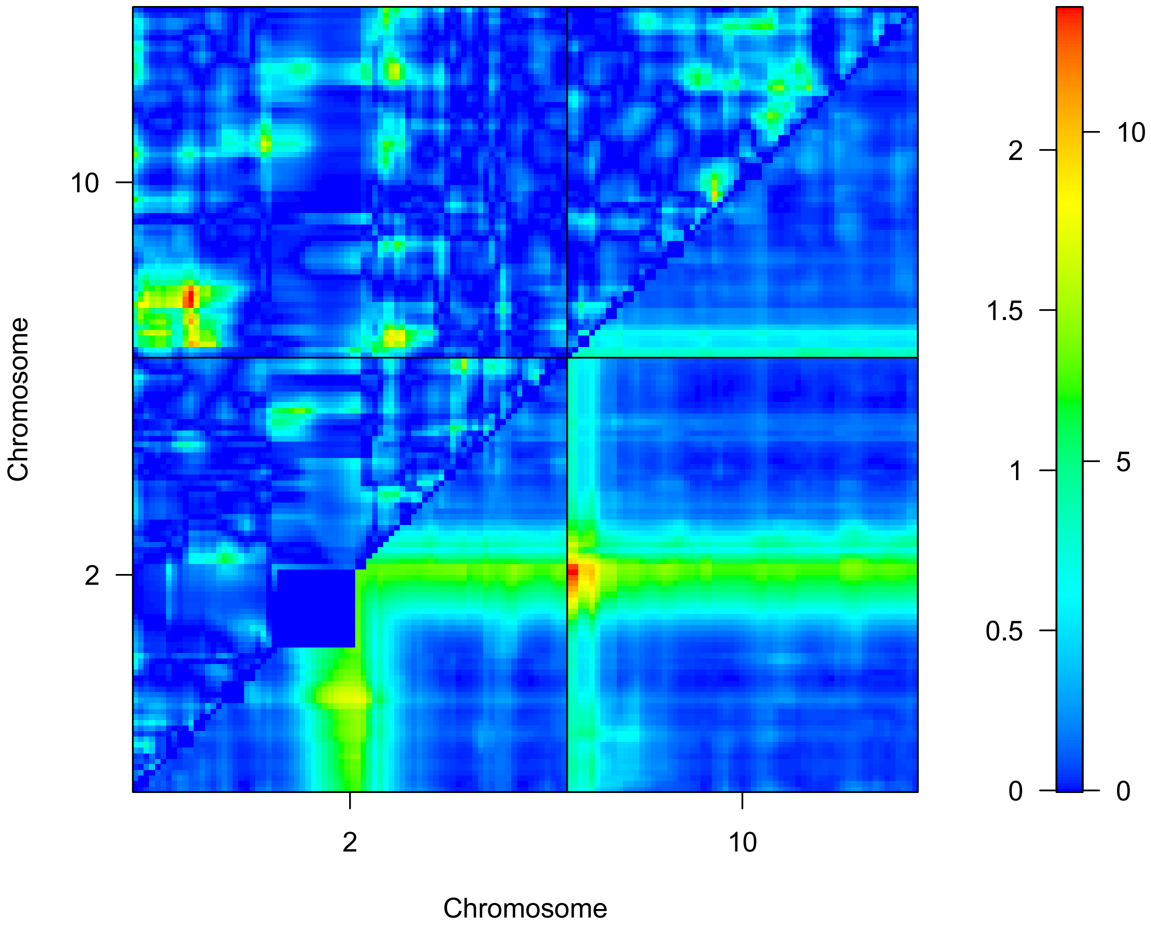


Figure 4.4 Heat map of two-locus genome scan LOD scores with linkage groups (chromosomes) of interest for lemma color. The upper left triangle and left scale bar indicate LOD scores when interaction between loci is considered. The bottom right triangle and right scale bar indicate the LOD scores of the full model.

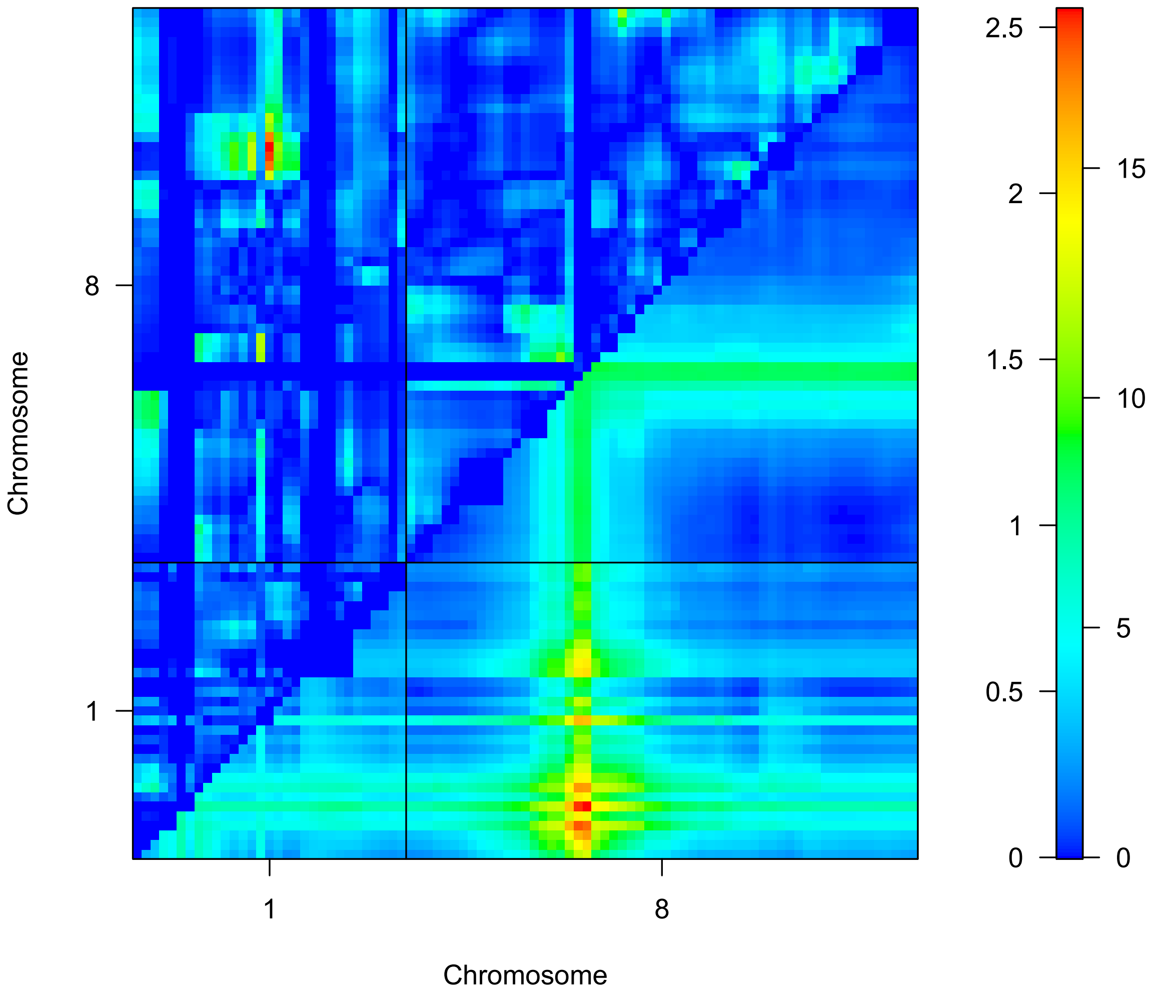


Figure 4.5 Heat map of two-locus genome scan LOD scores with linkage groups (chromosomes) of interest for pubescence. The upper left triangle and left scale bar indicate LOD scores when interaction between loci is considered. The bottom right triangle and right scale bar indicate the LOD scores of the full model.

Table 4.1 LOD score, % phenotypic variance explained, and pointwise chi-square p-value significance of markers closest to QTL for both traits, using MIM.

Model or marker combination	**df**	**LOD score**	**% Variance explained**	**χ^2 p-value**
Lemma color (n = 94)				
Full model	3	16.37	55.16	<0.001
LG 2 @ pos 98.5 cM	2	11.44	33.71	<0.001
LG 10 @ pos 2.0 cM	2	8.59	23.45	<0.001
Interaction	1	1.77	4.067	0.004
Leaf Sheath (n =86)				
Full model	3	19.88	65.52	<0.001
LG 1 @ pos 12.2 cM	2	11.09	27.98	<0.001
LG 8 @ pos 49.1 cM (leaf sheath)	2	13.07	34.93	<0.001
Interaction	1	8.45e-06	0	NS

4.4.2 Change in observed frequency of single trait(s) and trait combinations from 1970-2010

Both pubescence and light lemma increased significantly in overall frequency from 1970s to 2010, whereas the frequency of glabrousness and dark lemma decreased significantly (Fig. 4.6). The overall observed frequency change in leaf sheath pubescence/ glabrousness was 2x greater than the change observed in light lemma color. The frequency of pubescence changed from 0.254 in the 1970s to 0.582 in 2010 ($X^2 = 499.74$, $p < 0.0001$), and the frequency of light lemma color increased from 0.215 in the 1970s to 0.327 in 2010 ($X^2 = 138.58$, $p < 0.0001$)

For morphotype combinations there was a significant increase in the frequency of the classic mesic trait combination (pubescent and light lemma), and the pubescent leaf sheath and dark lemma combination (Fig. 4.7). The frequency of the mesic morphotype combination changed from 0.172 in the 1970s to 0.354 in 2010 ($X^2 = 499.74$, $p < 0.0001$), and the frequency of the dark and pubescent morphotype increased from 0.082 in the 1970s to 0.227 in 2010 ($X^2 = 138.58$, $p < 0.0001$). Changes in the frequency for the mesic morphotype and the dark pubescent morphotype did not significantly differ from each other (Fig. 4.7).

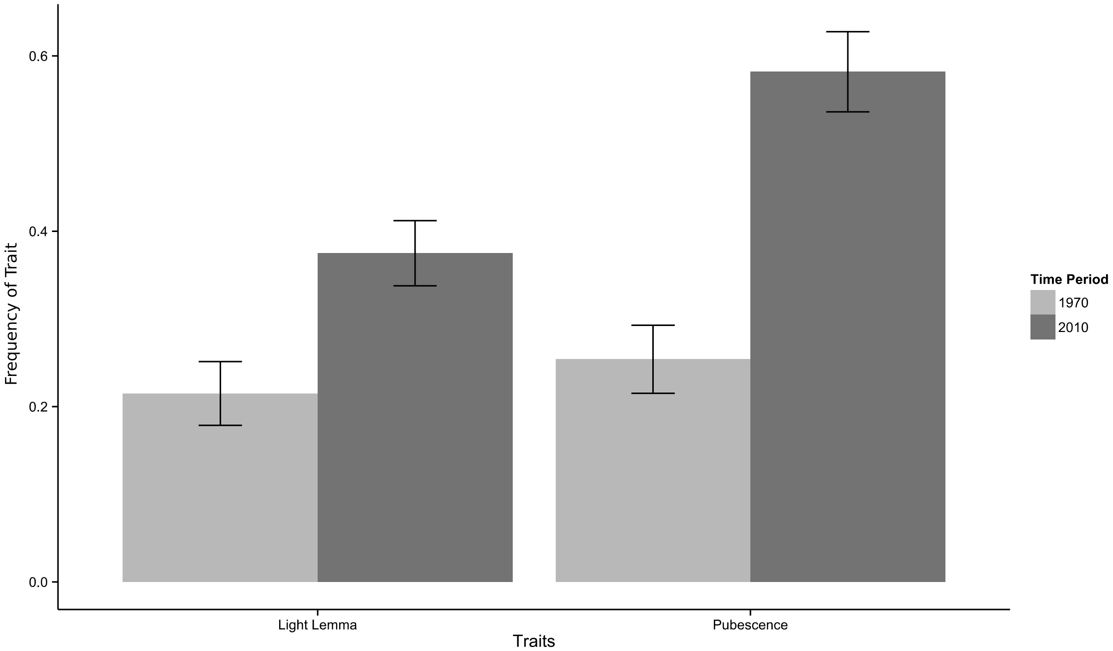


Figure 4.6 Barplot of trait frequencies for light lemma and pubescence in 1970 and 2010 and their standard errors (SE).

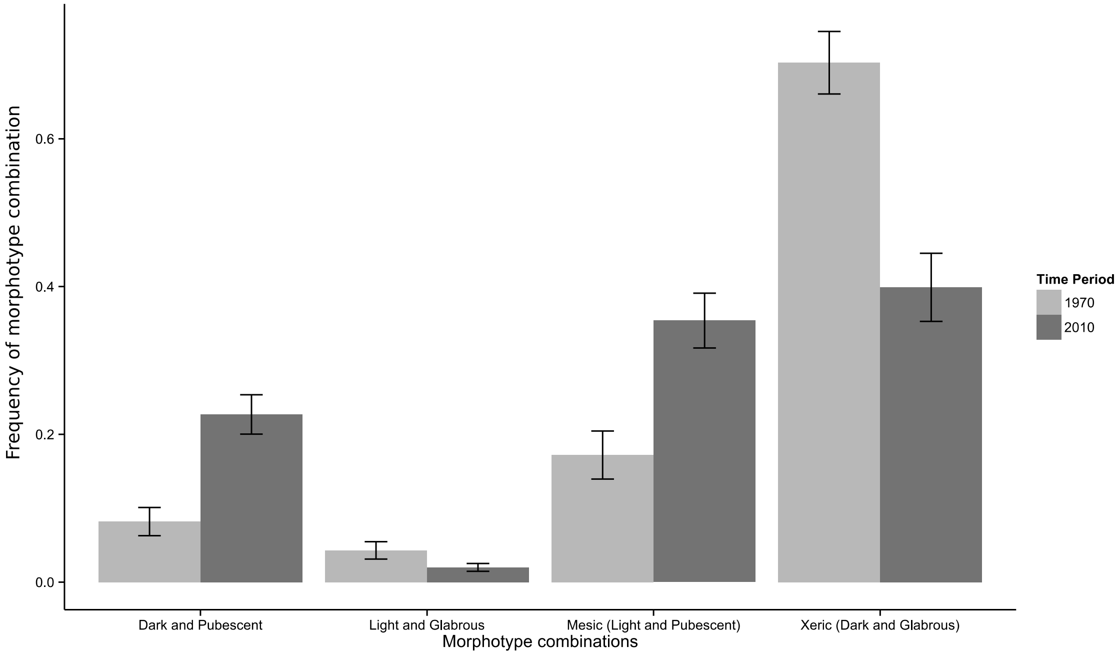


Figure 4.7 Barplot of frequencies of morphotype trait combinations in 1970 and 2010 and their standard errors (SE).

For each time period examined separately, latitude was associated with the frequency of light lemma (Fig. 4.8, Table 4.2). The cline was further north in the 1970s than in 2010, indicating that the trait of light lemma has moved toward more southerly latitudes since the 1970s. The greatest improvement in AIC and model deviance for each decade occurred with latitude relative to the base model (Table 4.2). There was still a fair amount of uncertainty around the estimated fixed effect of latitude for both time periods (Fig. 4.8). However, when data from both time periods were combined, and decade added to the model as a fixed effect, the result was a much lower AIC (Table 4.2), which suggests that the shift in cline through time for light lemma frequency is highly significant (χ^2 =179.65, df =1, p < 0.001).

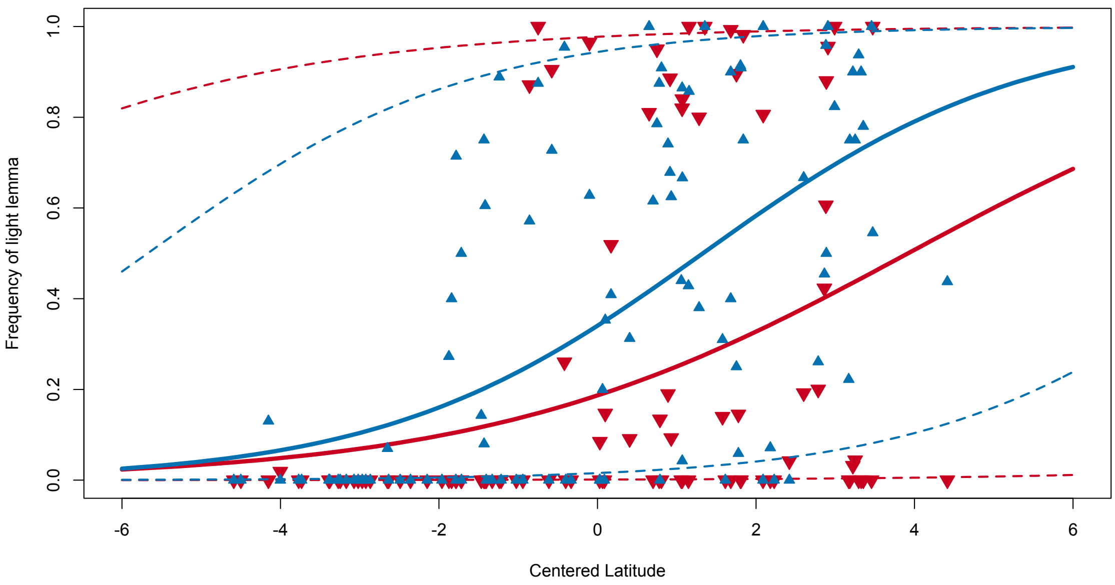


Figure 4.8 Binomial logistic regressions of the frequency of light lemma color at sites against centered latitude. Blue lines and upward triangles represent sites in the year 2010, and red lines and downward triangles sites the 1970s. Hatched lines are the 95% confidence intervals as calculated from probability density function of the logistic distribution.

4.4.3 Simulating random genetic drift with occasional outcrossing

In the absence of selection, some change in the frequency of each trait would still be expected due to occasional recombination and random genetic drift. Thus, the spread of the distribution of simulated trait frequencies is the range of expected outcomes under drift and this occasional recombination. At all but two bands of latitude examined, (in southernmost regions of California where there was almost no variation to begin with) the observed change in frequency from 1970s-2010 differed from the expectations of neutral change for both traits (Figs. 4.9 & 4.10).

Not only did observed 2010 frequencies for traits associated with mesic ecotype differ significantly from the simulated drift distributions; the observed frequencies shifted in the same direction in each band of latitude, shifted in the same direction as predicted by previous common garden results, and fell outside the distribution of outcomes expected under drift (Figs. 4.9 & 4.10). Statewide, Fisher's method revealed that the probability these changes were due drift was miniscule - for both traits: light lemma ($\chi^2 = 63.25$, p = 1.47e-07) and pubescence ($\chi^2 = 73.10$, p = 2.84e-09). This provides evidence that selection, and not drift is largely responsible for the observed trait changes for sites at latitudes greater than 35.0 N in California.

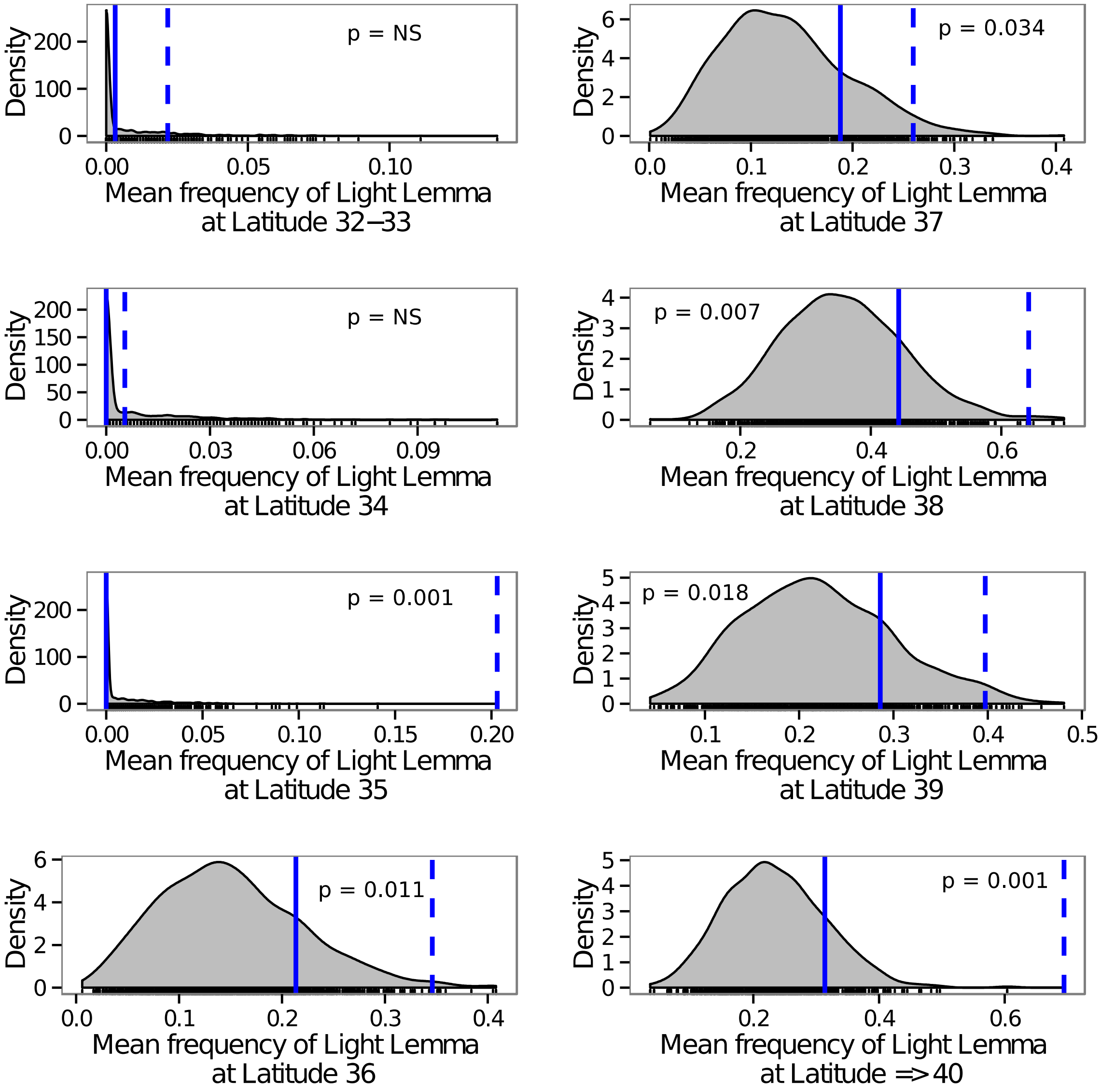


Figure 4.9 Frequency of light lemma observed at each band of latitude group in California for both time period, and distribution of expected frequencies simulation results (grey shaded region and black rug underlying shaded grey region). Solid lines represent observed frequency of light lemma in the 1970s, and dashed lines represent mean frequencies observed in 2010.

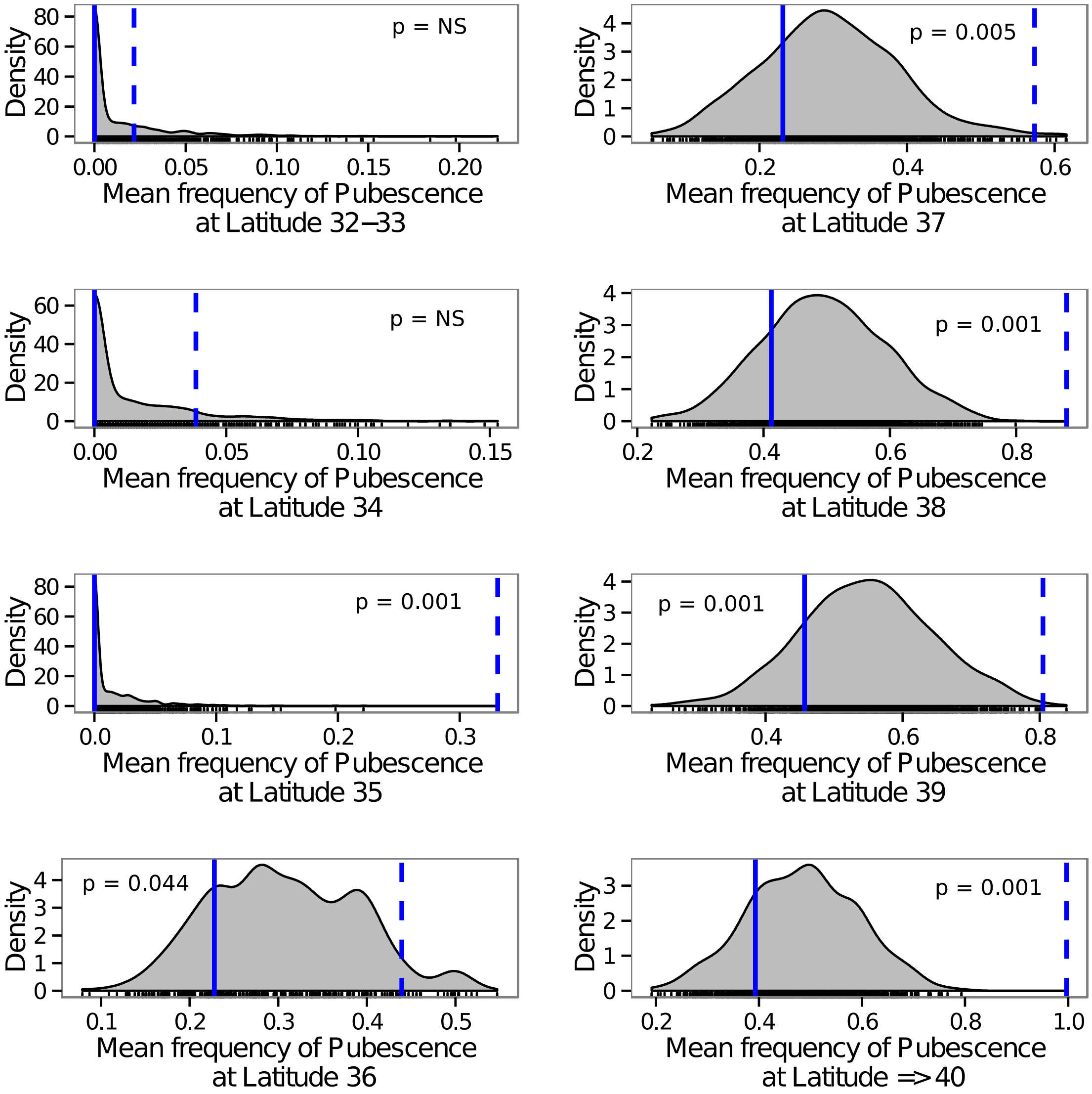


Figure 4.10 Frequency of pubescence observed at each band of latitude group in California for both time period, and distribution of expected frequencies simulation results (grey shaded region and black rug underlying shaded grey region). Solid lines represent observed frequency of pubescence in the 1970s, and dashed lines represent frequencies observed in 2010.

Table 4.2 Repeated measures ANOVA of generalized linear mixed models for both time periods and combined time periods models of lemma color and latitude.

Model	Time Period	AIC	Log-likelihood	Deviance	χ^2	χ^2 df	χ^2 p-value
Site	2010	1991.5	-993.75	1987.5	-	-	-
Site + Latitude	2010	1946.5	-970.23	1940.5	47.03	1	<0.001
Site	1970	1024.5	-510.23	1020.5	-	-	-
Site + Latitude	1970	1013.0	-503.48	1007.0	13.49	1	<0.001
Site	Both	3683.4	-1839.7	3679.4	-	-	-
Site + Latitude	Both	3639.6	-1816.8	3630.8	45.74	1	<0.001
Site + Latitude + Time Period	Both	3462.0	-1727.0	3448.7	179.65	1	<0.001

4.5 Discussion

There were considerable directional shifts in the frequency of both traits statewide since the 1970s. There was an increase in single-trait frequencies associated with the mesic genotype: light lemma and leaf sheath pubescence. The classic xeric traits of dark lemma and glabrous leaf sheath, which once occupied close to 60% of sites (monomorphic for these traits) in California in the 1970s, are being displaced. The change in the frequency of light lemma and pubescence are in the direction as predicted by the previous common garden experiment carried out in the field (Latta 2009).

Both polymorphic traits also have an underlying genetic basis, and are heritable with two QTL underlying each trait. These traits probably have an underlying epistatic model of inheritance, but strong epistasis could not be confirmed with the smaller subset of data used in this analysis (see methods). More GBS data from the complete set of RILs originally used in the experiments of 2003-2007 is needed in order to firmly establish epistatic inheritance, and this is in queue for the near future.

In the absence of selection, under the model of epistatic inheritance (with only occasional recombination and random genetic drift to influence change in frequency) there should be an increase in both the frequency of dark lemma and pubescent leaf sheaths. This is because under occasional recombination and genetic drift, both dark lemma and pubescent leaf sheaths are produced 3:1 over light lemma and glabrous leaf sheaths. In any case, the change in light lemma color is in the opposite direction of what would be predicted under drift (it increases in the field), and further the observed frequency in the field lies outside the distribution of expected outcomes in the simulations. The change in pubescence is in the predicted direction, but significantly greater than the expected change in frequency under drift. Based on these simulation results, neutral processes do not explain the observed changes in trait frequencies, and selection remains the primary evolutionary force responsible for contemporary change in California.

It is particularly remarkable that these observations are directly in line with predictions from earlier common garden experiments where mesic had highest fitness in the field (Latta 2009), and certain RILs had highest fitness in novel environments (glasshouse) (Johansen-Morris and Latta 2006; 2008). However, it is unlikely that there was selection directly for light lemma color and leaf sheath pubescence. Instead, light lemma and pubescent leaf sheaths merely serve as markers for tracking the spread of the mesic genotype, and mesic-recombinant genotypes.

It is interesting to note that one of the QTL for lemma color occurs close (within 3cM) to a QTL for RIL fitness in the field (Latta *et al.* 2010), and is suggestive of linkage disequilibrium (LD). The mesic allele at this QTL is associated with light lemma and higher mean fitness (than those RILs with the xeric allele – Appendix 2). As there is only occasional recombination for *A. barbata*, LD for these two QTL would only breakdown once a new genotype arrived in a population. Otherwise under high selfing, this association would likely be maintained as a positive genetic correlation through time, allowing lemma color to be dragged along by hitchhiking.

As stated in the introduction, there was no reason to hypothesize that either morphological marker was the direct target of selection, and a brief selection gradient analysis with mean fitness data of the RILs from the multi-year common garden reciprocal transplant experiment, shows no association between mean fitness and pubescence nor light lemma color.

Speculatively, it was observed that individuals with light lemma color appear to have larger seed sizes than dark lemma individuals, but no formal measurements were made with the RILs or accessions collected from the field in 2010, nor are the seed collections from 1970s available. Seed size has been observed to be associated with seed color in the common bean (Sax 1923), and it is known that large seed size can have a substantial effect on fitness (Stanton 1984). However, this conjecture must not go too far, as the relationship between seed (offspring) size and fitness may be quite complicated (Rollinson and Hutchings 2013).

Under linkage equilibrium, individuals with a dark lemma and pubescent phenotype combination have increased from their initially low frequencies since the 1970s. Increases in both dark lemma and leaf sheath pubescence are predicted under occasional recombination due to the suspected epistatic QTL model for both traits. Past work with RILs suggested that novel recombinants have higher fitness in novel environments, and thus could be spreading throughout California (Johansen-Morris and

Latta 2006; Johnson et al. 2008). However, this spread cannot be differentiated from selection for a novel recombinant over and above the selective spread of mesic-like traits.

There were two bands of latitude where almost no change in frequency occurred for either trait. These populations were at the southernmost extent of both geographic surveys, where most populations are fixed for traits associated with the xeric genotype. Southern California may have been the site of the initial introduction and founding population during the Spanish missionary period. Following this introduction, genotypes with dark lemmas and glabrous leaf sheaths likely spread northward, with the statewide mean frequency for the dark and glabrous trait combinations being close to 0.78 in the 1970s (Supp. Material 1). Since the 1970s, the latitudinal cline has shifted through time with the dark lemma moving further south. The shift in cline suggests a possible second introduction of genotypes with light lemma has occurred in the North (close to the latitude of San Francisco – a major port), and spread – either displacing genotypes with the dark lemma due to selection or recombining with them as seeds migrated.

Aside from natural selection there are few possibilities that can explain such a large increase in the frequency of genotypes with mesic traits at latitudes greater than 35.0 N. One remote possibility is that the seed bank at many sites has been built up by past migration events and these genotypes have recently emerged in the past 35-40 years increasing the genetic diversity in Northern California. This increase in genetic diversity via the seedbank has been observed in another Californian annual *Clarkia springvillensis* (McCue and Holtsford 1998), whose seed bank contains seeds from multiple cohorts that persist for multiple years. Another possibility is that of idiosyncratic or asymmetric migration events, where extreme propagule pressure has increased the density of individuals of one or two types so much so that in our 2010 survey as the most frequent morph that was present at the site in the 1970s was not detected. It cannot be known whether these scenarios occurred nor where, and both are perhaps plausible for a few of the sites, but very improbable for over a hundred sites in California.

Overall, it is the large geographic scale of this adaptive change that is both notable and unique. Few studies examining selection have had access to historical datasets like those of *A. barbata* populations in California, and studies that do have access to large temporally detailed datasets have been confined to smaller geographic areas such as islands (Grant and Grant 1993; Milner et al. 1999; Grant and Grant 2002), or a small pond (Cousyn et al. 2001), or have focused on one (Franks et al. 2007) or a few sites across a wide geographic range (Carroll et al. 1997). The only other study to have estimated selection on a heritable character over a short time period in a large geographic range are Kettlewell's moths (Kettlewell 1956; Cook 2000; Mathieson and McVean 2013). While the unidirectional shift in two heritable characters is conspicuous, these changes in frequency are not beyond the realm of modest selective pressures that have been commonly reported in the literature (Kingsolver et al. 2001; Kingsolver and Diamond 2011). Overall, the shift observed for leaf sheath pubescence (+32.7%) was over 2x greater than the shift toward light lemma color (+16.0%). In conclusion, this study may have be the first to have characterized an adaptive change in genotypic frequency in two characters, in wild populations, over a relatively short period (40 generations), and throughout a very large geographic area – nearly the entire state of California.

4.6 References

Allard, R. W., G. R. Babbel, M. T. Clegg, and A. L. Kahler. 1972. Evidence for coadaptation in Avena barbata. Proc. Natl. Acad. Sci. U.S.A. 69:3043–3048.

Arends D, Prins P, Jansen RC, Broman KW (2010) R/qtl: high-throughput multiple QTL mapping. *Bioinformatics* **26**, 2990–2992.

Broman KW, Wu H, Sen S, Churchill GA (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889-890.

Carroll SP, Hendry AP, Reznick D, Fox CW (2007) Evolution on ecological time-scales. *Functional Ecology* **21**, 387–393.

Carroll SP, Dingle H, Klassen SP (1997) Genetic differentiation of fitness-associated traits among rapidly evolving populations of the soapberry bug. *Evolution* **51,** 1182–1188.

Clegg MT (1972) Patterns of Genetic Differentiation in Natural Populations of Wild Oats. University of California, Davis. Doctoral Dissertation.

Clegg M, Allard R (1972) Patterns of genetic differentiation in the slender wild oat species *Avena barbata*. *PNAS*, **69**, 1820–1824.

Coffman FA (1964) Inheritance of morphologic characters in *Avena*. Technical Bulletin 1308. US Dept. of Agriculture.

Cook L (2000) Changing views on melanic moths. *Biological Journal of the Linnean Society* **69,** 431–441.

Cousyn C, DeMeester L, Colbourne JK, Brendonck L, Verschuren D, Volckaert F (2001) Rapid, local adaptation of zooplankton behavior to changes in predation pressure in the absence of neutral genetic changes. *PNAS* **98**, 6256–6260.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic markerdiscovery and genotyping usingnext-generation sequencing. *Nature Reviews Genetics* **12**, 499–510.

Dobzhansky T, Levene H (1948) Genetics of natural populations. XVII. Proof of operation of natural selection in wild populations of *Drosophila pseudoobscura*. *Genetics* **33**, 537-547.

Franks SJ, Sim S, Weis AE (2007) Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *PNAS* **104**, 1278–1282.

Gardner KM, Latta RG (2006) Identifying loci under selection across contrasting environments in *Avena barbata* using quantitative trait locus mapping. *Molecular Ecology* **15**, 1321–1333.

Grant BR, Grant PR (1993) Evolution of Darwin's Finches Caused by a Rare Climatic Event. *Proceedings of the Royal Society B: Biological Sciences* **251**, 111–117.

Grant PR, Grant BR (2002) Unpredictable Evolution in a 30-Year Study of Darwin's Finches. Science **296**, 707–711.

Hutchinson ES (1982) Genetic Markers and Ecotypic Differentiation of *Avena barbata* Pott ex Link. University of California, Davis. Doctoral Dissertation.

Johansen-Morris AD, Latta RG (2008) Genotype by environment interactions for fitness in hybrid genotypes of *Avena barbata*. *Evolution* **62**, 573–585.

Johansen-Morris AD, Latta RG (2006) Fitness consequences of hybridization between ecotypes of *Avena barbata*: hybrid breakdown, hybrid vigor, and transgressive segregation. *Evolution* **60**, 1585–1595.

Johnson MTJ, Dinnage R, Zhou AY, Hunter MD (2008) Environmental variation has stronger effects than plant genotype on competition among plant species. Journal of Ecology **96**, 947–955.

Kettlewell H (1956) A Resume of Investigations on the Evolution of Melanism in the Lepidoptera. *Proceedings of the Royal Society of London Series B* **145**, 297–303.

Kingsolver J, Diamond S (2011) Phenotypic selection in natural populations: what limits directional selection? *American Naturalist* **177**, 346–357.

Kingsolver J, Hoekstra H, Hoekstra J, Berrigan D, Vignieri S, Hill C, Hoang A, Gibert P, Beerli P (2001) The strength of phenotypic selection in natural populations. *American Naturalist* **157**, 245–261.

Latta RG (2009) Testing for local adaptation in Avena barbata: a classic example of ecotypic divergence. Molecular Ecology **18**, 3781-3791

Latta RG, Gardner KM, Staples DA (2010) Quantitative trait locus mapping of genes under selection across multiple years and sites in *Avena barbata*: epistasis, pleiotropy, and genotype-by-environment interactions. *Genetics* **185**, 375–385.

Latta RG, MacKenzie J, Vats A, Schoen D (2004) Divergence and variation of quantitative traits between allozyme genotypes of *Avena barbata* from contrasting habitats. *Journal of Ecology* **92**, 51–71.

Mathieson I, McVean G (2013) Estimating Selection Coefficients in Spatially Structured Populations from Time Series Data of Allele Frequencies. *Genetics,* **193**, 973-984.

McCue K, Holtsford T (1998) Seed bank influences on genetic diversity in the rare annual Clarkia springvillensis (Onagraceae). *Am. J. Bot.* **85**, 30-36.

Miller, R. D. 1977. Genetic Variability in the Slender Wild Oat Avena barbata in California. University of California, Davis. Doctoral Dissertation.

Milner JM, Albon SD, Illius AW, Pemberton JM, Clutton-Brock TH (1999) Repeated selection of morphometric traits in the Soay sheep on St Kilda. *Journal of Animal Ecology* **68**, 472–488.

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE* **7**, e32253.

Reznick D, Ghalambor CK (2001) The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. *Genetica* **112-113**,183–198.

Rollinson N, Hutchings JA (2013) The relationship between offspring size and fitness: integrating theory and empiricism. *Ecology* **94**, 315–324.

Sax, K (1923) The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. *Genetics* **8**, 552–560.

Stanton ML (1984) Seed variation in wild radish: effect of seed size on components of seedling and adult fitness. *Ecology* **65,** 1105–1112.

Chapter 5 A refutation of the reproductive economy hypothesis

5.1 Abstract

Under crowded conditions, plant populations typically exhibit L-shaped distributions of size. The hypothesis of “reproductive economy” proposes that since many plants remain small and suppressed in these populations, more offspring for the next generation come from smaller plants than from large plants, and therefore that the ability to reproduce while still small is favoured over the ability to become large. I tested this idea using four years of field data at two sites in California, USA from an experiment with the annual grass *Avena barbata*. Despite strongly skewed size distributions, the frequencies of the genotypes giving larger size always increased between parents and offspring while those of the smallest size class decreased. Directional selection gradients were also always positive and significant, indicating that natural selection indeed favours the few larger plants over their more numerous smaller neighbours. I develop a simple model using exponential distributions of size within morphs that differ in their growth potential and size threshold for reproduction. This model shows that selection consistently favours the morph with greater potential to become large, even if that potential comes at the cost of a larger size threshold for reproduction. Neither the empirical nor theoretical findings support the reproductive economy hypothesis.

5.2 Introduction

Size is usually positively related to fecundity in plants. Thus, bigger plants typically contribute more offspring than smaller plants to the next generation (Harper 1977). However, plant size is strongly affected by environmental conditions (e.g. competition, pathogens, predation, drought). In crowded, competitive environments, plants typically show L-shaped size distributions, with many more small plants than large plants.

Recently, Aarssen (2008) has hypothesized that selection favours ‘reproductive economy’, the ability of suppressed plants to reproduce despite small size. In support of this view, Chambers and Aarssen (2009) noted that smaller, more abundant plants collectively contribute more offspring to the next generation than the fewer, larger individuals in the population, and interpret this observation as evidence that large size is not favoured by selection under crowded conditions. Instead, they argue, natural selection favours plants which forego some growth (and hence reproductive potential) in return for a greater probability of reproducing at least some seeds. Since crowded conditions are the norm for most plant populations they hypothesize that this reproductive economy strategy (Aarssen 2008, Neytcheva and Aarssen 2008) will be favored in many plant populations. While it is certainly true that foregoing some growth, which is often costly, in order to reproduce earlier can be an optimal strategy under some conditions; this does not translate into natural selection favouring plants of smaller size. Rather, it is one example of selection on a life history trade-off (e.g. Stearns, 1992)

Moreover, by focusing on the total contribution of offspring by a large pool of suppressed small plants, Aarssen's (2008) reasoning conflicts with the standard model of natural selection. Selection does not hinge on the class of individuals that contribute the majority of the offspring *in toto*. Rather it is the phenotype that contributes more offspring *per capita* which will be favored, because this drives the increase in frequency between generations (Fisher 1930; Endler 1986; Orr 2009). Because a large plant will typically leave more offspring than a small plant, the progeny of large plants will make up a higher proportion of the progeny *than did their parents,* even if this proportion is nevertheless still small.

To illustrate this idea consider the fecundity data presented by Chambers and Aarssen (2009) for *Cardamine parviflora* across individuals of different size. If it is assumed for simplicity that all seeds are equally likely to germinate in the following growing season, one can estimate the contribution of each size class to the next generation. Comparing Chambers and Aarssen's (2009) Figure 1a and 1d shows that change in the relative frequencies is increasing in the largest class sizes and decreasing in the smallest class sizes (Fig. 5.1). This shows that selection indeed favours larger individuals in this generation, despite the fact that smaller individuals are far more common. Inspection of the direction of change in the relative frequencies for all 21 species examined in Chambers and Aarssen (2009) shows an overall increased frequency of the largest class sizes and overall decreased frequency of the smallest class sizes.

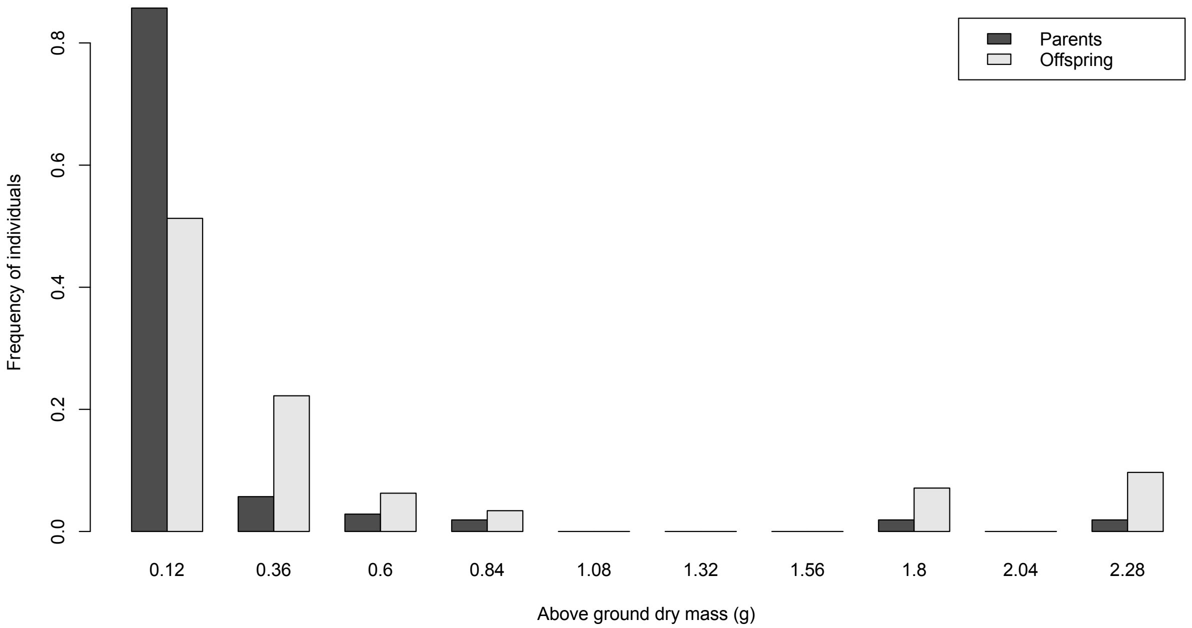


Figure 5.1 Selection on size in *Cardamine parviflora* (redrawn from Chambers and Aarssen 2009). Size distribution of parents (black bars) and of each of the ten parental size class' contribution to total fecundity (white bars). Values underneath the bars are the midpoint values of each of the same ten size classes specified by Chambers and Aarssen (2009).

Here, I examine whether skewed size distributions support the reproductive economy hypothesis in plant populations using two different approaches. First, I analyze field data from recombinant inbred lines (RILs) of an annual grass, *Avena barbata* Pott ex Link (the slender wild oat) to determine whether smaller or larger plants leave more offspring to the next generation. The central advantage of using RILs here is that each individual belongs to one of 190 known genotypes, allowing us to examine the change in genotype frequency between generations, and hence the direction of evolutionary change.

Second, I use exponential distributions to model selection on the trade-off between two competing life-history strategies. I seek to determine whether the skewed size distribution favours the morph which trades off some growth potential in order to reproduce at a smaller size – i.e. the strategy described as reproductive economy.

5.3 Methods

5.3.1 Avena barbata field data

Avena barbata Pott ex Link (the slender wild oat) is a highly-selfing annual grass that is native to the Mediterranean and introduced to other Mediterranean climate zones around the world. Data for the analysis were obtained from a study on local adaptation of the oats in California, USA using a set of recombinant inbred lines (RILs). These had been created from a cross between the moist-associated (mesic) and dry-associated (xeric) ecotypes described by Allard *et al*. (1972). Full details of the crossing and experimental design are given in Gardner and Latta (2006) and Latta (2009).

Common gardens were established at two sites – Hopland (moist) and Sierra Foothills (dry), and fitness was estimated in each of four growing seasons spanning a range of inter-annual variation (Latta 2009). Plants were sown into otherwise undisturbed vegetation from which they were fully exposed to competition. The same numbers of each genotype (RIL) were planted in a randomized block design within each common garden plot. It is thus possible to compare each genotype's frequency among the parents to its respective frequency in the progeny. Our dataset included years and sites of particularly high (Hopland 2006) and low (Hopland 2004 and 2007) viability, size, and fecundity. *A. barbata* produces seeds in spikelets, each containing two self-pollinated seeds and the glumes are retained on the plant after the seeds have dropped allowing them to be counted. The number of spikelets on an individual thus gives a good estimate of lifetime reproductive output (Marshall and Jain 1969). I can therefore test whether genotypes of large or small parents tend to increase in frequency in the next generation.

I began with a phenotypic analysis in which I included only those individuals that survived to reproduce. Within each year-site combination I counted individuals within ten equal size classes, spanning the range of aboveground dry mass. To calculate expected frequencies for the next generation, I summed the fecundity of plants within each size bin, and then expressed these as a proportion of the total fecundity across all size bins. I then compared the frequency of large size classes among the parents to the frequency of their seeds among the progeny.

To examine the change in genotype frequencies (i.e., the response to selection) I used the mean size of reproductive plants in each RIL as a measure of that genotype's size, and grouped these means into ten size classes as above. The total number of seeds produced by each genotype gives its contribution to the next generation. Since each genotype was sown into the gardens at an equal frequency, I can measure the total change in frequency for genotypes of each size class in the progeny. By comparing frequencies in the next generation to the frequency of the genotypes at sowing, this measure of the response to selection includes any tendency for some genotypes to survive to reproduce more readily than others. Positive changes for a given size bin would be indicative of selection favouring genotypes that tend to produce plants of that size.

Finally, I assessed whether directional selection acts on size for each year-site combination by regressing the standardized above ground dry mass against relative fitness (Lande 1982). I did this for phenotypic and genotypic means (Rausher 1992), where a positive slope value for the selection coefficient (*β*) indicates natural selection favours larger plants.

5.3.2 Theoretical simulations

I modeled selection under skewed size distributions, assuming an exponential distribution of adult plant sizes. I considered two strategies, in which one ('economy') reproduces at a lower threshold, but in order to do so, foregoes some potential to attain large size. Therefore 'economy' morphs are drawn from an exponential distribution with a smaller mean size ($\bar{X} = 1/\lambda$, where λ is the rate parameter of the exponential distribution) than the alternative 'full size' morph. Each morph reproduces if it exceeds a size threshold proportional to its mean (Fig. 5.2). Fecundity is proportional to the amount by which the final plant size exceeds this threshold. The fitness of each morph is the expected lifetime reproductive success integrated over the distribution of size classes.

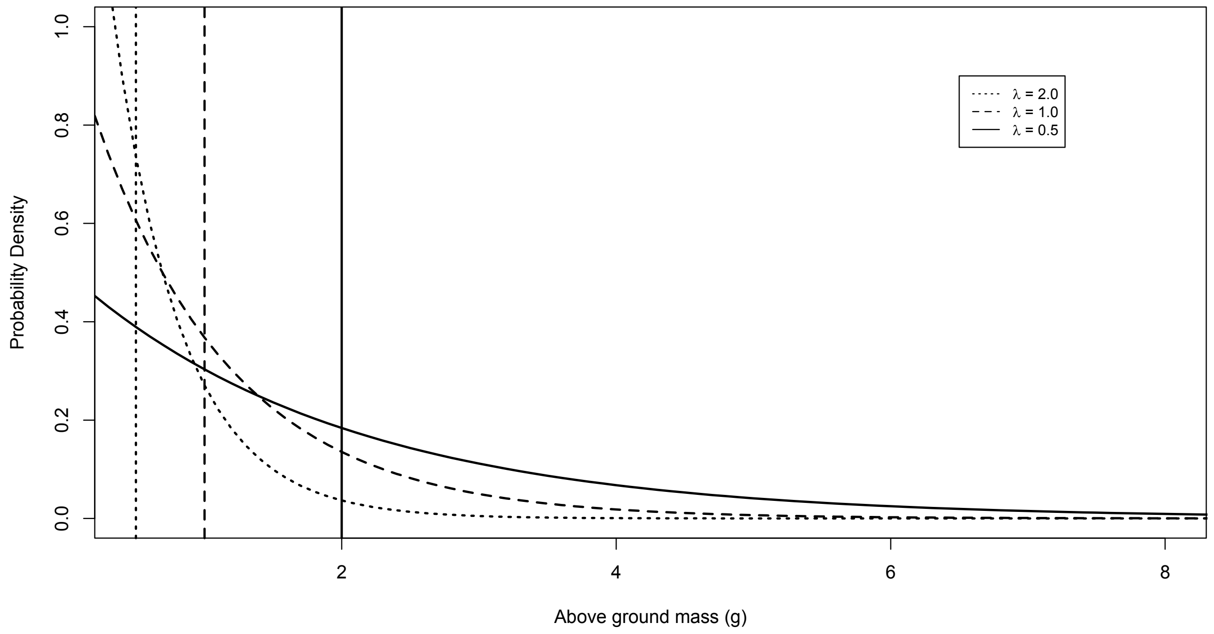


Figure 5.2 Examples of exponential functions modeling the L-shaped size distribution used in the simulation. Curves represent size distributions of potentially different life-history morphs. Corresponding vertical lines represents the minimum size threshold required for reproduction in each morph. The individuals with sizes to the left of the vertical lines are those small individuals that die without reproducing. λ = 2 represents a low average size and a low threshold for reproduction, λ = 0.5 represents a large average size and a large threshold for reproduction.

This model is illustrated by simulating the case for λ = 1.100 (lower mean size – economy) and λ = 0.900 (larger mean – full size), setting the threshold for reproduction equal to the mean size for each distribution. It was assumed that reproductive plants produce *k* seeds per unit mass above the threshold size for reproduction (i.e., a plant 1 unit larger than the threshold produces *k* seeds, where k is set large enough that the population can replace itself despite many individuals failing to reproduce). I simulated the seeds of 1000 plants, assuming for simplicity that seeds were of the same strategy as their parent. 1000 seeds were sampled without replacement to propagate into the next generation and recorded the changes in frequency of the morphs over 50 generations, running 1000 independent simulations. All simulations were conducted in the R v. 2.1.12 base package (R Core Development Team, 2011) (see Appendix 4 for script).

5.4 Results

5.4.1 Empirical

For every year, at each site, counts of individuals in the *A. barbata* plots were highly right skewed toward plants of small size (Fig. 5.3). Consequently, the majority of the seeds were derived from parents who fell in the smallest size classes. However, these progeny of small parents made up a lower proportion of the progeny pool than the small plants accounted for among the parents. The opposite was true of large plants. These accounted for only a small proportion of either the population, or the seed pool, but their relative frequency increased between the parents and the progeny.

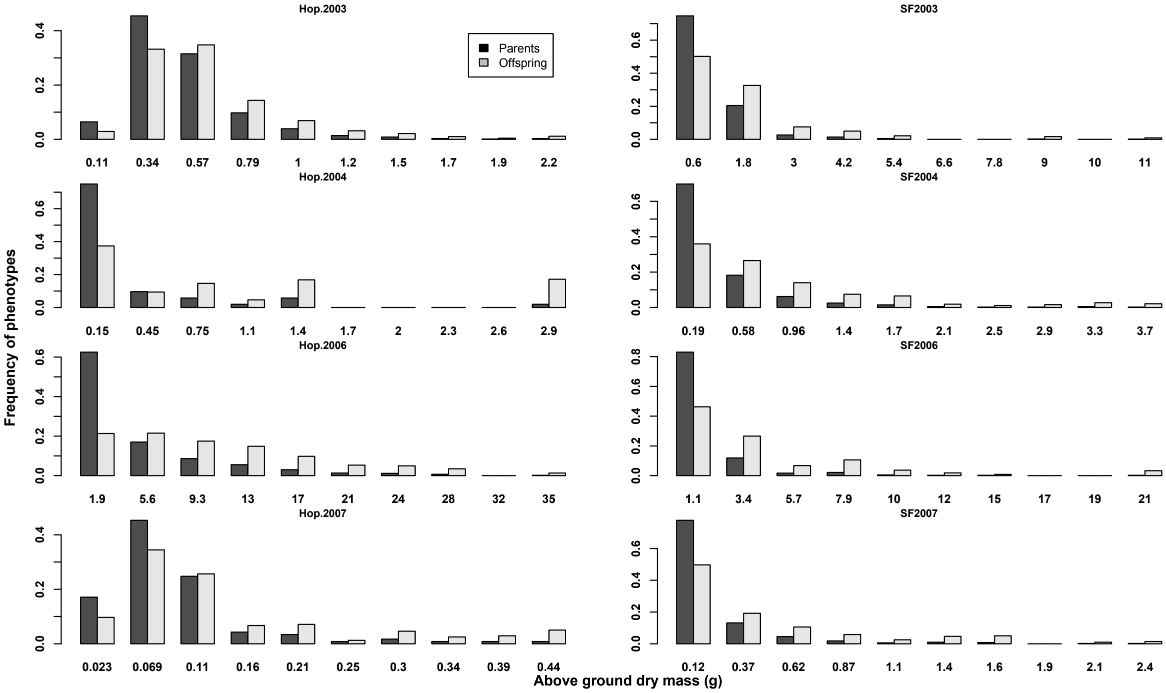


Figure 5.3 Phenotypic size distribution for parent plants (black bars) and of each parental size class' contribution to total fecundity (white bars) in each year at each site in *A. barbata*.

The distribution of mean size in the genotypes was less skewed, but still showed a pronounced right skew (Fig 5.4). As for phenotypes, the majority of the progeny had genotypes associated with small plant size. However, the frequency of genotypes producing small plants declined between parents and offspring while the frequency of genotypes producing large plants (though fewer) increased in frequency. Additionally, all standardized linear selection gradients are positive and largely significant (Table 5.1). Therefore selection clearly favours larger plant size, and increases the frequency of those genotypes that produce it.

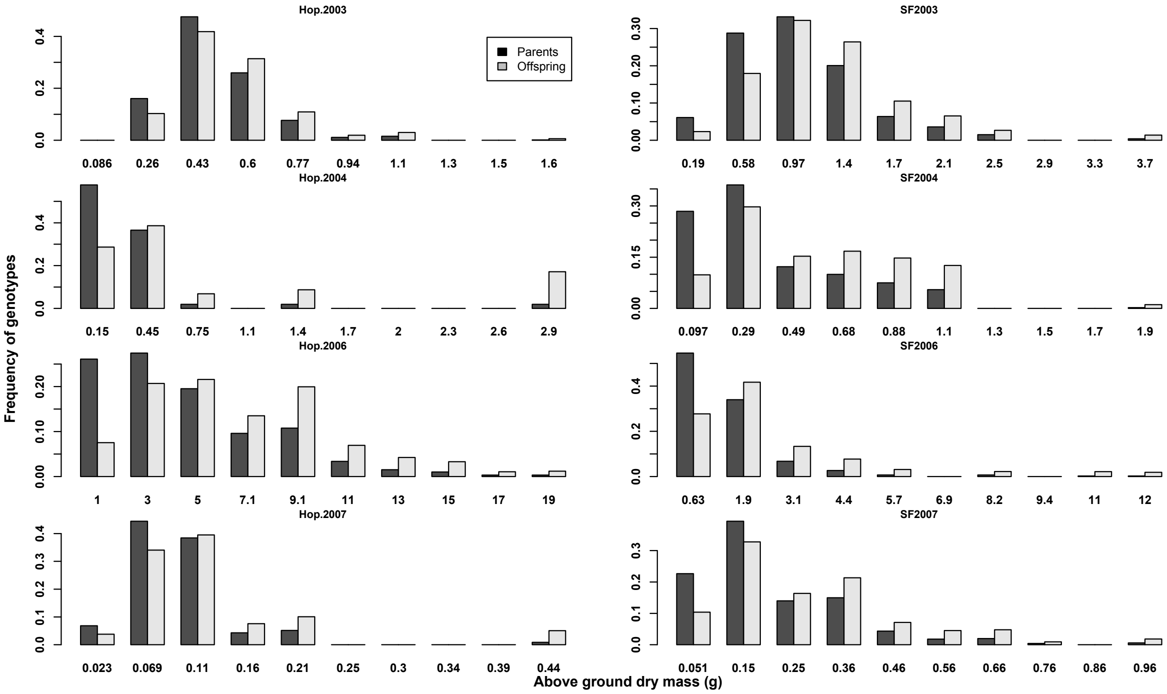


Figure 5.4 Distribution of genotypic mean size in parent plants (black bars) and offspring (grey bars) in each year at site in 190 recombinant inbred lines of *A. barbata*. Note different size scales (x-axis) than those displayed for phenotypes in Fig 5.3.

5.4.2 Theoretical

Exponential distributions have the property that a left truncated distribution is another exponential distribution with the same rate (λ) parameter (Bolker 2008). Thus, the distribution of the size above the reproductive threshold is also exponentially distributed and has the same mean as the distribution of the sizes themselves (Fig 5.2). Since fecundity is proportional to size above this reproductive threshold, fecundities of the full size and economy morphs also have means proportional to the mean size potential of each morph. In addition, if the thresholds are proportional to the mean size of each morph then the survival to reproduce is equal for the two morphs, because an equal proportion of the individuals will fall above the threshold in each case. Since an equal proportion of each of the morphs are reproducing, but the larger morph has higher mean fecundity, I expect the larger morph to have higher mean fitness. In other words, the lower threshold for reproduction cannot offset the reduced potential for growth unless these two are disproportionate to each other – that is, unless the difference between the reproductive thresholds of the two morphs is greater than the difference between their mean size potentials.

The results of our simulations indicated that the frequency of the full sized morph almost always reached fixation (*p* =1.0) in 50 generations in each of the 1000 simulated runs (Fig. 5.5). The frequency of *p* reached fixation on average after 24.3 generations. In eight runs which did not reach fixation *p* was close to fixation by the 50th generation ($p \geq 0.98$). In all cases, the right skewed size distribution was evident throughout the run. As observed in A. barbata (Fig. 5.3), the larger size classes increased in frequency, while the smaller size classes decreased. Because the full size morph made up a larger proportion of these larger size classes (Fig. 5.5, lower panels), natural selection drove it to fixation.

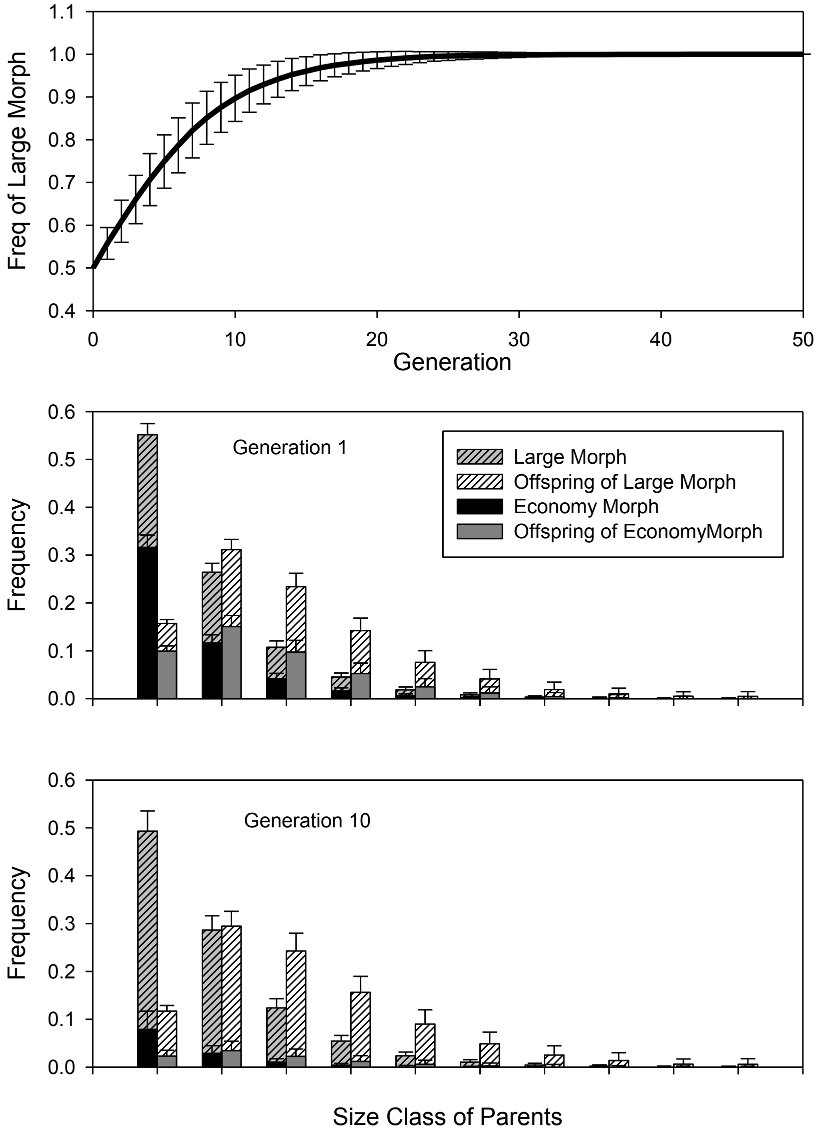


Figure 5.5 Simulation results. Top panel: trajectory of change in the frequency of the 'large' morph over time. Middle panel: Size distribution among parents and contribution to offspring in the first generation of the simulation. Morph frequencies within each size class are indicated. Bottom panel: Size and morph distribution after ten generations of selection. Error bars represent one standard deviation over 1000 runs of the simulation with λ of 0.9 and 1.1, and k =5 seeds per unit mass above the threshold for reproduction.

5.5 Discussion

Plant populations tend to consist of very many small plants and few large plants (Harper 1977), and it is from this observation that Chambers and Aarssen (2009) argue: "Natural selection therefore undoubtedly favours larger plant size, but just not most of the time". Our analyses show that despite persistently L-shaped distributions for size, natural selection nevertheless favours larger size plants under a wide range of conditions (Table 5.1). Relative change between parents and offspring in phenotype and genotype frequencies was positive in the larger size classes and negative for the very small size classes for every year site combination (Figs. 5.3 and 5.4). This occurred despite a wide range of conditions experienced by the plants in different years and at the different sites (Latta 2009). The trait of above ground mass is a target of natural selection in *A. barbata*, with larger plants clearly being favoured across all year-site combinations (Table 5.1).

Our simulations indicate that when fecundity is related to size, selection acts against plants with lower size threshold for reproduction if that lower threshold is matched by a lower size potential. Of course if the lowered threshold for reproduction is less than the reduced potential for large size, then selection might well favour a morph showing reproductive economy. In the extreme if a lower threshold for reproduction invoked no reduction in size potential it would obviously be favoured by selection (it would be tantamount to an increase in survival which, *ceteris paribus,* is advantageous). But if the traits are thus uncoupled, selection for the lower threshold would in no way over-ride or negate selection for large size as Aarssen (2008; Chambers and Aarssen, 2009) argues.

I suggest that the hypothesis of "reproductive economy" is better modeled as a life-history trade-off, wherein selection favours large size, but size may be negatively associated with other components of fitness. Increased size often comes at the cost of delayed maturity (Roff, 2001). Reproductive economy – the ability to produce at least some seeds even though small – is likely the outcome of this trade-off, since selection presumably acts against plants which delay reproduction indefinitely in order to grow larger. Alternatively, if growing larger entails a greater risk of mortality, then the total reproductive success of genotypes with higher growth might be outweighed by their lower survival. However, these are not the case in *A. barbata*, because genotypes with large size also show high survival and early flowering (Latta and McCain 2009).

Table 5.1 Average above ground mass, average fecundity, percent survival, and the standardized linear selection gradients (ßs) and their standard errors (SE) on size for both phenotypes and genotypes of *A. barbata* for each year-site combination. Significant values for ßs are indicated in bold. *P < 0.00001

Site/Year	Mean Mass	Mean Fecundity	Percent Survival	Phenotypic Selection (β)	Phenotypic Selection (SE)	Genotypic Selection (β)	Genotypic Selection (SE)
Hopland 2003	0.494	8.6	75%	**+0.44***	0.015	**+0.302***	0.008
Hopland 2004	0.344	6.2	6%	**+1.33***	0.066	**+1.306***	0.05
Hopland 2006	4.506	83.3	82%	**+1.07***	0.011	**+0.653***	0.008
Hopland 2007	0.096	2	15%	**+0.65***	0.034	**+0.541***	0.028
Sierra 2003	1.015	18.1	97%	**+0.79***	0.01	**+0.414***	0.005
Sierra 2004	0.374	7.9	55%	**+0.99***	0.021	**+0.630***	0.013
Sierra 2006	1.401	30.1	83%	**+1.26***	0.022	**+0.831***	0.015
Sierra 2007	0.218	5.5	48%	**+0.95***	0.014	**+0.488***	0.009

Other tradeoffs with size can be imagined. If highly fecund plants produced seeds that were less viable, some optimal intermediate would be favoured (Smith and Fretwell, 1974). However this size-number trade-off applies only among plants with a similar pool of total resources and is not expected when high fecundity is the result of some plants having greater resource acquisition (as in competitive scenarios) (de Jong and van Noordwijk 1992; Reznick *et al.* 2000). Moreover, each of these potential trade-offs are entirely compatible with selection for large size under crowded or competitive conditions.

Whatever the outcome of these trade-offs, I must re-emphasize that it is the *per capita* contribution of different genotypes or phenotypes to the next generation that drives natural selection, and not the total contribution. Therefore the observations that most plants are small (Aarssen, 2008) and most seeds come from small plants (Chambers and Aarssen, 2009) do not indicate selection. For any modest selection pressure (and selection can act on very minor fitness differences), the more abundant phenotype will invariably produce a greater total number of progeny. If this were the driving factor in adaptation, novel advantageous morphs could never increase from their initial low frequencies. Since rare advantageous traits clearly do increase in frequency, the total reproduction of a common form does not imply that natural selection favours that form.

5.6 References

Aarssen LW (2008) Death without sex—the ‘problem of the small’ and selection for reproductive economy in flowering plants. *Evolutionary Ecology* **22**, 279-298

Allard RW, Babbel GR, Clegg MT, Kahler AL (1972) Evidence for coadaptation in *Avena barbata*. *PNAS* **69**, 3043–3048

Bolker BM (2008) Ecological Models and Data in R. Princeton University Press, Princeton, New Jersey.

Chambers J, Aarssen LW (2009) Offspring for the next generation: most are produced by small plants within herbaceous populations. *Evolutionary Ecology* **23**, 737-751

De Jong G, van Noordwijk, AJ (1992) Acquisition and allocation of resources: genetic (co)variances, selection, and life histories. *American Naturalist* **139**, 749–770

Endler JA (1986) Natural Selection in the Wild. Princeton University Press, Princeton, New Jersey.

Fisher RA (1930) The Genetical Theory of Natural Selection. Clarendon Press, Oxford [Complete Variorum edition, Bennett J. H. (Editor), 1999, Oxford University Press, Oxford]

Gardner KM, Latta RG (2006) Identifying loci under selection across contrasting environments in *Avena barbata* using quantitative trait locus mapping. *Molecular Ecology* **15**, 1321-1333

Harper JL (1977) Population biology of plants. Academic Press, London

Lande R (1982) A quantitative genetic theory of life history evolution. *Ecology* **63,** 607-615

Latta RG (2009) Testing for local adaptation in Avena barbata: a classic example of ecotypic divergence. *Molecular Ecology* **18**, 3781-3791

Latta RG, McCain C (2009) Path analysis of natural selection via survival and fecundity across contrasting environments in *Avena barbata*. *Journal of Evolutionary Biology* **22**, 2458-2469

Marshall DR, Jain SK (1969) Interference in pure and mixed populations of *Avena fatua* and *A. barbata*. *Journal of Ecology* **57**, 251-270

Neytcheva MS, Aarssen LW (2008) More plant biomass results in more offspring production in annuals, or does it? *Oikos* **117**, 1298-1307

Orr HA (2009) Fitness and its role in evolutionary genetics. *Nature Reviews Genetics* **10**, 531-539

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Rausher, MD (1992). The measurement of selection on quantitative traits: biases due to environmental covariances between traits and fitness. *Evolution* **46**, 616-626

Reznick D, Nunney L, Tessier A (2000) Big houses, big cars, superfleas and the costs of reproduction. *Trends in Ecology and Evolution* **15**, 421-425

Roff DA (2001) Age and size at maturity. In: Fox CW Roff DA Fairbairn DJ (eds) Evolutionary Ecology: Concepts and Case studies. Oxford University Press, New York, pp 99-112

Smith CC, Fretwell SD (1974) The optimal balance between size and number of offspring. *American Naturalist* **108**, 499-506

Stearns SC (1992) The Evolution of Life Histories. Oxford Univ. Press, New York

Chapter 6 Conclusion

Biological invasions offer a unique opportunity to study contemporary evolution and the process of adaptation. Much of the work in this field has focused on assessing whether hybridization or recombination between multiple introductions of different genetic variants promotes invasiveness in new geographic ranges and environments (Ellstrand & Schierenbeck 2000; Schierenbeck & Ellstrand 2009). Although recombination may be one of the most important ways of generating novel adaptive genetic combinations, successful invasion may also depend on the degree of niche "conservatism" or "overlap" between the native and introduced range (*Guisan et al*. 2014). In my thesis, I evaluated the number of introductions of the highly selfing annual grass (*A. barbata*) to California, characterized the climatic niche in both ranges, and assessed whether there was recombination and selection on two heritable characters. Finally, I presented a study of selection on plant size using data collected from RILs in the field.

Multiple introductions of different genetic variants may help to alleviate the effects of genetic bottlenecks often incurred by small founding populations (Barrett & Colautti 2008). Similar studies have shown that multiple introductions may be especially important for outcrossing species (Chun *et al.* 2009), which may suffer more from inbreeding depression than do selfers, such as *A. barbata* (Husband & Schemske 1996). Ploidy and genomic changes may also produce novel adaptive genetic variants in a new environment and contribute to invasiveness (Mable 2013). In the second chapter of this thesis, I compared the genome size between invaders and native *A. barbata*. I found evidence that there have been multiple introductions of *A. barbata* into California, and that it is possible that recombination between different selfing lineages of smaller genome size could occur in the new range.

Local adaptation and or divergent selection may occur on different geographic scales depending on the scale of the difference in micro - and macro –environments (Yoder *et al.* 2014). In general, a large amount of climatic niche overlap has been observed for plant species occupying native and invaded ranges (Petitpierre *et al.* 2012). When there is niche overlap between invaded and native environments, selective pressures (at least the abiotic ones) are expected to be similar. Thus, adapted genotypes ought to persist in both ranges so long as the founding populations contain the same subset of the genetic variation found in the home range. In my third chapter, I found evidence that *A. barbata* occupies a different set of climatic conditions in its native Iberian range, compared to the climatic conditions that it occupies in California. This result is in contrast to the conventional finding of climatic niche overlap and niche conservatism in plants (Petitpierre *et al.* 2012; Guisan *et al.* 2014). Further, it suggests the possibility that evolution may be occurring in *A. barbata* populations within California.

There have been incredibly few studies documenting contemporary (short-term) evolution by selection in many populations over a large geographic area. Surveys of genotype frequency taken at multiple time points provide the information necessary to evaluate whether evolutionary change has been adaptive or neutral. Typically, datasets containing information on genotype frequency have been confined to small geographic spaces such as islands (Grant & Grant 1993; Milner *et al.* 1999; Grant & Grant 2002), with one notable exception being industrial melanic morphs observed in *Biston betularia* (Kettlewell 1956a; b; Cook 2000; Mathieson & McVean 2013). In my fourth chapter, I coupled datasets from the 1970s to my own and characterized how the frequency of two heritable characters, lemma color and leaf sheath pubescence has changed. I found that the frequency light-colored lemma and pubescence increased. This change was in the direction predicted by a previous common garden experiment (Latta 2009), and occurred statewide.

It is challenging to ascertain whether changes in the frequency of genotypes or morphs have occurred due to neutral versus adaptive processes. In the case of adaptive evolution, it is equally challenging to understand how selection has acted to shape the distribution of phenotype observed in contemporary populations. A classic example of this is with the Californian annual, *Linanthus parryae* where it was originally concluded that the spatial distribution of flower color was the result of random genetic drift, but later with evidence from reciprocal transplant experiments Schemske & Bierzychudek (2007) demonstrated selection for each of the resident ecotypes. Most of the previous observations from Allard's group show strong associations with of the two genotypes to different micro-and macro- environments, yet past (Hamrick & Allard 1975; Jain and Rai

1980) and more recent common garden experiments (Latta 2009) suggest that directional selection was acting to favour one type – the mesic genotype across environments. Both study systems illustrate the fallacy of initially using patterns of population structure as evidence for either neutral or adaptive evolutionary change.

Based on the striking observation that there are many small plants and few large plants it has been suggested in the literature that natural selection acts against larger individuals (Aarssen 2007; Chambers & Aarssen 2008). This is a puzzling suggesting given that there is generally a positive correlation between survival and fecundity and an individual's size (Chambers & Aarssen 2008). Using data from the same common garden field experiment (Latta 2009), and from simulations, I found evidence that genotypes of larger parents produce more seed per capita than smaller parents.

Identifying the genetic and morphological characters that facilitate colonization and invasion in novel environments is important for understanding the spatial distribution of species and the evolutionary processes shaping those distributions. The studies presented in my thesis contribute to our understanding of how genetic variation and ecological factors influence the colonization and expansion of populations into novel habitats. I found evidence that certain genetic variants are spreading in the introduced range and that the invasive populations have changed morphologically over the past 40 years.

References

Aarssen L (2008) Death without sex—the “problem of the small” and selection for reproductive economy in flowering plants. *Evolutionary Ecology*, **22**, 279–298.

Ackerly DD (2009) Evolution, origin and age of lineages in the Californian and Mediterranean floras. *Journal of Biogeography*, **36**, 1221–1233.

Allard RW, Babbel GR, Clegg MT, Kahler AL (1972) Evidence for coadaptation in *Avena barbata*. *PNAS* **69,** 3043–3048

Allard RW, Garcia P, Saenz-de-Miera LE, la Vega de MP (1993) Evolution of multilocus genetic structure in Avena hirtula and Avena barbata. *Genetics*, **135**, 1125–1139.

Arends D, Prins P, Jansen RC, Broman KW (2010) R/qtl: high-throughput multiple QTL mapping. *Bioinformatics* **26**, 2990–2992.

Baker HG (1955) Self-compatibility and establishment after“long-distance”dispersal. *Evolution*, **9**, 347–349.

Baker HG (1967) Support for Baker's law-as a rule. *Evolution*, **21**, 853–856.

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Barrett SCH, Colautti RI (2008) Plant reproductive systems and evolution during biological invasion. *Molecular Ecology*, **17**, 373–383.

Bennett MD, Leitch IJ (2005) Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Annals of Botany* **95**:45–90.

Bolker BM (2008) Ecological Models and Data in R. Princeton University Press, Princeton, New Jersey.

Blumler MA (2000) Spatial analysis to settle an unresolved question in genetics, with both theoretical and applied implications. *Research in Contemporary and Applied Geography: A Discussion Series*, **24**, 1–42.

Broennimann O, Treier UA, Müller-Schärer H *et al.* (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10**, 701–709.

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889-890.

Carroll SP, Hendry AP, Reznick D, Fox CW (2007) Evolution on ecological time-scales. *Functional Ecology* **21**, 387–393.

Carroll SP, Dingle H, Klassen SP. (1997) Genetic differentiation of fitness associated traits among rapidly evolving populations of the soapberry bug. *Evolution*, **51,** 1182–1188.

Chambers J, Aarssen LW (2008) Offspring for the next generation: most are produced by small plants within herbaceous populations. *Evolutionary Ecology*, **23**, 737–751.

Chamberlain S, Boettiger C, Ram K, Barve V, McGlinn D (2014) Package "rgbif" Interface to the Global Biodiversity Information Facility API. R package *version 0.5.0. 1–67.*

Charlesworth D, Wright SI (2001) Breeding systems and genome evolution. *Current Opinion in Genetics & Development*, **11**, 685–690.

Chun YJ, Nason JD, Moloney KA (2009) Comparison of quantitative and molecular genetic variation of native vs. invasive populations of purple loosestrife (Lythrum salicariaL., Lythraceae). *Molecular Ecology*, **18**, 3020–3035.

Clegg MT (1972) Patterns of Genetic Differentiation in Natural Populations of Wild Oats. University of California, Davis. Doctoral Dissertation.

Clegg MT, Allard RW (1972) Patterns of genetic differentiation in the slender wild oat species Avena barbata. *PNAS*, **69**, 1820–1824.

Coffman FA (1964) Inheritance of morphologic characters in Avena. 1308. US Dept. of Agriculture.

Colautti RI, Ricciardi A, Grigorovich IA, MacIsaac HJ (2004) Is invasion success explained by the enemy release hypothesis? *Ecology Letters*, **7**, 721–733.

Colautti RI, Eckert CG, Barrett SCH (2010) Evolutionary constraints on adaptive evolution during range expansion in an invasive plant. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1799–1806.

Cook L (2000) Changing views on melanic moths. *Biological Journal of the Linnean Society*, **69**, 431–441.

Cousyn C, DeMeester L, Colbourne JK, Brendonck L, Verschuren D, Volckaert F (2001) Rapid, local adaptation of zooplankton behavior to changes in predation pressure in the absence of neutral genetic changes. *PNAS* **98**, 6256–6260.

Crosby K, Latta RG (2013) A test of the reproductive economy hypothesis in plants: more offspring per capita come from large (not small) parents in Avena barbata. *Evolutionary Ecology*, **27**, 193–203.

D'Antonio CM, Vitousek PM (1992) Biological invasions by exotic grasses, the grass/fire cycle, and global change. *Annual Review of Ecology and Systematics*, **23**, 63–87.

Daehler C, Strong D (1997) Hybridization between introduced smooth cordgrass (Spartina alterniflora; Poaceae) and native California cordgrass (S. foliosa) in San Francisco Bay, California, USA. *American Journal of Botany*, **84**, 607–611.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic markerdiscovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499–510.

De Jong G, van Noordwijk, AJ (1992) Acquisition and allocation of resources: genetic (co)variances, selection, and life histories. *American Naturalist* **139**, 749–770

Dlugosch K, Parker M (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology*, **17**, 431–449.

Dobzhansky T, Levene H (1948) Genetics of natural populations. XVII. Proof of operation of natural selection in wild populations of Drosophila pseudoobscura. *Genetics* **33**, 537-547.

Doležel J, Číhalíková J, Lucretti S (1992) A high-yield procedure for isolation of metaphase chromosomes from root tips of Vicia faba L. *Planta*, **188**, 93–98.

Doležel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, **2**, 2233–2244.

Dubuis A, Giovanettina S, Pellissier L, Pottier J, Vittoz P, Guisan A (2012) Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. *Journal of Vegetation Science*, **24**, 593–606.

Ellstrand N, Schierenbeck K (2000) Hybridization as a stimulus for the evolution of invasiveness in plants? *PNAS*, **97**, 7043–7050.

Emery N.C. & Ackerly D.D. (2014) Ecological release exposes genetically based niche variation. *Ecology Letters*, **17**, 1149–1157.

Endler JA (1986) Natural Selection in the Wild. Princeton University Press, Princeton, New Jersey.

Fisher RA (1930) The Genetical Theory of Natural Selection. Clarendon Press, Oxford [Complete Variorum edition, Bennett J. H. (Editor), 1999, Oxford University Press, Oxford]

Fitzpatrick MC, Weltzin JF, Sanders NJ, Dunn RR. (2007) The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Global Ecology and Biogeography*, **16**, 24-33

Franks SJ, Sim S, Weis AE (2007) Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *PNAS*, **104**, 1278–1282.

Garcia P, Vences F, Vega M (1989) Allelic and Genotypic Composition of Ancestral Spanish and Colonial Californian Gene Pools of *Avena barbata*: Evolutionary Implications. *Genetics*, **122**, 687–694.

Gardner KM, Latta RG (2006) Identifying loci under selection across contrasting environments in Avena barbata using quantitative trait locus mapping. *Molecular Ecology*, **15**, 1321–1333.

Gillingham PK, Huntley B, Kunin WE, Thomas CD (2012) The effect of spatial resolution on projected responses to climate warming. *Diversity and Distributions*, **18**, 990–1000.

Godsoe W, Strand E, Smith CI, Yoder JB, Esque TC, Pellmyr O. (2009) Divergence in an obligate mutualism is not explained by divergent climatic factors. *New Phytologist*, **183**, 589–599.

Golding N. (2013) Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications. 1–247. University of Oxford, UK. Doctoral Dissertation.

Grant BR, Grant PR (1993) Evolution of Darwin's Finches Caused by a Rare Climatic Event. *Proceedings of the Royal Society B: Biological Sciences*, **251**, 111–117.

Grant P, Grant B (2002) Unpredictable Evolution in a 30-Year Study of Darwin's Finches. *Science*, **296**, 707–711.

Guisan A, Petitpierre B, Broennimann O, Daehler C, Kueffer C (2014) Unifying niche shift studies: insights from biological invasions. *Trends in Ecology & Evolution*, **29**, 260–269.

Hamrick JL, Allard RW (1972) Microgeographical Variation in Allozyme Frequencies in *Avena barbata*. *PNAS*, **69**, 2100–2104.

Hamrick JL, Allard RW (1975) Correlations between quantitative characters and enzyme genotypes in *Avena barbata. Evolution*, **29**, 438–442.

Harper JL (1977) Population biology of plants. Academic Press, London

Harwood TD, Mokany K, Paini DR (2014) Microclimate is integral to the modeling of plant responses to macroclimate. *PNAS*, **111**, E1164–E1165.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hutchinson ES (1982) Genetic Markers and Ecotypic Differentiation of *Avena barbata* Pott ex Link. University of California, Davis. Doctoral Dissertation.

Hutchinson G (1959) Homage to Santa Rosalia or why are there so many kinds of animals? *American Naturalist*, **93**, 145–159.

Husband BC, Schemske DW (1996) Evolution of the magnitude and timing of inbreeding depression in plants. *Evolution*, **50**, 54–70.

ISRIC – World Soil Information, 2014. SoilGrids: an automated system for global soil mapping. Available for download at http://soilgrids1km.isric.org.

Jackson LE (1985) Ecological origins of California's Mediterranean grasses. *Journal of Biogeography*, **12**, 349–361.

Jain SK, Rai KN (1980) Population biology of Avena. VIII. Colonization experiment as a test of the role of natural selection in population divergence. *American Journal of Botany*, **67**, 1342–1346.

Johnson MTJ, Dinnage R, Zhou AY, Hunter MD. (2008) Environmental variation has stronger effects than plant genotype on competition among plant species. *Journal of Ecology* **96,** 947–955.

Johansen-Morris AD, Latta RG (2006) Fitness consequences of hybridization between ecotypes of *Avena barbata*: hybrid breakdown, hybrid vigor, and transgressive segregation. *Evolution* **60**, 1585–1595.

Johansen-Morris AD, Latta RG (2008). Genotype by environment interactions for fitness in hybrid genotypes of *Avena barbata*. *Evolution* **62**, 573–585.

Kettlewell H (1956a) A Resume of Investigations on the Evolution of Melanism in the Lepidoptera. *Proceedings of the Royal Society of London Series B*, **145**, 297–303.

Kettlewell H (1956b) Further selection experiments on industrial melanism in the Lepidoptera. *Heredity*, **10**, 287–301.

Kingsolver J, Diamond S (2011) Phenotypic selection in natural populations: what limits directional selection? *American Naturalist,* **177,** 346–357.

Kingsolver J, Hoekstra H, Hoekstra J, Berrigan D, Vignieri D, Hill C, Hoang A, Gibert P, Beerli P (2001) The strength of phenotypic selection in natural populations. *American Naturalist,* **157,** 245–261.

Lande R (1982) A quantitative genetic theory of life history evolution. *Ecology* **63**, 607-615

Latta RG, MacKenzie J, Vats A, Schoen D (2004) Divergence and variation of quantitative traits between allozyme genotypes of *Avena barbata* from contrasting habitats. *Journal of Ecology* **92**, 51–71.

Latta RG (2009) Testing for local adaptation in Avena barbata: a classic example of ecotypic divergence. *Molecular Ecology*, **18**, 3781–3791.

Latta RG, McCain C (2009) Path analysis of natural selection via survival and fecundity across contrasting environments in *Avena barbata*. *Journal of Evolutionary Biology* **22**, 2458-2469

Latta RG, Gardner KM, Staples DA (2010) Quantitative trait locus mapping of genes under selection across multiple years and sites in Avena barbata: epistasis, pleiotropy, and genotype-by-environment interactions. *Genetics* **185**, 375–385.

Levin DA (2009) Flowering-time plasticity facilitates niche shifts in adjacent populations. *New Phytologist*, **183**, 661–666.

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Mable BK (2013) Polyploids and hybrids in changing environments: winners or losers in the struggle for adaptation? *Heredity*, **110**, 95–96.

Marshall DR, Jain SK (1969) Interference in pure and mixed populations of *Avena fatua* and *A. barbata*. Journal of Ecology 57: 251-270

Mathieson I, McVean G (2013) Estimating Selection Coefficients in Spatially Structured Populations from Time Series Data of Allele Frequencies. *Genetics,* **193**, 973-984.

McCue K, Holtsford T (1998) Seed bank influences on genetic diversity in the rare annual Clarkia springvillensis (Onagraceae). *Am. J. Bot.* **85**, 30-36

Miller, R. D. 1977. Genetic Variability in the Slender Wild Oat Avena barbata in California. University of California, Davis. Doctoral Dissertation.

Milner JM, Albon SD, Illius AW, Pemberton JM, Clutton-Brock TH (1999) Repeated selection of morphometric traits in the Soay sheep on St Kilda. *Journal of Animal Ecology*, **68**, 472–488.

Minnich R. (2008) *California's Fading Wildflowers.* University of California Press, Berkeley and Los Angeles.

Neytcheva M, Aarssen L (2008) More plant biomass results in more offspring production in annuals, or does it? *Oikos*, **117**, 1298-1307.

Orr HA (2009) Fitness and its role in evolutionary genetics. *Nature Reviews Genetics* **10**, 531-539.

Pearman PB, Randin CF, Broennimann O, Vittoz P, Knaap WOVD, Engler R, Lay GL, Zimmermann NE, Guisan A (2008) Prediction of plant species distributions across six millennia. *Ecology Letters*, **11**, 357–369.

Pérez de la Vega M, Garcia P, Allard RW (1991) Multilocus genetic structure of ancestral Spanish and colonial Californian populations of Avena barbata. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 1202–1206.

Peterson AT (2011) Ecological niche conservatism: a time-structured review of evidence. *Journal of Biogeography*, **38**, 817–827.

Petitpierre B, Kueffer C, Broennimann O *et al.* (2012) Climatic niche shifts are rare among terrestrial plant invaders. *Science*, **335**, 1344–1348.

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

Pinero D. (1982) *Correlations between enzyme phenotypes and physical environment in California populations of Avena barbata and Avena fatua.* University of California, Davis. Doctoral Dissertation.

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE* **7**, e32253.

Prentis PJ, Wilson JRU, Dormontt EE, Richardson DM, Lowe AJ (2008) Adaptive evolution in invasive species. *Trends in Plant Science*, **13**, 288–294.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Rausher, MD (1992). The measurement of selection on quantitative traits: biases due to environmental covariances between traits and fitness. *Evolution* **46**, 616-626

Reznick D, Nunney L, Tessier A (2000) Big houses, big cars, superfleas and the costs of reproduction. Trends in Ecology and Evolution 15: 421-425

Reznick D, Ghalambor CK (2001) The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. *Genetica* **112-113**,183–198.

Roff DA (2001) Age and size at maturity. In: Fox CW Roff DA Fairbairn DJ (eds) Evolutionary Ecology: Concepts and Case studies. Oxford University Press, New York, pp 99-112

Rollinson N, Hutchings JA (2013) The relationship between offspring size and fitness: integrating theory and empiricism. *Ecology* **94**, 315–324.

Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics*, **8**, 552–560.

Schemske DW, Bierzychudek P (2007) Spatial differentiation for flower color in the desert annual *Linanthus parryae*: was Wright right? *Evolution*, **61**, 2528–2543.

Schierenbeck KA, Ellstrand N (2009) Hybridization and the evolution of invasiveness in plants and other organisms. *Biological Invasions*, **11**, 1093–1105.

Sherrard ME, Maherali H, Latta RG (2009) Water stress alters the genetic architecture of functional traits associated with drought adaptation in *Avena barbata*. *Evolution*, **63**, 702–715.

Smith CC, Fretwell SD (1974) The optimal balance between size and number of offspring. *American Naturalist* **108**, 499-506

Stanton, ML (1984) Seed variation in wild radish: effect of seed size on components of seedling and adult fitness. *Ecology* **65**, 1105–1112.

Stearns SC (1992) The Evolution of Life Histories. Oxford Univ. Press, New York

Václavík T, Meentemeyer RK (2011) Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. *Diversity and Distributions*, **18**, 73–83.

Yoder JB, Stanton-Geddes J, Zhou P *et al.* (2014) Genomic signature of adaptation to climate in Medicago truncatula. *Genetics*, **196**, 1263–1275.

Appendix 1a. Site coordinates from 2010, trait frequencies observed at a site, and estimates of morphotype combinations.

Code	Lat	Lon	SS	Pub 2010	Glab 2010	Dark 2010	Light 2010	Glab & Dark 2010	Pub & Light 2010	Pub & Dark 2010	Light & Glab 2010
ARC	40.903	124.072	10	1.000	0.000	0.455	0.545	0.000	0.545	0.455	0.000
BEL	37.013	121.348	12	1.000	0.000	0.045	0.955	0.000	0.955	0.045	0.000
BER	38.499	122.121	48	0.063	0.938	0.957	0.043	0.898	0.003	0.060	0.040
BLF	40.421	122.191	17	1.000	0.000	0.176	0.824	0.000	0.824	0.176	0.000
BOD	38.32	123.035	58	1.000	0.000	0.259	0.741	0.000	0.741	0.259	0.000
BSR	36.214	121.747	12	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
CAL	35.281	118.624	11	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
CAS	34.507	118.603	8	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
CDR	38.578	122.529	49	0.980	0.020	0.571	0.429	0.012	0.420	0.560	0.009
CHI	38.501	122.353	49	1.000	0.000	0.333	0.667	0.000	0.667	0.333	0.000
CHO	35.711	120.31	15	0.867	0.133	0.500	0.500	0.067	0.433	0.433	0.067
CHP	38.585	122.57	49	1.000	0.000	0.143	0.857	0.000	0.857	0.143	0.000
CLR	39.042	122.341	8	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
COA	36.101	120.416	40	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
COM	39.269	123.643	41	1.000	0.000	0.250	0.750	0.000	0.750	0.250	0.000
CUY	35.07	119.99	8	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
DAL	40.312	122.007	24	1.000	0.000	0.042	0.958	0.000	0.958	0.042	0.000
DLO	40.783	123.335	50	1.000	0.000	0.220	0.780	0.000	0.780	0.220	0.000
DMR	32.931	117.237	9	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
EDS	34.941	118.925	7	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
ELK	39.61	122.532	14	0.857	0.143	0.929	0.071	0.133	0.061	0.796	0.010
ETO	33.66	117.657	18	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
FID	38.487	120.805	50	1.000	0.000	0.560	0.440	0.000	0.440	0.560	0.000
FOR	40.6	124.17	9	1.000	0.000	0.778	0.222	0.000	0.222	0.778	0.000
GEY	38.712	122.882	75	1.000	0.000	0.620	0.380	0.000	0.380	0.620	0.000
GOR	34.794	118.842	10	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000

Code	Lat	Lon	SS	Pub 2010	Glab 2010	Dark 2010	Light 2010	Glab & Dark 2010	Pub & Light 2010	Pub & Dark 2010	Light & Glab 2010
GUA	38.785	123.554	10	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000
HIX	38.131	122.713	12	1.000	0.000	0.385	0.615	0.000	0.615	0.385	0.000
HKR	40.293	122.277	11	1.000	0.000	0.545	0.455	0.000	0.455	0.545	0.000
HOP	39.007	123.08	100	1.000	0.000	0.690	0.310	0.000	0.310	0.690	0.000
ING	40.756	122.026	10	1.000	0.000	0.100	0.900	0.000	0.900	0.100	0.000
ISB	35.643	118.463	10	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
JAK	38.36	120.745	16	0.625	0.375	0.375	0.625	0.141	0.391	0.234	0.234
JCT	40.725	123.05	16	1.000	0.000	0.063	0.938	0.000	0.938	0.063	0.000
JEN	38.498	123.208	37	1.000	0.000	0.135	0.865	0.000	0.865	0.135	0.000
JOL	35.963	121.185	14	0.429	0.571	0.857	0.143	0.490	0.061	0.367	0.082
KRN	35.467	118.755	9	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
LAL	34.78	120.315	50	0.500	0.500	0.930	0.070	0.465	0.035	0.465	0.035
LCS	34.41	119.366	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
LHD	40.888	122.384	10	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000
LIK	37.329	121.681	86	1.000	0.000	0.372	0.628	0.000	0.628	0.372	0.000
LMN	36.406	119.055	11	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
LUC	35.999	121.47	27	0.111	0.889	0.920	0.080	0.818	0.009	0.102	0.071
LYT	34.261	117.499	14	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
MBU	34.041	118.892	40	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
MEN	39.24	123.181	11	1.000	0.000	0.091	0.909	0.000	0.909	0.091	0.000
MIL	37.028	119.702	10	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
MOS	37.528	122.513	17	1.000	0.000	0.647	0.353	0.000	0.353	0.647	0.000
MPS	37.46	119.943	11	1.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
MQR	35.555	120.888	11	0.636	0.364	0.727	0.273	0.264	0.174	0.463	0.099
MSH	38.18	122.909	40	1.000	0.000	0.214	0.786	0.000	0.786	0.214	0.000
NAC	35.998	121.383	16	0.875	0.125	0.250	0.750	0.031	0.656	0.219	0.094
NAP	38.209	122.187	8	1.000	0.000	0.125	0.875	0.000	0.875	0.125	0.000

Code	Lat	Lon	SS	Pub 2010	Glab 2010	Dark 2010	Light 2010	Glab & Dark 2010	Pub & Light 2010	Pub & Dark 2010	Light & Glab 2010
NAV	39.111	123.507	10	1.000	0.000	0.600	0.400	0.000	0.400	0.600	0.000
OAK	40.652	122.599	20	1.000	0.000	0.100	0.900	0.000	0.900	0.100	0.000
OMD	32.841	117.043	33	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
ORO	39.518	121.513	21	1.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
PAI	36.681	121.259	16	1.000	0.000	0.125	0.875	0.000	0.875	0.125	0.000
PAS	39.851	122.615	19	0.947	0.053	1.000	0.000	0.053	0.000	0.947	0.000
PAT	37.457	121.191	14	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
PFR	36.816	119.385	10	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
PIN	36.488	121.153	13	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
PJC	40.681	122.352	7	1.000	0.000	0.333	0.667	0.000	0.667	0.333	0.000
POR	36.026	118.922	11	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
PRB	35.589	120.7	9	0.667	0.333	0.600	0.400	0.200	0.267	0.400	0.133
PST	36.19	120.706	9	1.000	0.000	0.111	0.889	0.000	0.889	0.111	0.000
PYN	40.338	121.914	10	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000
RBF	40.215	122.178	23	1.000	0.000	0.739	0.261	0.000	0.261	0.739	0.000
RED	40.613	122.355	50	1.000	0.000	0.250	0.750	0.000	0.750	0.250	0.000
REF	34.563	120.091	9	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
REY	38.081	122.96	6	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000
RIC	40.032	123.784	20	0.950	0.050	0.333	0.667	0.017	0.634	0.316	0.033
SAD	39.183	123.754	16	1.000	0.000	0.750	0.250	0.000	0.250	0.750	0.000
SAN	38.218	120.69	14	0.286	0.714	1.000	0.000	0.714	0.000	0.286	0.000
SAR	36.012	120.924	40	0.725	0.275	0.395	0.605	0.109	0.439	0.286	0.166
SAU	37.836	122.488	10	0.600	0.400	0.688	0.313	0.275	0.188	0.413	0.125
SBA	34.462	119.771	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
SBR	34.184	117.329	6	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
SCL	33.428	117.61	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
SFH	39.234	121.295	58	1.000	0.000	0.086	0.914	0.000	0.914	0.086	0.000

Code	Lat	Lon	SS	Pub 2010	Glab 2010	Dark 2010	Light 2010	Glab & Dark 2010	Pub & Light 2010	Pub & Dark 2010	Light & Glab 2010
SFN	34.358	118.555	40	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
SGB	34.159	117.909	15	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
SIL	38.345	122.283	28	1.000	0.000	0.321	0.679	0.000	0.679	0.321	0.000
SJB	36.853	121.569	10	1.000	0.000	0.273	0.727	0.000	0.727	0.273	0.000
SLD	37.097	121.121	8	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
SLR	33.278	117.222	46	0.130	0.870	0.870	0.130	0.756	0.017	0.113	0.113
SMO	34.051	118.53	3	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
SMR	41.844	124.018	10	1.000	0.000	0.563	0.438	0.000	0.438	0.563	0.000
SMT	37.494	122.371	10	0.700	0.300	0.800	0.200	0.240	0.140	0.560	0.060
SMV	39.208	121.256	17	0.294	0.706	0.941	0.059	0.664	0.017	0.277	0.042
SNE	37.517	120.448	15	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
SNL	37.601	121.87	22	1.000	0.000	0.591	0.409	0.000	0.409	0.591	0.000
SON	38.237	122.514	11	1.000	0.000	0.091	0.909	0.000	0.909	0.091	0.000
SPR	36.572	121.741	6	1.000	0.000	0.429	0.571	0.000	0.571	0.429	0.000
SSM	35.646	121.192	14	0.714	0.286	0.286	0.714	0.082	0.510	0.204	0.204
STO	39.659	122.526	50	0.500	0.500	1.000	0.000	0.500	0.000	0.500	0.000
SUN	33.697	117.177	8	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000
UKH	39.111	123.197	10	1.000	0.000	0.100	0.900	0.000	0.900	0.100	0.000
WEO	40.317	123.917		NA	NA	0.500	0.500	NA	NA	NA	NA
WLT	39.521	123.393	18	0.667	0.333	0.000	1.000	0.000	0.667	0.000	0.333
WOD	35.714	118.86	13	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000

Appendix 1b. Site codes, coordinates from 1970s, mean trait frequencies observed at a site, and estimates of morphotype combinations. Dist = inferred distance to collection site in 2010, By = M, C, or H (Miller, Clegg, or Hutchinson).

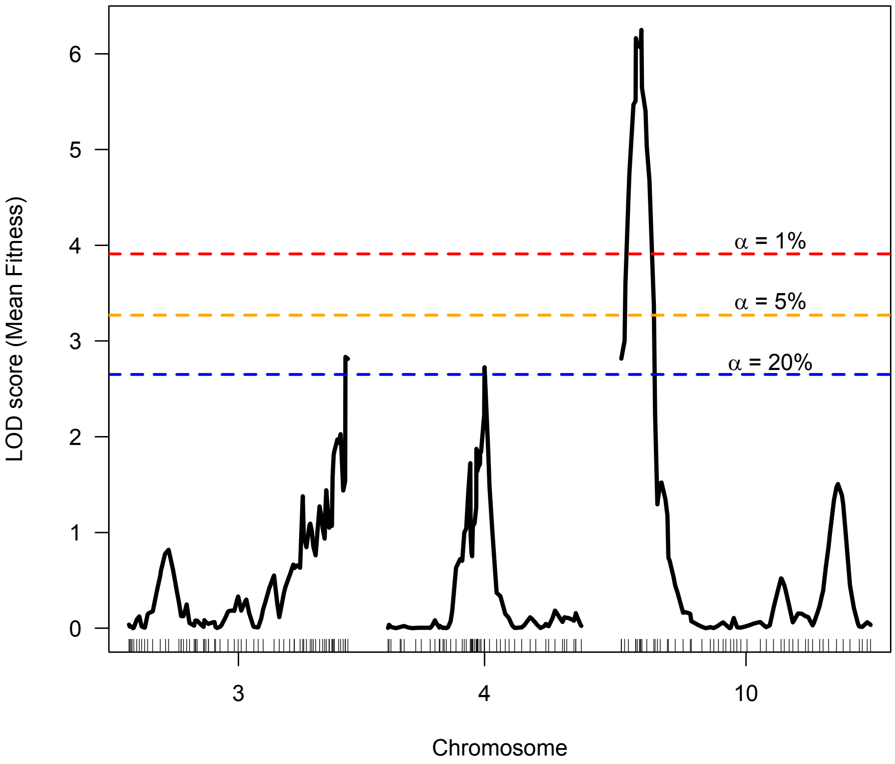
Code	Date	Lat	Long	Dist	Inferred Sample Size	Pubescent 1970	Glabrous 1970	Dark 1970	Light 1970	Glab & Dark 1970	Pub & Light 1970	Pub & Dark 1970	Light & Glab 1970	By
ARC	1972	40.883	124.083	2.4	5	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	M
BEL	1972	37.017	121.35	0.48	55	0.854	0.146	0.739	0.261	0.108	0.223	0.631	0.038	M
BER	1970	38.5	122.12		75	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	C
BLF	1972	40.417	122.183	0.81	10	0.900	0.100	0.000	1.000	0.000	0.900	0.000	0.100	M
BOD	1972	38.333	123.05	1.9	32	0.312	0.688	0.810	0.190	0.557	0.059	0.253	0.131	CM
BSR	1972	36.217	121.75	0.43	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
CAL	1972	35.3	118.6	3	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
CAS	1972	34.533	118.633	4	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
CDR	1977	38.575	122.533	0.48	20	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	H
CHI	1972	38.5	122.35	0.28	20	0.857	0.143	0.160	0.840	0.023	0.720	0.137	0.120	HM
CHO	1972	35.717	120.317	0.9	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
CHP	1977	38.583	122.567	0.34	20	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	H
CLR	1972	39.05	122.333	1.1	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
COA	1972	36.133	120.367	5.7	47	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	CM
COM	1977	39.283	123.65	1.7	20	1.000	0.000	0.018	0.982	0.000	0.982	0.018	0.000	H
CUY	1972	35.083	119.983	1.6	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
DAL	1972	40.317	122	0.81	33	0.830	0.170	0.394	0.606	0.067	0.503	0.327	0.103	M
DLO	1972	40.783	123.333	0.17	20	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	HM
DMR	1972	32.933	117.233	0.43	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
EDS	1972	34.933	118.85	6.9	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
ELK	1972	39.6	122.533	1.12	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
ETO	1970	33.645	117.655		81	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	C
FID	1970	38.5	120.798		62	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	C

Code	Date	Lat	Long	Dist	Inferred Sample Size	Pubescent 1970	Glabrous 1970	Dark 1970	Light 1970	Glab & Dark 1970	Pub & Light 1970	Pub & Dark 1970	Light & Glab 1970	By
FOR	1972	40.6	124.167	0.25	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
GEY	1972	38.683	122.833	5.3	60	0.615	0.385	0.200	0.800	0.077	0.492	0.123	0.308	CHM
GOR	1972	34.8	118.85	1	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
GUA	1972	38.783	123.533	1.8	28	0.689	0.311	0.000	1.000	0.000	0.689	0.000	0.311	M
HIX	1972	38.133	122.717	0.42	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
HKR	1972	40.283	122.267	1.4	10	0.700	0.300	0.577	0.423	0.173	0.296	0.404	0.127	M
HOP	1977	39.008	123.083	0.28	60	0.664	0.336	0.860	0.140	0.289	0.093	0.571	0.047	H
ING	1972	40.75	122.017	1	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
ISB	1972	35.717	118.483	8.4	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
JAK	1972	38.367	120.75	0.89	5	0.000	1.000	0.907	0.093	0.907	0.000	0.000	0.093	M
JCT	1972	40.733	123.05	0.89	176	0.556	0.444	1.000	0.000	0.444	0.000	0.556	0.000	M
JEN	1972	38.483	123.2	1.8	29	0.979	0.021	0.180	0.820	0.004	0.803	0.176	0.017	HM
JOL	1972	35.967	121.2	1.4	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
KRN	1972	35.45	118.75	1.9	36	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
LAL	1970	34.757	120.341		59	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	C
LCS	1972	34.4	119.367	1.1	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
LHD	1972	40.883	122.383	0.56	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
LIK	1970	37.38	121.654		50	0.060	0.940	0.035	0.965	0.033	0.058	0.002	0.907	C
LMN	1972	36.4	119.05	0.8	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
LUC	1972	36	121.483	1.2	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
LYT	1972	34.25	117.483	1.9	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
MBU	1972	34.05	118.9	1.2	50	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	CM
MEN	1972	39.233	123.183	0.8	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
MIL	1972	37.017	119.7	1.2	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
MOS	1972	37.533	122.5	1.3	50	0.300	0.700	0.853	0.147	0.597	0.044	0.256	0.103	M
MPS	1972	37.467	119.95	1	5	0.000	1.000	0.915	0.085	0.915	0.000	0.000	0.085	M

Code	Date	Lat	Long	Dist	Inferred Sample Size	Pubescent 1970	Glabrous 1970	Dark 1970	Light 1970	Glab & Dark 1970	Pub & Light 1970	Pub & Dark 1970	Light & Glab 1970	By
MQR	1972	35.55	120.883	0.7	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
MSH	1977	38.175	122.9	0.96	60	0.966	0.034	0.050	0.950	0.002	0.918	0.048	0.032	H
NAC	1972	36	121.367	1.5	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
NAP	1972	38.217	122.2	1.4	10	0.308	0.692	1.000	0.000	0.692	0.000	0.308	0.000	M
NAV	1972	39.117	123.517	1.1	8	0.883	0.117	1.000	0.000	0.117	0.000	0.883	0.000	M
OAK	1972	40.633	122.567	3.4	36	0.022	0.978	0.968	0.032	0.947	0.001	0.021	0.031	M
OMD	1977	32.842	117.042	0.14	20	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	H
ORO	1972	39.517	121.517	0.36	55	0.986	0.014	1.000	0.000	0.014	0.000	0.986	0.000	M
PAI	1972	36.683	121.267	0.75	35	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	M
PAS	1972	39.85	122.617	0.2	5	0.000	1.000	0.958	0.042	0.958	0.000	0.000	0.042	M
PAT	1972	37.45	121.183	1.1	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
PFR	1972	36.817	119.383	0.21	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
PIN	1972	36.467	121.15	2.4	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
PJC	1972	40.683	122.35	0.28	5	0.000	1.000	0.956	0.044	0.956	0.000	0.000	0.044	M
POR	1972	36.033	118.833	8	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
PRB	1972	35.6	120.7	1.2	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
PST	1972	36.15	120.7	4.5	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
PYN	1972	40.333	121.9	1.3	5	1.000	0.000	0.044	0.956	0.000	0.956	0.044	0.000	M
RBF	1972	40.2	122.183	1.72	35	0.413	0.587	0.800	0.200	0.470	0.083	0.330	0.117	M
RED	1970	40.615	122.347		71	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	C
REF	1972	34.55	120.067	2.6	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
REY	1972	38.083	122.967	0.65	15	0.538	0.462	0.190	0.810	0.088	0.436	0.102	0.374	M
RIC	1972	40.033	123.783	0.14	26	0.867	0.133	0.808	0.192	0.107	0.166	0.701	0.026	M
SAD	1972	39.183	123.75	0.35	29	0.886	0.114	0.104	0.896	0.012	0.794	0.092	0.102	M
SAN	1972	38.2	120.667	2.8	5	0.000	1.000	0.866	0.134	0.866	0.000	0.000	0.134	M
SAR	1972	36.017	120.9	2.2	20	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	HM

Code	Date	Lat	Long	Dist	Inferred Sample Size	Pubescent 1970	Glabrous 1970	Dark 1970	Light 1970	Glab & Dark 1970	Pub & Light 1970	Pub & Dark 1970	Light & Glab 1970	By
SAU	1972	37.833	122.483	0.55	6	0.333	0.667	0.909	0.091	0.606	0.030	0.303	0.061	M
SBA	1972	34.467	119.767	0.67	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SBR	1972	34.15	117.333	3.8	36	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SCL	1972	33.45	117.617	2.5	5	0.000	1.000	0.981	0.019	0.981	0.000	0.000	0.019	M
SFH	1977	39.25	121.292	1.8	20	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	H
SFN	1970	34.373	118.561		84	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	C
SGB	1972	34.167	117.9	1.2	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SIL	1972	38.333	122.283	1.3	35	0.743	0.257	0.114	0.886	0.029	0.658	0.085	0.228	M
SJB	1972	36.85	121.567	0.38	37	1.000	0.000	0.095	0.905	0.000	0.905	0.095	0.000	M
SLD	1972	37.083	121.117	1.6	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SLR	1970	33.264	117.234		68	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	C
SMO	1972	34.067	118.533	1.8	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SMR	1972	41.85	124.033	1.4	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SMT	1972	37.483	122.367	1.3	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SMV	1972	39.217	121.25	1.1	5	0.000	1.000	0.855	0.145	0.855	0.000	0.000	0.145	M
SNE	1972	37.517	120.45	0.18	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SNL	1972	37.6	121.867	0.29	35	1.000	0.000	0.481	0.519	0.000	0.519	0.481	0.000	M
SON	1972	38.233	122.5	1.3	10	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
SPR	1972	36.567	121.733	0.9	33	0.962	0.038	0.129	0.871	0.005	0.838	0.124	0.033	M
SSM	1972	35.65	121.2	0.8	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M
STO	1970	39.662	122.526		66	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	C
SUN	1977	33.708	117.183	1.3	20	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	H
UKH	1972	39.117	123.2	0.72	5	1.000	0.000	0.007	0.993	0.000	0.993	0.007	0.000	M
WEO	1972	40.317	123.917	0.65	10	0.500	0.500	0.120	0.880	0.060	0.440	0.060	0.440	M
WLT	1972	39.517	123.4	0.75	31	0.978	0.022	0.194	0.806	0.004	0.788	0.190	0.018	M
WOD	1972	35.683	118.883	4	5	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	M

Appendix 2



Appendix 2. LOD plot of linkage group associated with mean fitness in the field (LG 3, 4, and 10) in the RILs carried out using CIM based on a subset of RILs (n = 94 for which there is sufficient data). Hatched blue, orange, and red lines are cut-offs for α-values = 0.2 (LOD = 2.60), 0.05 (LOD = 3.31), and 0.01 (LOD = 4.26) thresholds based on 1000 permutation tests.

Appendix 3. R script used to produce expected distributions of mean frequency of traits under drift.

Filename:simulation_two_genotypes_epistatic_basic.R

Kate Crosby

Sat Jun 28 10:10:23 2014

```
rm(list=ls())

library(doParallel)

## Loading required package: foreach
## Loading required package: iterators
## Loading required package: parallel

detectCores()

## [1] 4

cl <- makeCluster(4)  # Use 4 cores
registerDoParallel(cl)

# Run 1000 simulations,
# on a popsize of 1000 each simulation for 40 generations
totalSims <- 1000
popsize <- 1000
totalGen <- 40

# Define the trait genotypes as homozygotes
Dark <- c(1,1,1,1)
Light <- c(0,0,0,0)

# Specify outcrossing rate
outcrossingrate <- 0.02


# Outputs defined below
Output = matrix(0,totalGen,10)
OutputSims = matrix (0, totalSims, 11)
# Start simulation loop
for(isim in 1:totalSims)
    {
        # Keep track of pseudo-random seeds
        seeder<-round(runif(min=2, max = 80E4, n=1),2)
        set.seed(seeder)

        # Start close to the observed frequencies of trait
        Light <- matrix(data = Light, nrow = popsize*.5, ncol = 4, byrow = T)
        Dark <- matrix(data = Dark, nrow = popsize*.5, ncol = 4, byrow = T)


# Make the initial generation parents array of the four types
```

```
        parents <- rbind(Dark, Light)

  #Start the generation loop
          for(igen in 1:totalGen)
          {
                moms <- parents[sample(nrow(parents), popsize, replace=T),]
                num.outcrossers <- rbinom(1,popsize,outcrossingrate)
                selfers <- popsize - num.outcrossers

                Output[igen,1] <- igen
                Output[igen,2] <- popsize
                Output[igen,3] <- num.outcrossers
                Output[igen,4] <- selfers

                sex <- sample(1:popsize, num.outcrossers)
                selfing <- setdiff(1:popsize, sex)

                # Identify those rows
                sex.id <- moms[sex,]
                self.id <- moms[selfing,]

# Start with outcrossing MOTHERS (i.e. NOT POLLEN), i.e. the maternal allele at each locus
                locus1.color.maternal <- sex.id[,1:2]
                locus2.color.maternal <- sex.id[,3:4]

                        mom.sex.locus1.color.allele1 <- NULL
                        for(i in 1:nrow(locus1.color.maternal))
                            {
                            mom.sex.locus1.color.allele1[i] = sample(locus1.color.maternal[i,],1)
                            }

                        mom.sex.locus2.color.allele2 <- NULL
                        for(i in 1:nrow(locus2.color.maternal))
                            {
                            mom.sex.locus2.color.allele2[i] = sample(locus2.color.maternal[i,],1)
                            }

# Pollen
              dads <- sample(1:popsize,num.outcrossers)
              dads.id <- parents[dads,]

                locus1.color.pollen <- dads.id[,1:2]
                locus2.color.pollen <- dads.id[,3:4]

                    dad.sex.color.locus1allele1 <- NULL
                    for(i in 1:nrow(locus1.color.pollen))
                            {
                                dad.sex.color.locus1allele1[i] = sample(locus1.color.pollen[i,],1)
                            }

                    dad.sex.color.locus2allele2 <- NULL
                    for(i in 1:nrow(locus2.color.pollen))
                            {
                                dad.sex.color.locus2allele2[i] = sample(locus2.color.pollen[i,],1)
                            }
# Then cbind the alleles for locus 1 and locus 2 together making a new dataframe

                new.outcrossed.progeny <- data.frame(cbind( mom.sex.locus1.color.allele1,
                    dad.sex.color.locus1allele1, mom.sex.locus2.color.allele2,
                    dad.sex.color.locus2allele2))
```

```
# For the selfers - use "self.id" array, and define each locus
            locus1.color.selfer <- self.id[,1:2]
            locus1.color.selfer <- as.matrix(locus1.color.selfer)
            locus2.color.selfer <- self.id[,3:4]
            locus2.color.selfer <- as.matrix(locus2.color.selfer)

# Choose alleles

            locus1.color.allele1 <- NULL
                for(i in 1:nrow(locus1.color.selfer))
                  {
                  locus1.color.allele1[i] <- sample(locus1.color.selfer[i,],1,replace =T)
                  }

            locus1.color.allele2 <- NULL
                 for(i in 1:nrow(locus1.color.selfer))
                   {
                   locus1.color.allele2[i] <- sample(locus1.color.selfer[i],1,replace =T)
                   }

            locus2.color.allele1 <- NULL
                 for(i in 1:nrow(locus2.color.selfer))
                   {
                   locus2.color.allele1[i]  <- sample(locus2.color.selfer[i,],1, replace =T)
                   }

             locus2.color.allele2 <- NULL
                 for(i in 1:nrow(locus2.color.selfer))
                    {
                    locus2.color.allele2[i]  <- sample(locus2.color.selfer[i,],1, replace =T)
                    }


                    # Make the array of the selfed progeny
          new.selfed.progeny <- data.frame(cbind(locus1.color.allele1,
                                                 locus1.color.allele2,
                                                 locus2.color.allele1,
                                                 locus2.color.allele2))

                    # Combine arrays, just use column names from selfed progeny
           new.generation <- rbind(new.selfed.progeny, setNames(new.outcrossed.progeny, names(new.
selfed.progeny)))
                    new.generation <- as.matrix(new.generation)

                    #Get homozygotes
                    homs <- subset(new.generation, locus1.color.allele1==locus1.color.allele2 &
                        locus2.color.allele1==locus2.color.allele2)

                    dim.homs<-dim(homs)

                    sum.homs <- dim.homs[1]

                    #How many heterozygotes?
                    hets <- popsize - sum.homs

                    #Get the morphotypes
                    dark <- rowSums(new.generation[,1:4]) > 0
                    light <- rowSums(new.generation[,1:4]) == 0
                    dark <- sum(dark)
                    light <- sum(light)

                    Output[igen,5] <- sum.homs
                    Output[igen,6] <- hets
                    Output[igen,7] <- light
                    Output[igen,8] <- light/popsize
                    Output[igen,9] <- dark
                    Output[igen,10] <- dark/popsize


            parents = new.generation
        }

    #Output

# Define outputs for 1000 simulations

    OutputSims[isim,1] <- isim
    OutputSims[isim,2] <- seeder
    OutputSims[isim,3] <- Output[igen,2]
    OutputSims[isim,4] <- mean(Output[igen,3])
    OutputSims[isim,5] <- mean(Output[igen,4])
    OutputSims[isim,6] <- mean(Output[igen,5])
    OutputSims[isim,7] <- mean(Output[igen,6])
    OutputSims[isim,8] <- mean(Output[igen,7])
    OutputSims[isim,9] <- mean(Output[igen,8])
    OutputSims[isim,10] <- mean(Output[igen,9])
    OutputSims[isim,11] <- mean(Output[igen,10])



}

#OutputSims

# Rename columns
Sim_No <- OutputSims[,1]
Seed_No <- OutputSims[,2]
popsize <- OutputSims[,3]
mean_no_outcross <- OutputSims[,4]
mean_no_selfers <- OutputSims[,5]
mean_homozygotes <- OutputSims[,6]
mean_heterozygotes <- OutputSims[,7]
mean_light <- OutputSims[,8]
mean_light_freq <- OutputSims[,9]
mean_dark <- OutputSims[,10]
mean_dark_freq <- OutputSims[,11]


simResults <- data.frame(cbind(Sim_No, Seed_No, popsize, mean_no_outcross,
    mean_no_selfers, mean_homozygotes, mean_heterozygotes,
    mean_light, mean_light_freq, mean_dark, mean_dark_freq))

simResults

##   Sim_No Seed_No popsize mean_no_outcross mean_no_selfers mean_homozygotes
## 1      1   16761    1000               21             979              975
## 2      2    1504    1000               19             981              988
##   mean_heterozygotes mean_light mean_light_freq mean_dark mean_dark_freq
## 1                 25        506           0.506       494          0.494
## 2                 12        333           0.333       667          0.667

save(simResults, file = "simulationResults.RData")
```

Appendix 4 R script used for simulating reproductive economy in an annual plant.

```r

#First set the parameters of the model.
threshold = c(0.9,0.91) # The minimum size required to reproduce for the two
strategies with a "minimum"/small/big difference between them.
lambda = 1/threshold # Set the parameter of the exponential distribution
popsize = 1000 # preferably large to minimize drift, but also to indicate
competition of a finite resource in a closed space
k=5 # seeds per unit plant size above threshold
totalGen=50 # Maximum number of generations for one simulation to run
totalSim=1000 #System time for this is 10 mins on this machine, reduce to get the
gist


#Set up the arrays for results output
Output=matrix(0,totalGen,7) #the seven outputs just for a generationloop are defined
below
Outputfinal=matrix(0,totalSim,11) #eleven outputs defined below

###LOOP THE LOOP THE LOOP####
###Set random seed from a uniform distribution of values, and keep track of
seeds#######
for(isim in 1:totalSim)
{
  #For deterministic results see below for pseudo random seeds, otherwise these
  #next two lines can be commented out.
  seeder<-round(runif(min=2, max = 80E4, n=1),2)
  set.seed(seeder)

  # Set up a population with half one strategy, half the other (coded as "TRUE" and
"FALSE")
  x = c(rep(FALSE,popsize/2),rep(TRUE,popsize/2))
  x = sort(x) # False is sorted ahead of true
  p = length(x[x == TRUE])/length(x)


  ###Loop loop###
  ### To loop over multiple generations or just one simulation, start the loop
here.##########
  for (igen in 1:totalGen)
  {
    Output[igen,1]<- igen     # Store Generation number in 1st column of output
matrix
    #Calculate the proportion of each type for every generation and store in the
Output
```

```
    Output[igen,2] <- p      # proportion of larger morph
    Output[igen,3] <- 1-p    #proportion of smaller morph

    # Set up a vector of sizes drawn from an exponential distribution
    # The bigger lambda is, the smaller the average size is
    size = c(rexp(length(x[x==FALSE]),lambda[1]),rexp(length(x[x==TRUE]),lambda[2]))
    Output[igen, 4]<- mean(size)
    #Before getting into the reproduction, set up arrays to keep track of fecundity,
seed type and size of parents
    fecundity = rep(0,popsize)
    seeds = array()
    parentsize = array()

    # seeds and parentsize will start with an initial entry of NA which needs to be
stripped away below

    # Loop over the population creating an array of seeds
    for (i in 1:length(size)) {
      if (x[i])       {       # for the "TRUE" morph (larger)...
        if (size[i] > threshold[2])         {       # If it exceeds the minimum size
then...
          fecundity[i] = floor(k*(size[i]-threshold[2]))   # it produces k seeds for
every unit mass over the threshold
          # note that this produces a geometric distribution of fecundity (the
discontinuous equivalent of exponential)
          # with the same lambda parameter as the size distribution
          seeds = c(seeds,rep(x[i],fecundity[i]))  # We record its seeds in the
vector of progeny
          parentsize = c(parentsize,rep(size[i],fecundity[i]))    # and we keep
track of the seed output of plants of different sizes
        }
      }
      else     {       # If the plant is the "FALSE" morph (reproductive economy)
        if (size[i]>threshold[1])           {       # as above, but using hte lower
threshold
          fecundity[i] = floor(k*(size[i]-threshold[1]))
          seeds = c(seeds,rep(x[i],fecundity[i]))
          parentsize = c(parentsize,rep(size[i],fecundity[i]))
        }
      }

    }
    ###Close parent loop#####

    # Since we added all the seeds that started with an NA entry, we strip off the
first entry and just use entries 2-n.
```

```
    seeds = seeds[2:length(seeds)]
    parentsize = parentsize[2:length(parentsize)]

    # Take popsize seeds at random to start the next generation
    x= sample(seeds, popsize, replace=FALSE) #If I remove popsize, this will be more
than 1000 individuals and then the loop hangs itself by the seventh or 8th
generation- I think
    x = sort(x)       #Sort them (False comes ahead of true)
    p = length(x[x == TRUE])/length(x)      #And calculate the frequency of the
"TRUE" (larger) morph

    Output[igen,5]<-mean(fecundity)
    Output[igen,6]<-max(fecundity)
    Output[igen,7]<-max(size)

    ###Pulling other "useful" summary stats out, using logical conditions - make
dataframe
    Outputdf<-as.data.frame(Output)
    maxsize<-Outputdf[Outputdf$V4==max(Outputdf$V4),] #pull out the entire row of
max size
    genmaxsize<-maxsize[1,1] #pull out the Gen no in which max size occurs
    maxfec<-Outputdf[Outputdf$V5==max(Outputdf$V5),] #pull out the entire row of max
fecundity
    genmaxfec<-maxfec[1,1] #pull out the Gen no in which max fecundity occurs
    fixationall<-Outputdf[Outputdf$V2>=1.0,] #get all the rows where p=1
    fixationfirstgen<-fixationall[1,1] #isolate only the first instance (generation)
where p=1

  }

  #########################closes Igen
loop################################################################################
##################################



  ###The output with comments but no real names
  Outputfinal[isim,1] <-isim
  Outputfinal[isim,2] <-seeder #If deterministic only, otherwise no need to keep
track
  Outputfinal[isim,3] <- mean(Output[igen,2])     # proportion of larger morph p
  Outputfinal[isim,4] <- mean(Output[igen,3])             #proportion of smaller
morph q
  Outputfinal[isim,5] <- mean(Output[igen,4])             #mean size of that
simulation
  Outputfinal[isim,6] <- mean(Output[igen,5])             #mean fecundity that
```

```
simulation
  Outputfinal[isim,7] <- Output[igen,6]                    #max fecundity of that
simulation
  Outputfinal[isim,8] <- Output[igen,7]                    #max size of that
simulation
  Outputfinal[isim,9] <- genmaxsize         #in which generation of that simulation
is maximum size achieved?
  Outputfinal[isim,10] <- genmaxfec
  Outputfinal[isim,11] <- fixationfirstgen


}


####closes Isim loop - run 1000 - takes too much time on this machine####

#Output
Outputfinal ##prints [1000,11] matrix with no names

#Renaming the Outputfinal columns
Sim_No<-Outputfinal[,1]
Seed_No<-Outputfinal[,2]
final_p_freq<-Outputfinal[,3]
final_q_freq<-Outputfinal[,4]
mean_size_sim<-Outputfinal[,5]
mean_fec_sim<-Outputfinal[,6]
max_fec_sim<-Outputfinal[,7]
max_size_sim<-Outputfinal[,8]
gen_max_size_out<-Outputfinal[,9]
gen_max_fec_out<-Outputfinal[,10]
fixation_p_1stgen<-Outputfinal[,11]

#### Make a dataframe of these names
results.df<-data.frame(cbind(Sim_No, Seed_No,final_p_freq,
                             final_q_freq,mean_size_sim,mean_fec_sim,max_fec_sim,
                             max_size_sim,gen_max_size_out,gen_max_fec_out,
                             fixation_p_1stgen))

####Print the dataframe
results.df
summary(results.df)

####Write it out
write.csv(results.df, file="results_aarssen_very_little_difference.csv")
```