The Maintenance of Genetic Variation by Environmental Selection

by

Cheng-Ruei Lee

Department of Biology
Duke University

Date:_____
Approved:

_____
Thomas Mitchell-Olds, Supervisor

_____
Philip Benfey

_____
Mohamed Noor

_____
Mark Rausher

_____
John Willis

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biology in the Graduate School
of Duke University

2013

ABSTRACT

The Maintenance of Genetic Variation by Environmental Selection

by

Cheng-Ruei Lee

Department of Biology
Duke University

Date:_____
Approved:

_____
Thomas Mitchell-Olds, Supervisor

_____
Philip Benfey

_____
Mohamed Noor

_____
Mark Rausher

_____
John Willis

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology in the Graduate School of Duke University

2013

Abstract

Understanding forces creating or maintaining the vast amount of biodiversity has been a major task of biologists. Genetic variation plays a major role in the creation of biodiversity because in contrast to environmental influence, genetic variants can be inherited. For a species in natural environments, genetic variation is generated by mutation, eliminated by genetic drift or selective sweep, and maintained by balancing selection that favors different alleles in different environments or time. In my dissertation, I will address how spatially heterogeneous environmental selection maintains genetic variation in two aspects.

Genes in the genome vary vastly in their level of polymorphism. Previous studies have used features within the genome, such as recombination rate or expression level, to explain the variation in gene polymorphism. One factor, however, that has often been overlooked is the effect of environmental adaptation on gene polymorphism. Specifically, if different alleles of a gene are responsible for local adaptation to distinct environments, the polymorphism of this gene will be actively maintained by spatially heterogeneous environmental selection. In the first part (Chapter 2) of my dissertation, I used publicly available genomic data from *Arabidopsis thaliana* to address this question. I found that environmental relevance of a gene has a significantly positive relationship with the variation in polymorphism level among genes in the *Arabidopsis* genome, consistent with the hypothesis that environmental selection actively maintains the polymorphism of environmentally responsive genes.

A biological species is formed by a mating pool of individuals, and for two populations of the same species, differentiation is often homogenized by gene flow. Reproductive isolation between populations allows genetic differentiation, and therefore

speciation, the process in which full reproductive isolation is achieved between populations, plays important role in generating biodiversity. In the second part of my dissertation I used *Boechera stricta* to address how environmental selection contributes to speciation. In Chapter 3, I used niche modeling to show that environmental factors have more important roles than geographical distance in the genetic differentiation of EAST and WEST subspecies, and local water availability is the most important factor. In Chapter 4, I performed large-scale greenhouse experiments to identify key traits responsible for the EAST-WEST local adaptation, and that those traits have significantly larger differentiation between subspecies than neutral expectation. In Chapter 5, I performed quantitative trait loci mapping for those important traits and fitness in both parental environments and greenhouse. In summary, the second part of my dissertation provides an example to study ecological speciation from the environment, trait, to the genetic level.

To my wonderful family, Chin-Hsiung, Hsueh-Ying, Ju-Ting, and Jui-Ju.

Contents

Abstract iv

List of Tables xi

List of Figures xii

List of Appendix A Supplementary Tables xiv

List of Appendix B Supplementary Figures xv

Acknowledgements xvi

1. Introduction 1

1.1 Overview of part 1 (Chapter 2) 2

1.2 Overview of part 2 (Chapter 3 to 5) 3

2. Environmental adaptation contributes to gene polymorphism across the Arabidopsis thaliana genome 5

2.1 Materials and methods 7

2.1.1 Data source 7

2.1.2 Calculating environmental relevance at individual loci 8

2.1.3 Genome-wide analysis among loci 10

2.1.4 Analysis with different groups of environmental variables and accessions 12

2.1.5 Gene ontology term enrichment of high environmental relevance genes 12

2.2 Results 13

2.2.1 Environmental relevance predicts genomic patterns of polymorphism 13

2.2.2 Genes with high environmental relevance are enriched in unknown functions 15

2.2.3 Consistent results were obtained form different subsets of data 16

2.3 Discussion 17

2.3.1 Environmental relevance predicts polymorphism among genes 18

2.3.2 The polygenic nature of environmental adaptation 20

2.3.3 Relationship to other studies 21
2.3.4 Conclusion 22
2.4 Data availability 22
3. Quantifying Effects of Environmental and Geographical Factors on Patterns of Genetic Differentiation 23
3.1 Materials and methods 26
3.1.1 Study species 26
3.1.2 Genotyping 27
3.1.3 Genetic analysis 27
3.1.4 Environmental variables 29
3.1.5 Niche modeling 30
3.2 Results 33
3.2.1 Genetic structure in *Boechera stricta* 33
3.2.2 Contribution of environment versus geography to population structure 35
3.2.3 Identifying sources of environmental selection 37
3.3 Discussion 38
3.3.1 Contribution of environment versus geography to population structure 39
3.3.2 Identifying sources of environmental selection 43
3.3.3 Conclusion 45
3.4 Data availability 46
4. Complex trait divergence contributes to environmental niche differentiation in ecological speciation of *Boechera stricta* 47
4.1 Materials and methods 49
4.1.1 Plant material 49
4.1.2 Experiment 1. Short-term drought manipulation and phnology 50
4.1.3 Experiment 2. Long-term drought manipulation 53

4.1.4 Experiment 3. Vegetative-phase morphology without drought treatment 54
4.1.5 Principal component analysis 55
4.1.6 Calculation of univariate and multivariate Q_{ST} 56
4.1.7 Empirical SNP F_{ST} distribution 57
4.2 Results 59
4.2.1 No significant divergence in eco-physiological traits between subspecies 59
4.2.2 Trait divergence between EAST and WEST subspecies 60
4.2.3 Comparing F_{ST} to univariate and multivariate Q_{ST} 62
4.3 Discussion 66
4.3.1 Trait divergence corresponds to niche modelingpredictions 67
4.3.2 Lack of physiological differentiation 69
4.3.3 Lack of geographic effects 70
4.3.4 Comparing F_{ST} with multivariate Q_{ST} 72
4.3.5 Conclusion 74
4.4 Data availability 75
5. Quantitative trait loci mapping identifies genomic region controlling ecological speciation of *Boechera stricta* 76
5.1 Materials and methods 79
5.1.1 Plant materials, phenotypic measurements, and trait analyses 79
5.1.2 Genotyping by sequencing 83
5.1.3 Quantitative trait locus (QTL) mapping 85
5.2 Results and Discussion 87
5.2.1 Quantitative traits, heritability, and fitness components 87
5.2.2 Linkage map 88
5.2.3 Quantitative trait loci for important traits 90

5.2.4 Co-localization of fitness and trait QTL 96
5.2.5 Conclusion 98
Appendix A. Supplementary tables 104
Appendix B. Supplementary figures 126
References 129
Biography 141

List of Tables

Table 1: Different distribution of genes with high vs. low environmental relevance values in gene ontology Slim terms. Shown are the *P* values from chi-square tests[a] between genes with top 20% and lower 80% environmental relevance values. Asterisks denote *P* values less than 0.05. 16

Table 2: Proportion of genetic variation explained by environmental (ENV) and geographical (GEO) effects or their interaction in the EAST-WEST (species-wide, genetic PC1) and the NORTH-SOUTH (within-EAST, genetic PC2) genetic divergence patterns 36

Table 3: Proportion of EAST-WEST (genetic PC1) genetic variation explained by climatic (CLIM), topographical (TOPO), geographical (GEO), or the interaction effects in the allopatric or sympatric regions 37

Table 4: *P* values based on likelihood ratio tests in multiple logistic regressions on EAST-WEST genotypes (a binary response variable) in the allopatric or sympatric regions. 38

Table 5: Mixed model ANOVA results of water use efficiency in short-term and long-term drought experiment 60

Table 6: Divergence of the 'composite trait' for each trait category. For DFA scores from each trait category, this table shows the *P*-value of the subspecies effect in univariate ANOVA, the Q_{ST}, and the empirical *P*-value of Q_{ST} compared to genome-wide distribution of SNP F_{ST} (Figure 4). Data are from all 24 genotypes. 66

Table 7: Relative contribution of survival, bolting, and fecundity fitness components to the variation of overall fitness at the family level 88

Table 8: One-way ANOVA and power analysis of fitness QTL identified in one field environment on the corresponding fitness components in the other garden 94

List of Figures

Figure 1: Proportional contribution of each predictor category to the variation of gene polymorphism in *A. thaliana*. There are four predictor categories (PHY – physical properties; FUN – functional constraints; ENV – environmental relevance; DUP – duplication status) and three separate measures of genetic polymorphism (π – total polymorphism; π_N – nonsynonymous polymorphism; π_S – synonymous polymorphism). (A) All 80 accessions (B) 65 accessions, excluding Russia and Central Asia. 14

Figure 2: Proportional contribution of each predictor variable to the variation of gene polymorphism in *A. thaliana*. There are sixteen predictor variables (Appendix Table S1) and three separate measures of genetic polymorphism (π – total polymorphism; π_N – nonsynonymous polymorphism; π_S – synonymous polymorphism). The height of each bar represents total variation explained by the full model. Each colored box represents the partial variation explained by one factor, and the grey bars are variations explained by the correlation among predictor variables. (A) All 80 accessions (B) 65 accessions, excluding Russia and Central Asia. 15

Figure 3: Collection sites and STRUCTURE results for *Boechera stricta*. Each pie chart represents one individual randomly chosen from one location. Different colors in each pie chart represent STRUCTURE posterior probabilities that the individual belongs to each genetic group. A) The distribution of three genetic groups across western North America. Red = WEST; blue = NORTH; green = SOUTH. Notice the narrow contact zone between WEST and EAST (comprised of NORTH + SOUTH), and the clinal distribution between NORTH and SOUTH genetic groups. B) The distribution of WEST and EAST genetic groups around the contact zone. Red = WEST; blue = EAST. Region encompassed by the dashed line is regarded as 'sympatric zone'..... 34

Figure 4: Genetic principal component analysis (PCA) of 239 *Boechera stricta* accessions. PC1 explains 40.4% and PC2 explains 17.2% of total genetic variation. Accessions were colored based on STRUCTURE results with k = 3, and a genotype belongs to a 'pure genetic group' (W = WEST, N = NORTH, S = SOUTH) only when the corresponding posterior probability is higher than 0.8. 'NS' and 'WE' denote NORTH-SOUTH hybrids and WEST-EAST (EAST = NORTH + SOUTH) hybrids, respectively. Notice the distinct distribution patterns between WEST-EAST along PC1 (discrete) and NORTH-SOUTH along PC2 (continuous). 35

Figure 5: Collection sites of 24 genotypes used in this study. The region is denoted as a black star on the state boundary map. Blue circles – allopatric EAST. 49

Figure 6: Principal components of genotype-level trait values. EAST genotypes - closed circles. WEST genotypes - open circles. A - all traits. B – four physiology traits. C – eight phenology traits. D – five stalk morphology traits. E – thirteen rosette morphology traits. F – nine leaf shape traits. Refer to Table 2 for the traits within each category. 61

Figure 7: Relationship between trait Q_{ST} and (A) negative log *P*-value of subspecies effect in ANOVA (B) absolute value of correlation with discriminant function analysis (DFA) score from each trait category. Traits with high Q_{ST} generally have low *P*-values (high

negative log *P*) and high correlation with DFA score. Consistent with Figure 2, many morphological and phenological traits are highly diverged. Shown are data from all 24 genotypes. 62

Figure 8: Empirical SNP F_{ST} distribution between 11 WEST and 8 EAST genotypes. 64

Figure 9: Average leaf shape of EAST and WEST genotypes (n = 60 from each subspecies). EASTERN leaf - closed circles connected by dashed line. WESTERN leaf - open circles connected by solid line. For every leaf, landscape points were rotated and scaled to obtain equal length among all leaves (a standardized length of 100 units across the horizontal axis), and points Y1 to Y9 separate the central leaf axis (dotted line) into ten sections of equal length. The Y coordinates of Y1 to Y9 were used in the statistical analyses. 64

Figure 10: Linkage map of *Boechera stricta*. Horizontal lines on each linkage group represent genetic markers. 89

Figure 11: Multivariate least square interval mapping (MLSIM) result for each trait category. Around a QTL peak, the region where the statistical value is higher than the permutation significance threshold is marked in black.....92

Figure 12: Quantitative trait loci (QTL) of univariate traits in three environments on seven *Boechera stricta* chromosomes. Each graph represents chromosome 1 to 7 in order. Within each graph, columns are univariate traits where three environments are separated by two vertical black lines, and rows are centi-Morgan on the linkage map. QTL and confidence intervals are presented as colored bars, where blue means the Parker (EAST subspecies) allele has higher trait value and red means the Ruby (WEST subspecies) allele has higher trait value. Darker red or blue region represents 1-LOD confidence interval, and lighter red or blue region represents 2-LOD confidence interval. 103

List of Appendix A Supplementary Tables

Table S 1: Predictor variable used in Chapter 2 104

Table S 2: Twenty environmental variables used to estimate the environmental relevance of each gene in Chapter 2 105

Table S 3: Seventeen microsatellite loci and their primer sequences used in Chapter 3. 106

Table S 4: Environmental variables used in Chapter 3 107

Table S 5: Trait divergence between subspecies. For all traits in 24 genotypes, shown are trait, category, *P*-value for subspecies in ANOVA, the subspecies with higher trait value, Q_{ST}, *P*-value of Q_{ST} compared to empirical F_{ST} distribution, and the correlation with discriminant function analysis (DFA) score from each trait category. 108

Table S 6: Trait divergence between subspecies. For all traits in 19 genotypes, shown are trait, category, *P*-value for subspecies in ANOVA, the subspecies with higher trait value, Q_{ST}, *P*-value of Q_{ST} compared to empirical F_{ST} distribution, and the correlation with discriminant function analysis (DFA) score from each trait category. 110

Table S 7: Divergence of the 'composite trait' from each trait category. This table shows the data from 19 genotypes. For DFA scores from each trait category, this table shows the subspecies effect *P*-value in univariate ANOVA, the Q_{ST}, and the empirical *P* value of Q_{ST} compared to genome-wide distribution of SNP F_{ST}. 112

Table S 8: List of all univariate traits in Chapter 5 113

Table S 9: List of all composite traits used in Chapter 5 116

Table S 10: Adaptor oligos and PCR primers in the modified Andolfatto *et. al.* (2011) protocol 117

Table S 11: List of all QTL, allelic direction, and proportional genetic variation explained in Chapter 5 121

List of Appendix B Supplementary Figures

Figure S 1: The 24 genotypes represent most of the (A) geographical and (B) genetic variation among all *Boechera stricta* accessions in my study area (Latitude: 43.50 to 46.00 N, Longitude: 111.00 to 116.00 W). In both panels, white stars represent 24 core genotypes used in this study, blue dots represent EASTERN genotypes, red dots represent WESTERN genotypes, and pink dots represent hybrids. All data are obtained from Lee and Mitchell-Olds (2011). Genetic groups (EAST/WEST/hybrid) were assigned by STRUCTURE.....126

Figure S 2: Example of multivariate trait divergence in phenology, assuming natural selection favors the divergence in 'total reproduction time' between the red and blue population. Each point represents one genotype. (A) This trait, although not directly measured, is a linear combination of flowering time and duration. The two populations may diverge in either flowering time (B), duration (C), or both (D). In examples (B) and (C), the traits under divergent selection could be identified via their high Q_{ST}. In case (D), however, no univariate trait has Qst higher than the significance threshold, and the divergent selection on phenology as a whole might not be identified. Nevertheless, these three examples all have the same amount of divergence in total reproduction time. In case (D), the composite trait under strongest divergent selection (and therefore its Q_{ST}) could be identified via discriminant function analysis or MANOVA between the two populations. Notice that this method only involves a rotation of axis and does not produce an upward bias in multivariate Qst. Finally, in (E) if none of the univariate or multivariate traits has diverged, the multivariate Qst also will be low.....127

Figure S 3: Relationship between trait Q_{ST} and (A) P value of subspecies effect in ANOVA (B) absolute value of correlation with discriminant function analysis (DFA) score from each trait category. Shown are data from 19 genotypes. All axes and scales are equivalent to Figure 7.....128

Acknowledgements

I offer my deepest thanks to my advisor, Tom Mitchell-Olds, for your intellectual and funding support. You are always open-minded and give me a lot of freedom to investigate various subjects, most of which turn out to be my valuable side projects. Thank you for bringing in those valuable chances of collaborations. You are always a perfect role model of being highly collaborative and versatile, and without you, I can never understand that besides ecology and evolution, we can use our knowledge to help improve important crops for those in need. Equally important is your help and advise outside of science: you showed me how to change a flat tire and taught me how to survive the world of ticks, rattlesnakes, bears, and thunderstorms during our fieldwork. These are all invaluable experience for a city kid from another culture.

I also offer my thanks to my committee members. Thank you, Mark Rausher, for always pointing out critical flaws in my logic and willing to spend time answering my statistical and scientific questions. Thanks to John Willis for always keeping eyes on the newest literature and willing to share with all of us, and your passion and happiness always encourage me. Thank you, Mohamed Noor, for your valuable advise. Thank you, Philip Benfey, for representing the molecular biology part of my committee and sharing with me the chance to work on rice.

I want to thank all members of the Mitchell-Olds lab, especially Kasavajhala Prasad. Thank you for always spending time teaching me knowledge and techniques in molecular biology. Thank you Jill Anderson for answering my endless questions and always provide me with invaluable chances for collaboration. I thank Carrie Olson-Manning for teaching me molecular cloning techniques and answering my endless questions through email even after you leave the lab. Thank you Rob Colautti for your

always-valuable insights in quantitative genetics. Thank you Bao-Hua Song. Your earlier works on *Boechera stricta* form the basis of my PhD study. I also thank Eric Schranz, and although we have never overlapped, the large amount of plant resources and information you left are invaluable for we young investigators. I also want to thank other members in the lab who have helped me with lab or field works: Antonio, Kathy, Chun-Lin, Cathy, Maggie, Rose, Nadeehsa, Sara, Evan, Katie, Michael, Kate, Marshall, Tim, Kara, and Nadia.

I also thank the supports from the University and Biology Department: the DSCR team, Biology IT team, greenhouse team especially John Mays, and also Anne Lacey, Jim Tunney, Jo Bernhardt, and all other administrative supports.

For genomic analyses, I certainly cannot generate all data on my own. For chapter 2, I thank the Weigel laboratory at the Max Planck Institute of Developmental Biology and the Gaut laboratory at University of California Irvine for making their data publicly available. For chapter 4 and 5, I thank Kasavajhala Prasad and many in the Joint Genome Institute, especially Kerrie Barry, Uffe Hellsten, and Stephen Fairclough, for their work on the *Boechera stricta* reference genome.

My family in Taiwan, more than 8,000 miles away, has been very supportive for my pursuit of PhD alone in the US. Thank you my parents Chin-Hsiung and Hsueh-Ying, for your emotional support. Thank you Ju-Ting, my sister, you are always the best friend I ever have. Thank you Hui-Ju, my dear wife, for your understanding and emotional support through all these years. Hearing your lovely voice through the phone always gives me strength to go forward.

1. Introduction

Understanding forces contributing to the existence of biological diversity is an important task in biological study. Biodiversity typically refers to the phenotypic variation among organisms, and the phenotypic variation is created by environmental and genetic influences. Genetic variation, in particular, is an active area of research because unlike the ephemeral environmental variation, genetic variation is heritable and can be passed down through evolutionary timescales.

Many factors affect the level of genetic variation (Hartl & Clark 2007). On one hand, genetic drift and selective sweeps reduce genetic variation. On the other hand, in some situations the amount of mutation or migration alone may not be sufficient to replenish lost variation. Therefore, factors maintaining genetic variation may play an important role in shaping the patterns of biodiversity in nature. Spatially heterogeneous environmental selection is one of those factors (Turelli & Barton 2004). Genetic variation might be maintained by differential local adaptation, where distinct natural environments favor different phenotypes and therefore different allelic combinations of underlying genes. The migration and homogenization of individuals or alleles is restricted by natural selection against unfit genotypes, and as a consequence, the variation of ecologically important genes, traits, or associated lineages may be maintained. In my thesis I will address two aspects of spatially heterogeneous environmental selection's effect on genetic variation. Part 1 (Chapter 2) uses a genomic approach to study the contribution of environmental selection to the variation of polymorphism levels across genes. Part 2 (Chapter 3 to 5) uses ecological and quantitative genetic approaches to study the influence of environmental selection on the accumulation of overall within-species genetic variation.

1.1 Overview of part 1 (Chapter 2)

The level of within-species polymorphism differs greatly among genes in a genome. Many genomic studies have investigated the relationship between gene polymorphism and factors such as recombination rate or expression pattern (Comeron *et al.* 1999; Lercher & Hurst 2002; Pal *et al.* 2001). However, the polymorphism of a gene is affected not only by its physical properties or functional constraints, but also by natural selection on organisms in their environments (Hedrick 2006). Specifically, if functionally divergent alleles enable adaptation to different environments, locus-specific polymorphism may be maintained by spatially heterogeneous natural selection. Therefore, I expect that genes under spatially balancing selection will have higher variation than the rest of the genome. Few studies have investigated whether or how much the 'environmental relevance' of each gene contributes to the difference in polymorphism levels across genes in a genome. In Chapter 2 I use publicly available data from 80 sequenced Arabidopsis thaliana genomes to test this hypothesis and estimate the extent to which environmental selection shapes the pattern of genome-wide polymorphism. I calculated the 'environmental relevance' of each gene and found substantial effects of environmental relevance on patterns of polymorphism among genes. In addition, the correlation between environmental relevance and gene polymorphism is positive, consistent with the expectation that balancing selection among heterogeneous environments maintains genetic variation at ecologically important genes. These results suggest an important role for environmental factors in shaping genome-wide patterns of polymorphism, and this chapter is one of the first successful attempts to use environmental factors to explain the variation of polymorphism levels across genes in the genome.

1.2 Overview of part 2 (Chapter 3 to 5)

In Chapter 2, I focus on the variation in polymorphic levels of genes in the genome. Part two of my dissertation is focused on how environmental selection affects the other aspect of genetic variation. In a species with high gene flow among populations, neutral polymorphism could be eliminated by genetic drift, fixing one allele species-wide. On the other hand, low gene flow or reproductive isolation among lineages allows the possibility of fixing different alleles among lineages, thereby allowing the accumulation of species-wide genetic variation. Therefore, spatially heterogeneous environmental selection could contribute to the overall accumulation of within-species polymorphism by creating reproductive isolation among lineages. This process, often termed ecological speciation (Rundle & Nosil 2005) or isolation by adaptation (Nosil *et al*. 2008), is generated by the interaction of many aspects in nature, such as heterogeneous natural selection in distinct environments, the distinct phenotypes suitable for each environment, the fitness as a consequence of environment-phenotype interaction, and the genetic basis of this ecological speciation. However, only in a few organisms have each of these processes been examined jointly. In the second part my dissertation, I will use *Boechera stricta* as model to investigate the extent to which environmental selection contributes to genetic variation (Chapter 3), identify the selection force and phenotypic response (Chapters 3 and 4), and the loci controlling these important traits (Chapter 5).

In Chapter 3, I estimate the quantitative contributions of environmental adaptation and isolation by distance on genetic variation in *Boechera stricta*. Between two subspecies (EAST and WEST), environmental factors have larger contribution than geography. I further identify water availability as the possible cause of differential local

adaptation in both geographic regions. This chapter shows that geographical and environmental factors together created stronger and more discrete genetic differentiation than isolation by distance alone, which only produced a gradual, clinal pattern of genetic variation. These findings emphasize the importance of environmental selection in shaping patterns of species-wide genetic variation in the natural environment.

In Chapter 4, I perform several large-scale greenhouse experiments to investigate the divergence of various physiological, phenological, and morphological traits. The WEST subspecies has faster growth rate, larger leaf area, less succulent leaves, delayed reproductive time, and longer flowering duration. These trait differences are concordant with previous results that habitats of the WEST genotypes have more consistent water availability. By comparing univariate and multivariate divergence of complex traits (Q_{ST}) to the genome-wide distribution of SNP F_{ST}, I conclude that aspects of phenology and morphology (but not physiology) are under divergent selection.

After identifying water availability as an important selective factor responsible for the local adaptation in Chapter 3 and the important traits in Chapter 4, in Chapter 5 I conduct quantitative trait loci mapping. Several QTL are identified for fitness in the field environments (two environments corresponding to the two parental subspecies) and for important traits such as rosette leaf succulence. The QTL for field fitness show signs of conditional neutrality – those in the WEST garden do not co-localize with those for EAST garden, and I find no sign of reciprocal changes in rank fitness. The detailed mechanism responsible for this ecological speciation process remains to be investigated.

2. Environmental adaptation contributes to gene polymorphism across the Arabidopsis thaliana genome

Evolutionary biologists have long been interested in factors influencing genetic variation among and within species. With the availability of whole-genome sequences, I can now investigate both genetic variation among individuals within a clade and among genes within a genome. Between species, Yang and Gaut (2011) examined the factors that contribute to evolutionary rate variation among genes by modeling the pattern of divergence between *Arabidopsis thaliana* and *A. lyrata* using 14 properties of each gene. Within species, many intrinsic factors of a genome contribute to the polymorphism of a gene, such as local recombination rate (Comeron *et al.* 1999; Lercher & Hurst 2002), local gene density (Flowers *et al.* 2012), expression pattern (Pal *et al.* 2001), and chromosome (Andolfatto *et al.* 2011b; Bachtrog & Charlesworth 2002). However, to my knowledge the role of environmental heterogeneity in maintaining gene polymorphism has not been investigated at the whole-genome level.

If the biological function of a gene controls environmental adaptation, the geographic distribution of different alleles may be associated with spatially heterogeneous environmental factors, such as temperature or precipitation. Several recent studies have used similar logic to identify SNPs or genes responsible for environmental adaptation in humans (Hancock *et al.* 2010), pine trees (Eckert *et al.* 2010a; Eckert *et al.* 2010b), and *Arabidopsis* (Hancock *et al.* 2011). In addition, a gene responsible for differential environmental adaptation may also be more polymorphic, because balancing selection might actively maintain locus-specific polymorphism, making it harder for one allele to fix across the species range either through drift or selective sweep (Gillespie & Langley 1974; Hedrick 1986). Although spatially

heterogeneous environmental selection has been the focus of many single-gene studies (Hedrick 2006), the importance of such environmental selection in maintaining genetic polymorphism has not been examined on patterns of polymorphism across the genome.

In this study, I used genome sequences from 80 *A. thaliana* accessions (Cao *et al.* 2011) to estimate the extent to which spatially heterogeneous environmental selection shapes the level of polymorphism in individual loci across the genome. For each gene, I calculated the 'environmental relevance': the proportion of genetic variation explained by the local environments of each accession, after controlling for population structure. This environmental relevance is an estimate of the association between a gene's biological function and environmental conditions. The environmental relevance of each gene is then used as a predictor variable to model its effect on the pattern of total, synonymous, and nonsynonymous polymorphism within *Arabidopsis thaliana*. If heterogeneous environments maintain polymorphism in particular genes, environmental relevance may predict the variation of nonsynonymous polymorphism in the genome. In addition, incorporating data from Yang and Gaut (2011), I also compare the importance of environmental relevance vs. variables representing the physical properties and functional constraints of a locus, which are crucial in shaping the evolutionary pattern of genes (Andolfatto *et al.* 2011b; Bachtrog & Charlesworth 2002; Comeron *et al.* 1999; Flowers *et al.* 2012; Lercher & Hurst 2002; Pal *et al.* 2001).

My major goal is to identify the extent to which environmental influences shape the different levels of polymorphism among genes. In addition, because heterogeneous environmental selection would maintain polymorphism in corresponding genes, I further test the prediction that gene polymorphism and environmental relevance should be positively correlated.

2.1 Materials and methods

2.1.1 Data source

Genome sequences of eighty *Arabidopsis thaliana* accessions were downloaded from the MPICao2010 subset (Cao *et al.* 2011) of the *Arabidopsis* 1001 genome project (http://1001genomes.org/). From the annotation information in TAIR10 and the genome matrix containing 80 accessions (Cao *et al.* 2011), I extracted coding sequence alignments of the specific splicing variant from *A. thaliana* genes used in the Yang and Gaut (2011) data set. Further filtering removed individual sequences meeting any of the following criteria: 1) pre-mature stop codons, or 2) lengthy 'bad bases' (ambiguous sites, alignment gaps, and regions affected by frame-shift mutation) exceeding 20% of full length.

From Yang and Gaut (2011), I adopted 13 variables representing the physical properties and functional constraints of 11,492 *A. thaliana* protein coding genes. I used my calculation of 'coding sequence length' rather than Yang and Gaut's 'gene length'. In addition, although the four states in the 'duplication status' variable were originally used as integers ranging from 1 to 4 in the analysis of Yang and Gaut (2011, which assumed a numeric relationship among the four duplication states), in my statistical model I treated duplication status as a categorical, nominal variable with four distinct states (singletons or early/ recent/ non-whole genome duplications). See Appendix Table S1 for detailed description of each variable in my main model.

Based on the geographical coordinates of 80 *Arabidopsis thaliana* accessions (Cao *et al.* 2011), I extracted elevation and 19 climatic variables (Appendix Table S2) from the Worldclim database (Hijmans *et al.* 2005). Those 20 variables were used to estimate the environmental relevance of each gene.

2.1.2 Calculating environmental relevance at individual loci

For each gene, I define its environmental relevance as the proportion of genetic variation explained by environmental factors while simultaneously controlling for population structure. Therefore, environmental relevance is undefined for monomorphic genes. To estimate the population structure within *Arabidopsis thaliana*, I used SMARTPCA (Patterson *et al*. 2006) to calculate the genomic background principal component (PC) scores of each accession, using all available SNPs in the genome.

I first created the polymorphic codon matrix of each gene. Rows of the matrix represent individual accessions, and columns represent polymorphic codon sites. Each cell has a value of 0 or 1, denoting whether an accession in a specific codon site has the same allele as the reference genome or the alternative allele. Three separate environmental relevance values were calculated based on the total, synonymous, or nonsynonymous polymorphic codon matrices. The three environmental relevance values were later used for three independent genome-wide analyses (with π, π_N, and π_S as response variables, described below). Because some genes have only synonymous or nonsynonymous polymorphism, the number of genes with available environmental relevance values differs among the three analyses. To estimate the proportion of within-locus genetic variation explained by these twenty environmental variables and five genomic background PC values, I first performed principle component analysis (PCA) on the polymorphic codon matrix separately for each gene. PCA gave p principal component axes (PCi, where i ranges from 1 to p) for each gene, where p equals the number of polymorphic codons in a gene and varies among genes. With PC scores for each orthogonal axis as response variables in turn, I analyzed the following multiple linear regression models, using ~80 *A. thaliana* sequences as data points for each gene:

$$PC_i = ENVar(20) + GenomePC(5) + error,$$

where PC*i* is the score of each accession in one of the *p* PC axis of this gene, ENVar(20) are 20 environmental variables at accession collection sites, and GenomePC(5) are the scores on the first five genomic-background PC axes calculated from whole-genome SNP data (serving as a control for population structure). For each axis PC*i* of genetic variation at a locus, the proportional genetic variation explained by environmental factors is obtained by comparing this full model (with 25 predictors) to a reduced model (five predictors):

$$PC_i = GenomePC(5) + error,$$

which models the effect contributed only by population structure. The environmental contribution in this PC axis is further weighted by the proportional importance (the proportion of eigenvalues) of the current PC axis (PC*i*) in this gene, and environmental relevance is obtained by summing this weighted proportion from all *p* PC axes for this gene. The statistical steps were performed and automated in R (http://www.r-project.org/). The possibility of model over-fitting might be raised regarding the use of all 20 environmental variables (some of which are correlated) in the same model. However, here I merely estimated the joint contribution of all 20 variables rather than the specific effect from each, and the same procedure was applied to all PC axes in all genes. Therefore, this procedure does not cause gene-specific bias in the estimation of environmental relevance.

Two other methods may be used to estimate environmental relevance. The first one is canonical correlation analyses between the codon matrix and environmental variable matrix. However, in some genes the presence of codons with highly similar polymorphic patterns makes the correlation matrix singular, and therefore I were not able to perform canonical correlation analysis on many genes. The other method is based

on partial Mantel's test (Hancock *et al.* 2011), where pair-wise distance matrices among accessions were used, with gene-specific genetic distance, environmental distance, and genome-wide genetic distance (kinship matrix) in the model. I did not use this method because: 1) In partial Mantel's test, the same environmental distance matrix is used across all genes, which does not allow different environmental variables to have different contributions to different genes; 2) While a partial Mantel's test is suitable for determining the significance of predictor effects, some studies have shown that this method does not correctly estimate the proportion of total variation explained by predictor matrices (r^2), which is my main focus here (Balkenhol *et al.* 2009; Legendre & Fortin 2010); 3) my linear modeling approach provides statistical flexibility to compare a range of alternative models.

2.1.3 Genome-wide analysis among loci

To model influences on polymorphism among *A. thaliana* genes, I quantified the level of variation at each locus using three different response variables: mean pairwise difference per nucleotide (π), mean pairwise d_N (π_N), and mean pairwise d_S (π_S) between aligned sequences of each gene. I used the PopGen module (Stajich & Hahn 2005) in Bioperl (Stajich *et al.* 2002) to calculate the mean pairwise nucleotide difference of each gene, and π is obtained by scaling this value with the coding sequence length. For each gene, I calculated pairwise d_N and d_S between all sequence pairs using the likelihood-based program codeml (runmode -2) in PAML 4 (Yang 2007), and π_N and π_S are obtained by averaging all pairwise values.

Due to the highly skewed distribution of almost all quantitative variables, I log-transformed them before final analysis, which greatly improves the normality of residuals. From the 11,492 genes, I excluded genes with any missing data in the Yang

and Gaut (2011) data set, leaving 5,919 genes. I further excluded monomorphic genes and genes with available sequence data from less than 40 accessions, leaving 5873 genes for π, 5841 for π_S, and 5722 for π_N in the final analysis. The genome-wide analysis uses genes as data points in a fixed-effect ANCOVA model with both quantitative and categorical (chromosome and duplication status) predictor variables:

$$PI = PHY(9) + FUN(5) + DUP + ENV + error,$$

where PI is the univariate response variable (π, π_N, or π_S) for the three separate analyses, PHY(9) are nine variables reflecting physical properties of genes, FUN(5) are five functional constraint variables, DUP is a categorical variable indicating duplication status, and ENV is the environmental relevance of each gene. Appendix Table S1 provides detailed description of these variables. The full model consists of one response and 16 predictor variables. To estimate the variation of PI explained by each predictor category, I compare the proportional reduction of explained variation (i.e., the difference in r^2) between the full and reduced models (removing all variables for a given category). For example, the reduced model (with 7 predictors) to estimate the combined effect of all variables in the physical property category is:

$$PI = FUN(5) + DUP + ENV + error.$$

In addition, I performed a standard fixed-effect ANCOVA with all 16 predictors, and the proportional variation explained by each predictor (after accounting for effects from all other predictors) was estimated via type III sum of squares. The partial regression coefficients between ENV and PI in the three independent analyses are also recorded to test the prediction that heterogeneous environmental selection maintains genetic variation. The statistical models were performed in JMP 8 (SAS, Cary, NC).

2.1.4 Analysis with different groups of environmental variables and accessions

To investigate whether my result would change with different types of environmental variables, I separated the 20 environmental variables into six groups: altitude, temperature, temperature variation, precipitation, precipitation variation, and temperature-by-precipitation interaction (Appendix Table S2). I calculated environmental relevance separately for the six groups and then re-did the full analysis for each group.

Among the 80 *A. thaliana* accessions being sequenced, those from Southern Russia and Central Asia showed substantial divergence from others (Cao *et al.* 2011). To confirm whether this major pattern of population structure affects my conclusion, I removed 15 accessions from these regions and re-did the whole analyses.

2.1.5 Gene ontology term enrichment of high environmental relevance genes

To identify which functional categories of genes may be most associated with environmental adaptation, based on the analysis using all environmental variables and all 80 accessions, I compared the enrichment of gene ontology (GO) terms between genes with the top 20% highest environmental relevance values ('top-20' hereafter) versus the other genes in my data set ('lower-80' hereafter). The comparison was performed separately for three sets of environmental relevance values calculated from total, synonymous, and nonsynonymous polymorphism data. I used the GO Slim terms defined by TAIR (Berardini *et al.* 2004), which provides a concise summary of many hierarchical GO terms into major categories. Within each of the three classification systems in GO (molecular function, biological process, and cellular component), I first determined whether the distribution of genes across all GO Slim terms is homogeneous

between top-20 and lower-80 genes. Because one gene may simultaneously correspond to several GO terms, I use permutation tests for significance. Each gene was randomly assigned to the top-20 or lower-80 groups in each permuted data set, and the significance of the observed data was then determined by comparing the chi-square value to 1,000 permuted data sets.

I observed that, in some cases top-20 genes are enriched in the unknown molecular function, unknown biological process, or unknown cellular component category. To specifically test this enrichment between top-20 and lower-80 genes, I used Cochran–Mantel–Haenszel test to compare the distribution of genes in known vs. unknown function categories, controlling for the three GO classification systems.

2.2 Results

2.2.1 Environmental relevance predicts genomic patterns of polymorphism

I first report the result with all 80 *A. thaliana* accessions and 20 environmental variables. As expected, physical properties (mostly associated with mutation rate) dominate the patterns of total and synonymous polymorphism among genes (8.1% for π and 6.8% for π_S, Figure 1A). On the other hand, nonsynonymous polymorphism is mostly influenced by functional constraints (5.9%) and secondly by physical properties (4.8%). Environmental relevance alone explains 1.3% of nonsynonymous polymorphism, about one-fifth of the effect from functional constraint (Figure 1A). Although duplication status was shown to be important in the divergence between *A. thaliana* and *A. lyrata* (Yang & Gaut 2011), it has minor effect on the level of polymorphism among *A. thaliana* genes.

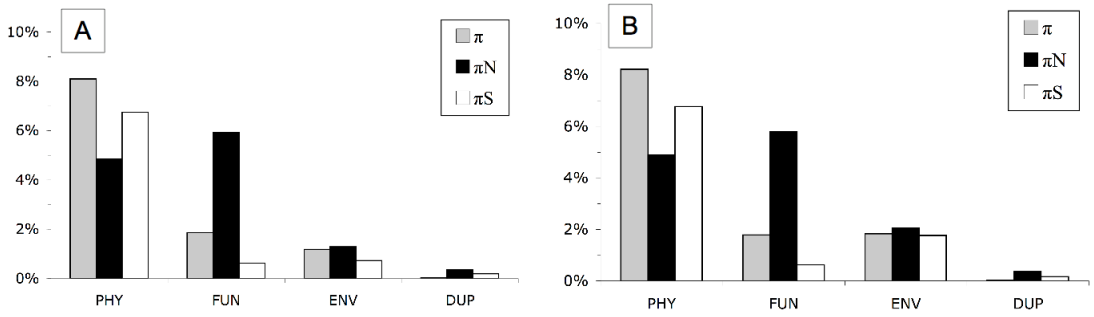


Figure 1: Proportional contribution of each predictor category to the variation of gene polymorphism in *A. thaliana*. There are four predictor categories (PHY – physical properties; FUN – functional constraints; ENV – environmental relevance; DUP – duplication status) and three separate measures of genetic polymorphism (π – total polymorphism; π_N – nonsynonymous polymorphism; π_S – synonymous polymorphism). (A) All 80 accessions (B) 65 accessions, excluding Russia and Central Asia.

At first glance, environmental relevance does not seem to be a major contributor to genetic variation, compared to physical properties or functional constraints. However, the large effects of physical properties and functional constraints represent the combined effects from multiple variables (9 for physical and 5 for functional, Appendix Table S1). Figure 2A shows the individual effects of each predictor variable, after accounting for the effect of all other predictors. Here, environmental relevance is the third most important predictor of total polymorphism (1.2% for π, after chromosome position and intron number), nonsynonymous variation (1.3% for π_N, after expression level and intron number), and synonymous polymorphism (0.7% for π_S, after chromosome position and intron number). Thus, environmental relevance is one of the most important among the 16 variables explaining polymorphism levels among *Arabidopsis thaliana* protein coding genes. In addition, the partial regression coefficients between environmental relevance and genetic polymorphism are positive in all three models. This is consistent with the prediction that spatially heterogeneous environmental selection maintains the polymorphism of responding genes. Furthermore,

while it is possible that relationships between environmental factors and genetic polymorphisms can be detected more easily at highly variable loci, greater statistical power at such genes cannot explain the consistently positive relationship that I find between environmental relevance and nucleotide variability.

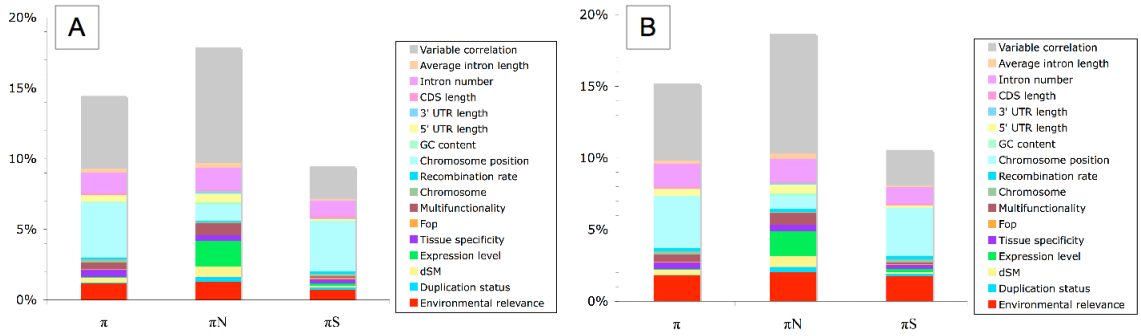


Figure 2: Proportional contribution of each predictor variable to the variation of gene polymorphism in *A. thaliana*. There are sixteen predictor variables (Appendix Table S1) and three separate measures of genetic polymorphism (π – total polymorphism; π_N – nonsynonymous polymorphism; π_S – synonymous polymorphism). The height of each bar represents total variation explained by the full model. Each colored box represents the partial variation explained by one factor, and the grey bars are variations explained by the correlation among predictor variables. (A) All 80 accessions (B) 65 accessions, excluding Russia and Central Asia.

2.2.2 Genes with high environmental relevance are enriched in unknown functions

Since environmental relevance is associated with the pattern of genome-wide polymorphism, I further tested whether the distribution of gene ontology (GO) terms differs between genes with high and low environmental relevance. I observed significant heterogeneity between the two groups of genes, especially when environmental relevance is calculated based on nonsynonymous polymorphism (Table 1). Interestingly, a further examination in each specific term showed that loci with high environmental relevance in the nonsynonymous data set are enriched for genes of unknown function. This enrichment is significant in Cochran–Mantel–Haenszel test (total polymorphism P =

$8x10^{-4}$; synonymous $P = 7x10^{-3}$; nonsynonymous $P = 2x10^{-7}$). Again, the enrichment in unknown functional categories is most obvious in nonsynonymous-based environmental relevance.

Table 1: Different distribution of genes with high vs. low environmental relevance values in gene ontology Slim terms. Shown are the *P* values from chi-square tests[a] between genes with top 20% and lower 80% environmental relevance values. Asterisks denote *P* values less than 0.05.

Data source[b]	Molecular	Biological	Cellular
Total	0.530	0.139	0.263
Synonymous	0.837	0.515	0.257
Nonsynonymous	0.003 *	0.022 *	0.003 *

a. Because one gene may simultaneously belong to multiple GO Slim categories, the *P* values were determined by 1,000 permutations.
b. Three sets of environmental relevance values were calculated based on total, synonymous, or nonsynonymous polymorphisms within each gene.

2.2.3 Consistent results were obtained form different subsets of data

With environmental relevance calculated from all 20 environmental variables, I have observed that environmental relevance explains variation in polymorphism level across genes (Figure 1A and 2A). The same pattern is also observed when environmental relevance was calculated from different groups of environmental factors, and each group exhibits distinct pattern. Consistent with previous result, in all cases environmental relevance explains more nonsynonymous variation than synonymous variation, and the partial regression coefficients between environmental relevance and gene polymorphism were positive.

To exclude the possible confounding effect from major population structure, I removed 15 genotypes from Southern Russia and Central Asia and re-did all analyses. The result with 65 accessions is qualitatively similar, and environmental relevance (from all 20 environmental variables) explains patterns of gene polymorphism (Figure 1B and

2B). The pattern still holds when environmental relevance was calculated from different groups of environmental variables, and all partial regression coefficients between environmental relevance and gene polymorphism are positive.

2.3 Discussion

Several approaches to genetic variation exist in biology: while quantitative genetics is focused on heritable variation for complex traits, molecular population genetics examines DNA or protein level polymorphism. Both fields have long histories in evolutionary genetics, and here I focus on intraspecific molecular polymorphism. In this area, three types of studies have investigated factors contributing to genetic variation: 1) At the single-locus level, genetic polymorphism may be maintained by balancing selection in heterogeneous environments (Hedrick 1986; Hedrick 2006). 2) At the whole-genome level, many non-ecological factors (such as recombination rate, GC ratio, or tissue-specific gene expression) can influence levels of polymorphism among genes (Andolfatto *et al.* 2011b; Bachtrog & Charlesworth 2002; Comeron *et al.* 1999; Flowers *et al.* 2012; Lercher & Hurst 2002; Pal *et al.* 2001). 3) Also at the whole-genome level, ecological factors can contribute to the average genome-wide divergence among populations or genotypes via local-adaptation-mediated reduction in immigrant or hybrid fitness, which may contribute to reproductive isolation (Lee & Mitchell-Olds 2011; Manel & Segelbacher 2009; Storfer *et al.* 2010).

In this study, I combined the first and second approaches: for every gene, I individually estimated its polymorphism and environmental relevance. I predict that, if a gene is more environmentally relevant (different alleles are associated with local adaptation to different environments), it is more likely to experience balancing selection, and thus it would be more polymorphic. Therefore, I examine whether environmental

relevance contributes to variation in polymorphism among genes while controlling for other aspects of gene function. The focus of this study is different from many studies in landscape genetics (the third category of studies, above), which use environmental differences to model the average genomic divergence among genotypes. In essence, the focus of this study is similar to the second category of studies, with a novel predictor variable (environmental relevance). This analytical approach became possible only recently, with the availability of whole-genome sequences of multiple accessions collected across a broad geographical range. Here I ask: how much do environmental factors influence the variation in polymorphism level of genes across the genome? To the best of my knowledge, I am not aware of other studies asking this biological question at a whole-genome level.

2.3.1 Environmental relevance predicts polymorphism among genes

In this study, I used 'environmental relevance' (the genetic variation within a locus explained by local environmental conditions, while controlling for population structure) to summarize the importance of each gene for environmental adaptation. I find that environmental relevance explains a significant portion of variation in functional polymorphism (π_N) among genes, and it is the third most important predictor among all 16 variables considered. Although environmental relevance is not the most important factor, it is remarkable that levels of environmental selection affect the pattern of polymorphism across the genome, considering the transient nature of environmental influences relative to the persistent long-term effects of physical properties, functional constraints, or duplication status of a gene. In addition, genes with high and low environmental relevance have significant differences in the distribution of gene ontology terms, and this difference is most obvious when environmental relevance was calculated

from nonsynonymous polymorphism. Environmentally relevant genes are enriched in the unknown functional categories (unknown biological processes, unknown molecular functions, and unknown cellular components). This may reflect the laboratory-based focus of most genomic studies and the paucity of genetic experiments in natural environment (Colbourne *et al*. 2011; Pena-Castillo & Hughes 2007), although other explanations are possible.

If most synonymous polymorphisms are neutral, then why does environmental relevance explain variation in π_S? It is possible that some synonymous mutations are selectively important (Hershberg & Petrov 2008; Kunstner *et al*. 2011). For example, a synonymous mutation might decrease the transcription or translation efficiency of a drought responsive gene, making an individual more susceptible to drought. The contribution of environmental relevance to π_S also may be due to the within-locus linkage disequilibrium between synonymous and nonsynonymous polymorphisms. Indeed, the correlation between π_S and π_N is 0.42 ($P < 0.001$) for these genes. In addition, the confounding effects of population structure, isolation by distance, and environmental variables may also affect the result. Although I have controlled for population structure when estimating environmental relevance and obtained similar results with a subset of accessions (which alleviates problems from major population structure), false positives or negatives may still be possible (Hancock *et al*. 2011).

Notice that environmental relevance quantifies the relationship between the functions of segregating alleles at a locus and local climatic conditions, and monomorphic genes were excluded from my analysis. Therefore, environmental relevance cannot detect genes that influenced environmental adaptation during the divergence between *A. thaliana* and *A. lyrata*. Consequently, in this study I restrict my analysis to the patterns of polymorphism but not divergence.

2.3.2 The polygenic nature of environmental adaptation

My observation of positive correlations between environmental relevance and gene polymorphism supports the hypothesis that spatially heterogeneous environmental selection may maintain genetic variation – if the function of a gene is more closely related to environmental adaptation, it may be more polymorphic. Population genetics theory states that balancing selection can maintain polymorphism of genes showing antagonistic pleiotropy, where genetic tradeoffs make alleles advantageous in one environment but unfit in another (Anderson *et al.* 2011b; Hedrick 1986; Mitchell-Olds *et al.* 2007). Under this view, my observed correlation between environmental relevance and level of polymorphism (Figure 1 and 2) might suggest an important role for antagonistic pleiotropy in environmental adaptation. On the other hand, a recent large-scale field experiment in *Arabidopsis thaliana* found that different loci control local adaptation in different locations (Fournier-Level *et al.* 2011). This may suggest conditional neutrality, in which an allele of a gene is adaptive in one location and neutral elsewhere (Anderson *et al.* 2011b; Hall *et al.* 2010). In the absence of trade-offs in local adaptation, a conditionally neutral allele is expected to gradually go to fixation in the absence of barriers of gene flow. Although there are many factors influencing the ability to detect antagonistic pleiotropy, especially the requirement of statistical power to detect the same loci in multiple environments (Anderson *et al.* 2012; Colautti *et al.* 2012), the results from Fournier-Level *et al.* (2011) still suggest that conditional neutrality is abundant in *Arabidopsis thaliana*. Therefore, the existence of antagonistically pleiotropic genes may not be the only cause of the observed relationship between environmental relevance and genetic variation. My observation may also reflect environmental

adaptation at many conditionally neutral loci, together with effects of limited gene flow and local demographic processes.

To date, most ecological genetic studies focus on single genes with large effects (Barrett & Hoekstra 2011; Hedrick 2006; Mitchell-Olds *et al.* 2007). However, several recent discussions emphasize the importance of polygenic adaptation, where evolution of a quantitative trait occurs via small changes in allele frequency at many loci (Pritchard *et al.* 2010; Rockman 2012). In this study, I have shown that environmental adaptation in *A. thaliana* shapes genome-wide variation, a pattern that would not occur if environmental adaptation involved only a few genes with large effects. Consistent with recent studies (Filiault & Maloof 2012; Fournier-Level *et al.* 2011; Hancock *et al.* 2011), my results may suggest a polygenic nature of environmental adaptation in this species.

2.3.3 Relationship to other studies

Environmental adaptation has long been known to influence patterns of genetic variation, especially in plants, which are sessile in nature. *Arabidopsis thaliana* and its relatives are good models not only for molecular genetics but also for investigating the role of environmental adaptation in shaping patterns of genetic variation (Gaut 2012; Lee & Mitchell-Olds 2011; Mitchell-Olds 2001; Rushworth *et al.* 2011; Weigel 2012). In addition to my analysis, two other studies have also investigated the gene-environment relationship in a whole-genome scale in *A. thaliana*. Fournier-Level et al. (2011) used genome-wide association study to identify SNPs influencing fitness components in four common gardens, and these SNPs together explained about 9 to 24% of local fitness variation. Hancock et al. (2011) identified SNPs associated with environmental factors, and these SNPs explain about 12 to 18% of local fitness in a common garden in France.

However, the focus of their analysis (variation in fitness explained by environmentally-relevant SNPs) is different from ours (variation in gene polymorphism explained by environmental relevance).

My method of calculating environmental relevance has parallels to Hancock et al. (2011) and a few other studies (Eckert *et al.* 2010a; Eckert *et al.* 2010b; Hancock *et al.* 2010), but instead of focusing on statistical significance of individual SNPs, I quantitatively estimated the proportion of genetic variation explained by environmental factors. My analysis asks a different biological question than these studies – rather than trying to identify specific SNPs or genes underlying environmental adaptation, here I ask whether and how much environmental adaptation shapes the variation in polymorphism levels across genes.

2.3.4 Conclusion

Environmental adaptation has long been known to affect genetic variation among genomes – among species, populations, or genotypes (Lee & Mitchell-Olds 2011; Manel & Segelbacher 2009; Storfer *et al.* 2010). Here, I estimate its influence on patterns of genetic variation among genes within a genome. Although environmental relevance is not the most important predictor in my investigation, my study introduces a new approach to analyzing genome-wide diversity data. My results suggest that the patterns of genome-wide polymorphism may be affected both by the innate properties of a genome and factors from the extrinsic environment.

2.4 Data availability

Data are deposited at Dryad doi:10.5061/dryad.q9p4s

3. Quantifying Effects of Environmental and Geographical Factors on Patterns of Genetic Differentiation

Elucidating the processes underlying the origin and maintenance of genetic variation in natural populations is a fundamental task in biology. The detailed characterization of genetic variation may reveal the demographic history and population structure of a species (Bryc *et al.* 2010; Novembre *et al.* 2008; Novembre & Stephens 2008; Platt *et al.* 2010; Sharbel *et al.* 2000; van Heerwaarden *et al.* 2011). This information also enables further analyses, such as association mapping for complex traits (Atwell *et al.* 2010; Huang *et al.* 2010; Yu *et al.* 2006) and the identification of genes that co-vary with specific environmental factors (Coop *et al.* 2010; Eckert *et al.* 2010a; Hancock *et al.* 2010; Manel *et al.* 2010), both aiming at understanding the genetic basis of local adaptation and the mechanisms underlying evolutionary changes. However, despite the fundamental importance of studying natural genetic variation and the availability of diverse methods of describing patterns of genetic variation, (Engelhardt & Stephens 2010; Gao *et al.* 2007; Jombart *et al.* 2009; Pritchard *et al.* 2000), still few studies have tried to investigate the relative contributions of factors affecting genetic differentiation across a species range.

It is widely acknowledged that genetic differentiation is strongly influenced by two processes: isolation by distance and differential local adaptation (Nosil *et al.* 2008; Nosil *et al.* 2005; Slatkin 1987; Wright 1931; Wright 1943). Under isolation by distance, the major factor limiting interbreeding is the physical distance, and populations diverge via genetic drift or clinal selective factors correlated with geographical distance. Because neighboring populations often have only minor differences in local environments (for example, day-length across latitude) and therefore minor reductions of immigrant or

hybrid fitness, substantial gene flow could occur among adjacent populations. As a consequence, the amount of gene flow is mainly restricted by geographical distance, and genome-wide divergence, as revealed by neutral genetic markers, is expected to be clinally correlated with geographical distance. In contrast, when migration occurs between nearby populations adapted to distinct environments, fitness of immigrants or hybrids may be reduced by natural selection (Nosil *et al.* 2005), and the resulting reduction of genetic exchange may facilitate or maintain genetic divergence (Thibert-Plante & Hendry 2010). Under this process, an abrupt change in local environment (for example, elevation change over a few kilometers) may cause substantial reduction of immigrant fitness, resulting in discrete, rather than continuous pattern of genetic differentiation. Therefore, the degree of genetic differentiation inferred from neutral loci is expected to correlate more with differences in local environment than with geographical distance. Although examples, theories, and reviews exist for the two processes (Engelhardt & Stephens 2010; Nosil *et al.* 2008; Nosil *et al.* 2005; Orr & Smith 1998; Rundle & Nosil 2005; Schluter 2001; Schluter & Conte 2009; Templeton 2008; Thibert-Plante & Hendry 2010; Wang & Summers 2010), few studies have jointly considered the relative importance of isolation by distance and local adaptation on genetic variation at a species-wide scale (but see Cushman *et al.* 2006; Freedman *et al.* 2010; Pease *et al.* 2009). By combining population structure estimation and niche modeling, here I statistically separate and quantify the effects of isolation by distance and local adaptation on genetic divergence patterns in the wild mustard species *Boechera stricta*.

For divergent selection to facilitate or maintain population differentiation, the environmental differences between lineages should be higher than within species or populations (Coyne & Orr 2004). Therefore, niche modeling has been used to identify

possible environmental factors contributing to population differentiation (Hübner *et al.* 2009; Kozak *et al.* 2008; Kozak & Wiens 2006; Nakazato *et al.* 2008). However, many environmental factors are highly correlated with each other and with geographical distance. To avoid spurious correlations, it is necessary to control for neutral processes when estimating the relationship between environment and genetic structure (Dyer *et al.* 2010; McCormack *et al.* 2010). Using geographical distance as a covariate, I investigate the contribution of environmental factors to independent axes of genetic differentiation in *Boechera stricta*. With isolation by distance as the null model (Novembre *et al.* 2008; Novembre & Stephens 2008; Platt *et al.* 2010; Sharbel *et al.* 2000), I attribute an axis of genetic differentiation to isolation by distance when only geographical distance has significant effect on this axis, or when I am unable to separate the effects of geography and environment due to their strong correlation. On the other hand, after controlling for geography, significant effects of environmental factors are expected when local adaptation drives or maintains genetic divergence.

Previous research has identified three major genetic groups within *Boechera stricta* (Song *et al.* 2009). A contact zone between the two most diverged groups (EAST and WEST) is found in the Rocky Mountains in Idaho, USA. During the last glacial maximum, this contact zone was mostly unsuitable habitat for this species or was covered by montane glaciers (Brunelle & Whitlock 2003; Hostetler & Clark 1997), suggesting that the current overlap is a zone of secondary contact after historical allopatry. Despite the existence of this contact zone, less than 3% of sampled genotypes were admixed (Song *et al.* 2009); nevertheless, fertile and healthy hybrids can be produced in the laboratory. Both observations suggest the existence of an extrinsic reproductive isolating mechanism other than isolation by distance or intrinsic hybrid inviability. If natural selection imposed by environmental factors contributes to divergence and prevents current

hybridization between the two genetic groups, I may be able to identify environmental factors as significant predictors of genotypic differentiation in both allopatric and sympatric regions. Additionally, the significant predictors should reflect the same underlying causal factors in both regions. In contrast, if reproductive isolation is caused by factors not related with environmental selection, while several environmental factors may be identified in the allopatric regions due to correlations among geography, genetic structure, and environments, no relationship between environmental factors and genetic divergence should exist in the contact zone.

In this study, I address the following questions: (i) What is the relative contribution of isolation by distance and environmental adaptation on independent genetic axes showing distinct patterns of differentiation? (ii) When environmental adaptation is inferred, can I further confirm this by identifying the same causal environmental variable in both allopatric and sympatric regions?

3.1 Materials and methods

3.1.1 Study species

Boechera stricta (Brassicaceae) is a wild perennial mustard species and a close relative of *Arabidopsis thaliana* (Mitchell-Olds 2001; Oyama *et al.* 2008). This species is native to western North America, occupying wide geographical, altitudinal, and environmental ranges (Song *et al.* 2006). Although polyploidy or apomixis occur in this genus (Schranz *et al.* 2005), *B. stricta* genotypes are predominantly diploid and sexual, with approximately 95% selfing rate (Song *et al.* 2006). With 46 genotypes, previous research has identified three genetic groups within this species (Song *et al.* 2009). To obtain more detailed information on genetic variation across the distribution range and to examine the multi-dimensional niche space of these genetic groups, I used 239

genotypes sampled from relatively un-disturbed environments in western North America.

3.1.2 Genotyping

Seeds of *Boechera stricta* were collected from about 250 locations across western North America and grown in the Duke Greenhouse. One individual was randomly chosen as representative of each collection site, a sampling scheme also used in previous studies (Manel *et al.* 2003; Platt *et al.* 2010). Because genetic variation within local populations is low (Song *et al.* 2006), this sampling scheme maximizes genetic diversity for a given sample size. Trichome morphology was examined for species confirmation (Rollins 1993), and the ploidy was estimated by flow cytometry (Partec, Munster, Germany) or the number of alleles in microsatellite loci, leaving 239 diploid individuals, each from different locations (Figure 3A). Seventeen microsatellite markers used in a previous study (Appendix Table S3, Song *et al.* 2006) were genotyped, and the PCR primers were modified for fluorescently-labeled M13-tailing (Boutin-Ganache *et al.* 2001). PCR products were processed with Applied Biosystems 3730, and alleles were called with GeneMarker (SoftGenetics, State College, PA, USA).

3.1.3 Genetic analysis

Two major methods have been employed to identify population structure (Engelhardt & Stephens 2010). Admixture-based models, such as STRUCTURE (Pritchard *et al.* 2000), estimate the proportion of each sample's genome derived from an ancestral genetic group. The other method, principal component analysis (PCA), uses multivariate statistics to depict the genetic structure and is free from many population genetics assumptions underlying STRUCTURE (Gao *et al.* 2007; Jombart *et al.* 2009). Although the two methods differ in model assumptions and methodologies, a recent

study (Engelhardt & Stephens 2010) showed that both approaches are special cases of matrix factorization with different constraints, and while admixture-based models are more suitable for discrete and partially admixed populations (such as secondary contact after historical allopatry), PCA is more useful with continuous patterns of differentiation (such as isolation by distance). Here, I employed advantages of both methods to investigate population structure within *Boechera stricta*. I have not employed methods that incorporate geographic information while assigning genetic structure (for example, Guillot *et al.* 2005) because my goal is to investigate the population structure based on genetic information per se, with the contributions from geography and environment to be estimated subsequently.

With STRUCTURE, three replicates were run for each k value (k = 2 and 3), following previous results (Song *et al.* 2006). I tried other k values (k from 4 to 10) but do not explicitly report the results here because I focused on the major genetic differentiation pattern in this study and other k values did not produce clear patterns (data not shown). Within each run, a total of two million iterations were conducted with the first one million as burn-in. In my definition, a genotype was regarded as belonging to a pure group if the Bayesian posterior probability was higher than 0.8. In addition, principal coordinate analysis (PCOA) was conducted with GenAlEx (Peakall & Smouse 2006). GenAlEx first calculated a pairwise genetic distance matrix based on the allele states. The PCOA axes and scores were then obtained by performing multidimensional scaling on this matrix. In theory, PCOA is equivalent to principal component analysis (PCA) if the initial distance matrix is calculated as Euclidean distance. Therefore, the PCOA result generated by GenAlEx can be viewed as the PCA of allele states within *Boechera stricta*.

I used customized Perl scripts to compare the range of F_{ST} values between genetic groups identified by STRUCTURE. Instead of bootstrapping among loci (Goudet 2001), my script performed bootstrap resampling of individuals within each genetic group. This approach gave us the advantage of retaining information from all loci while accounting for the spatial and temporal unevenness in field seed collection. One thousand bootstrapped data sets were generated by randomly resampling individuals from each group. Each data set was transformed into the input data format of FSTAT (Goudet 2001), and F_{ST} was calculated as the proportion of between-group to total genetic variation by package HIERFSTAT (Goudet 2005) in R (http://www.r-project.org/).

3.1.4 Environmental variables

Environmental variables with a resolution of 1 km^2 were downloaded from publicly available databases. Elevation and nineteen biologically-relevant climatic variables (Bioclim variables) were downloaded from WorldClim (Hijmans *et al.* 2005), and five topographical variables (aspect, slope, flow direction, flow accumulations, and compound topographical index) were downloaded from the HYDRO1k database of U.S. Geological Survey (USGS). Based on latitude and longitude, data layers were overlaid in ArcGIS 9 (ESRI, Redlands, CA, USA), and environmental factors from *Boechera stricta* collection sites were extracted with Hawth's Tools (http://www.spatialecology.com/htools/tooldesc.php). In addition, I manually measured 'distance to the nearest stream' with the resolution of one meter in Google Earth. Some environmental factors were excluded due to high correlation ($r > 0.9$ in some pairs of variables), finally leaving six climatic and four topographical variables (Appendix Table S4). The six climatic variables were chosen as the representatives of four major clusters in the hierarchical clustering analysis of climatic variables (data not

shown), and these variables represent the mean and variation of temperature and precipitation and their interaction effect. All environmental variables were log-transformed and standardized prior to statistical analyses due to their skewed distribution. Latitude and longitude were also transformed in the following regression-based but not distance-matrix-based analyses.

3.1.5 Niche modeling

The genetic analyses identified three major genetic groups, forming two contrasting patterns of genetic differentiation within *B. stricta* - the discrete EAST-WEST and the continuous NORTH-SOUTH divergence. To dissect the effect of natural selection (environment, isolation by adaptation) and genetic drift (geography, isolation by distance) on the two distinct patterns of genetic differentiation, I first performed Mantel tests to assess the correlations among genetic, environmental, and geographic distance matrices. Pairwise genetic distance among genotypes was calculated by GenAlEx (described above), and the environmental distance matrix was obtained by calculating the Euclidean distance between pairs of collection sites from the ten environmental variables. The great-circle geographic distance, the nearest distance between two points on the Earth surface, was obtained by package 'fields' (http://CRAN.R-project.org/package=fields) of R using un-transformed latitude and longitude values. I did not employ more complex geographical distance measurements, such as least-cost path (Storfer *et al*. 2007), because the dispersal distance of *B. stricta* is only a few meters (Mitchell-Olds, personal observation), a much smaller scale than the resolution of the environmental data layers used in this study. To account for the correlation among these three distance matrices, partial Mantel tests were further conducted to estimate the contribution of environmental distance to genetic distance while accounting for the effect

of geographic distance. Both Mantel and partial Mantel tests were performed with package 'vegan' (http://CRAN.R-project.org/package=vegan) of R, and significance was determined by 1000 permutations.

However, while partial Mantel tests can examine the significance of correlations among matrices, recent reports (Balkenhol *et al*. 2009; Legendre & Fortin 2010) show that such distance-based methods have less statistical power and do not correctly estimate the amount of total variation explained by predictor variables. To quantitatively estimate the relative influence of genetic drift and environmental adaptation on genetic differentiation, I combined the genetic principal component analysis (PCA) and geographical and environmental discriminant function analysis (DFA) into a multiple regression framework:

$$GEN = GEO + ENV + GEO*ENV,$$

where GEN, GEO, and ENV are the genetic, geographic, and environmental 'scores' of each genotype. Each genotype has its unique positions in the multivariate genetic, geographic, and environmental spaces, and the corresponding scores are projections on axes that best distinguish genetic groups in each multivariate space. Notice that I employed DFA rather than PCA for geographical and environmental factors because PCA axes only capture most variation among all samples, but not necessarily the geographical or environmental differences between genetic groups. These scores provide a metric to quantify how geographical and environmental factors predict genetic variation between *Boechera* genetic groups. Thus, the GEN score is simply the projection on the genetic PCA axes. For GEO and ENV, discriminant function analyses (DFA) were first performed between the inferred genetic groups being compared, and the geographic and environmental score of every individual (including hybrids) was calculated from the coefficients of each variable identified by DFA. DFA was performed with the 'MASS'

package (http://CRAN.R-project.org/package=MASS) in R, and multiple regression was performed with JMP 8 (SAS, Cary, NC). Proportion of genetic variation explained by GEO or ENV, after accounting for the effect of each other, was calculated from Type III sum of squares. The entire analysis was conducted separately for the EAST-WEST and NORTH-SOUTH comparisons. I chose genetic PCA values rather than STRUCTURE posterior probabilities as responses because PCA axes are independent by definition. This allowed us to model the contribution from environment and geography to one genetic differentiation pattern (e.g., EAST-WEST, PC1) with minimal interference from the other pattern (e.g., NORTH-SOUTH, PC2). In contrast, the posterior probabilities given by STRUCTURE are constrained so that all values sum to 1. Nevertheless, using STRUCTURE posterior probability as response variable yields qualitatively similar results (data not shown).

To further identify whether the two categories of environmental factors (climatic and topographical, Appendix Table S4) have different contributions to the spatial distribution of 'pure genotypes' in sympatric and allopatric regions, a similar regression analysis was performed by separating the ENV factor into CLIM (six climatic variables) and TOPO (four topographical variables):

$$GEN = GEO + CLIM + TOPO + GEO*CLIM + GEO*TOPO + CLIM*TOPO + GEO*CLIM*TOPO.$$

In these regression analyses, I were able to quantitatively estimate the contribution of each predictor variable to the genetic structure of *B. stricta* by using the genetic PCA scores as response variables. However, PCA scores reflect the genetic variation both within and between genetic groups. Therefore, I used multiple logistic regression to identify specific environmental variables contributing mainly to the between-group differentiation, with 'pure genetic group' (a binary categorical variable)

as response and twelve factors (latitude, longitude, and ten environmental factors) as predictor variables. Because putting all predictors in a full model simultaneously would cause over-fitting of the model, I first used automatic forward selection of predictors in JMP 8 and then manually removed non-significant variables. I set the alpha value for each iteration of the forward selection process as 0.01, a somewhat stringent significance criterion, to prevent type I error generated during multiple steps of model comparison and to limit the number of predictor variables in the final model.

In analyses involving the comparison between EAST and WEST genetic groups in the sympatric or allopatric regions, three collections from central Montana (MacDonald Pass Trailhead, Elkhorn, and Brackett Creek) were removed because, due to limited sampling, I were not certain about the existence of a contact zone there.

3.2 Results

3.2.1 Genetic structure in *Boechera stricta*

My larger sample confirms previous results (Song *et al.* 2009), in that STRUCTURE identified three major groups (NORTH, SOUTH, and WEST) when $k = 3$ (Figure 3A). When setting $k = 2$, NORTH and SOUTH merged into one group while WEST remained distinct. This result was consistent with PCA (Figure 4). While the PC axis explaining the largest fraction (40.43%) of genetic variation distinguished WEST versus the two other groups, the axis accounting for 17.23% of the variation separated NORTH from SOUTH groups. Both results were consistent with previous findings that WEST was most diverged from the two other genetic groups. Therefore, NORTH and SOUTH lineages will be referenced collectively as the 'EAST' genetic group at some points in the following discussion. This pattern was also supported by the F_{ST} distribution from bootstrap

resampling of 'pure genotypes' (mean F_{ST} between EAST and WEST = 0.30, with 95% CI from 0.28 to 0.32; NORTH and SOUTH = 0.18, with 95% CI from 0.16 to 0.21).

Noticeably, NORTH and SOUTH groups are distributed continuously along the second principal component axis (PC2, Figure 4). In contrast, although most of the WEST genotypes were sampled in the Idaho contact zone, they were genetically distinct from the NORTH group in PC1, suggesting mechanisms other than geographic isolation may contribute to their genetic differentiation. Therefore, my niche modeling focused on two distinct comparisons: a species-wide comparison of EAST vs. WEST, and a NORTH vs. SOUTH comparison within the more continuously distributed EAST group.

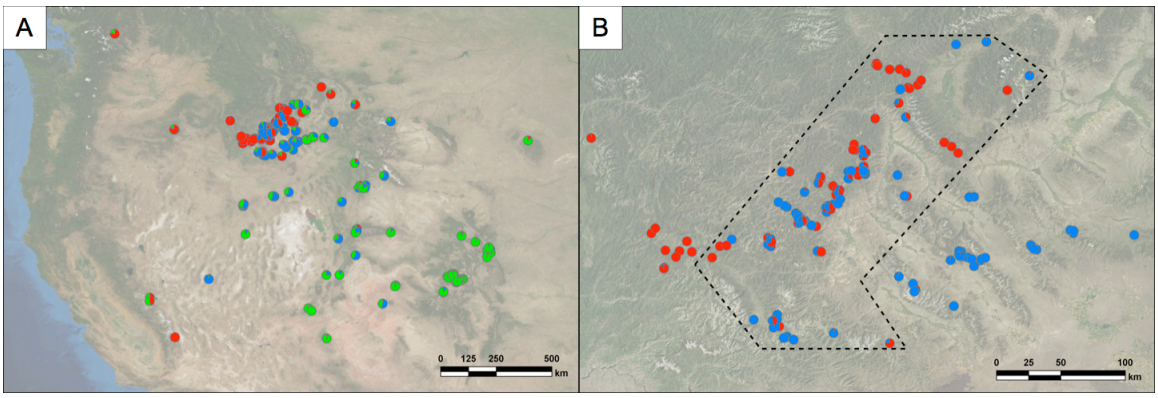


Figure 3: Collection sites and STRUCTURE results for *Boechera stricta*. Each pie chart represents one individual randomly chosen from one location. Different colors in each pie chart represent STRUCTURE posterior probabilities that the individual belongs to each genetic group. A) The distribution of three genetic groups across western North America. Red = WEST; blue = NORTH; green = SOUTH. Notice the narrow contact zone between WEST and EAST (comprised of NORTH + SOUTH), and the clinal distribution between NORTH and SOUTH genetic groups. B) The distribution of WEST and EAST genetic groups around the contact zone. Red = WEST; blue = EAST. Region encompassed by the dashed line is regarded as 'sympatric zone'.

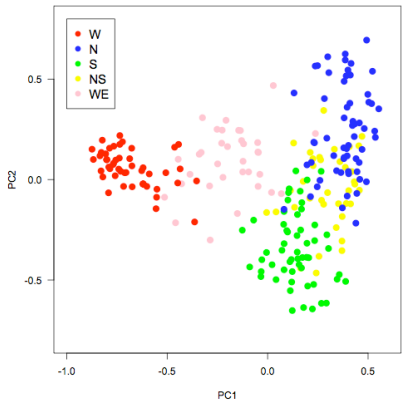


Figure 4: Genetic principal component analysis (PCA) of 239 *Boechera stricta* accessions. PC1 explains 40.4% and PC2 explains 17.2% of total genetic variation. Accessions were colored based on STRUCTURE results with k = 3, and a genotype belongs to a 'pure genetic group' (W = WEST, N = NORTH, S = SOUTH) only when the corresponding posterior probability is higher than 0.8. 'NS' and 'WE' denote NORTH-SOUTH hybrids and WEST-EAST (EAST = NORTH + SOUTH) hybrids, respectively. Notice the distinct distribution patterns between WEST-EAST along PC1 (discrete) and NORTH-SOUTH along PC2 (continuous).

3.2.2 Contribution of environment versus geography to population structure

Mantel tests showed that for both EAST-WEST and NORTH-SOUTH divergence, all three distance matrices (genetic, environmental, and geographic) were highly correlated (all $P \leq 0.002$). In partial Mantel tests, environmental distance remained a significant predictor of genetic distance after accounting for geographic distance only in the EAST-WEST ($P = 0.001$) but not in the NORTH-SOUTH comparison ($P = 0.185$).

The genetic PC1 values (from all samples) correspond to the genetic scores of EAST-WEST divergence. Within the EAST group, PC2 scores correspond to NORTH-SOUTH divergence (Figure 4). In both cases, quantitative results from multiple regression revealed similar pattern as the partial Mantel tests (Table 2). While the full models explained comparable amounts of total genetic variation in both contrasts between

groups (42.77% for EAST-WEST and 50.84% for NORTH-SOUTH), environmental factors gave significant prediction only for EAST vs. WEST divergence (21.60%, $P < 0.001$) but not between NORTH and SOUTH (0.87%, $P = 0.107$), while controlling for geographic effect. In the NORTH-SOUTH comparison (Table 2), any predictor only explained a small portion of genetic variation after accounting for the contribution of other predictors. This reflects the strong correlation between geography and environment in the NORTH-SOUTH comparison (Pearson's correlation coefficient r = 0.95, $P < 0.001$). In contrast, this correlation was less pronounced (r = 0.66, $P < 0.001$) in the EAST-WEST divergence pattern.

Table 2: Proportion of genetic variation explained by environmental (ENV) and geographical (GEO) effects or their interaction in the EAST-WEST (species-wide, genetic PC1) and the NORTH-SOUTH (within-EAST, genetic PC2) genetic divergence patterns.

	EAST-WEST		NORTH-SOUTH	
Predictors	Proportion explained (%)	*P* value	Proportion explained (%)	*P* value
Full model	42.77	< 0.001	50.84	<0.001
-ENV	21.60	< 0.001	0.87	0.107
-GEO	0.06	0.608	1.15	0.065
-ENV*GEO	4.80	< 0.001	0.74	0.139
Error	57.23		49.16	

These results suggest that isolation by distance played a fundamental role in the divergence between NORTH and SOUTH genetic groups. On the other hand, when controlling for geographical factors, the importance of environmental selection was highly significant in EAST-WEST divergence. Next, I focused on identifying the specific environmental factors contributing to the ecological differentiation between EAST and WEST lineages.

3.2.3 Identifying sources of environmental selection

By separating ten environmental variables into six climatic and four topographical variables (Appendix Table S4), similar regression analyses identified the relative contribution of the two categories of environmental variables to the genetic divergence between EAST and WEST genetic groups in sympatric and allopatric regions (Figure 3B, Table 3). In the allopatric region, climatic factors explained 8.17% ($P <$ 0.001) of total genetic variation, about three times the contribution of topographical factors (2.66%, $P = 0.001$). These results were reversed in the sympatric region, where topographical factors predicted 5.68% ($P = 0.002$) of EAST-WEST genetic divergence, but climatic factors alone had little effect (0.67 %, $P = 0.278$).

Table 3: Proportion of EAST-WEST (genetic PC1) genetic variation explained by climatic (CLIM), topographical (TOPO), geographical (GEO), or the interaction effects in the allopatric or sympatric regions.

	Allopatric		Sympatric	
Predictors	Proportion explained (%)	*P* value	Proportion explained (%)	*P* value
Full model	71.69	<0.001	41.39	<0.001
-CLIM	8.17	<0.001	0.67	0.278
-TOPO	2.66	0.001	5.68	0.002
-GEO	3.18	<0.001	3.32	0.017
-CLIM*TOPO	5.63	<0.001	0.44	0.381
-CLIM*GEO	0.35	0.231	<0.01	0.995
-TOPO*GEO	0.15	0.430	0.01	0.878
-CLIM*TOPO*GEO	3.25	<0.001	1.27	0.136
Error	28.31		58.61	

Logistic regression confirmed the importance of climate in allopatry and topography in sympatry for the genetic divergence between EAST and WEST lineages (Table 4). In the allopatric region, while most environmental variables differed significantly between EAST and WEST genotypes in simple logistic regression (data not shown), only 'winter precipitation' (a climatic variable, $P < 0.001$) and longitude ($P <$

0.001) were significant in multiple logistic regression. For sympatric genotypes, 'distance to the nearest stream' (a topographical variable, $P < 0.001$) and latitude ($P < 0.001$) were significant in multiple logistic regression. Noticeably, this pattern was also reflected by the significant interaction effect between environment and geography in the previous multiple regression (Table 2).

Table 4: *P* values based on likelihood ratio tests in multiple logistic regressions on EAST-WEST genotypes (a binary response variable) in the allopatric or sympatric regions.

Predictors[a]	Allopatric	Sympatric
Winter precipitation	<0.001	
Distance to the nearest stream		<0.001
Latitude		<0.001
Longitude	<0.001	

a. Only significant predictors in multiple logistic regression are reported. Refer to Appendix Table S4 for a full list of all variables used.

3.3 Discussion

Recent years have witnessed the rise of landscape genetics, a research area combining molecular population genetics and landscape ecology (Manel *et al.* 2003; Storfer *et al.* 2007; Storfer *et al.* 2010). As summarized by Storfer *et al.* (2007), the study of landscape genetics includes several major research categories, using a broad range of approaches to examine geographical patterns of genetic variation. Nevertheless, most studies focus on the effects of geographical and environmental factors on current gene flow among local populations. Phylogeography, on the other hand, differs from landscape genetics in the broader spatial and longer temporal scale considered (Manel *et al.* 2003; Storfer *et al.* 2007). However, despite its larger spatio-temporal scale, phylogeographic analyses to date have concentrated primarily on the effect of historical neutral processes on the pattern of genetic variation, and the role of environmental

adaptation is not often considered (Hickerson *et al.* 2010). Here I combine the consideration of environmental factors from landscape genetics and the broad spatio-temporal scale of phylogeography in order to separate the effects of neutral processes and environmental adaptation on the species-wide pattern of genetic variation. I regard the pattern of genetic variation within *Boechera stricta* as created via the long-term accumulation of reproductive isolation among the three major genetic groups, rather than the result of recent gene flow between local populations. Hence, this research has larger spatio-temporal scale than most landscape genetics studies. While most studies investigating within-species genetic variation are mainly exploratory rather than hypothesis driven (Storfer *et al.* 2010), my approach specifically tests whether different patterns of genetic differentiation (distinct or continuous) are driven by heterogeneous contributions from geography and environment.

In this study, I investigated the population structure of *Boechera stricta* and then performed sequential tests to examine the role of environmental factors in shaping the pattern of species-wide genetic variation. First, I investigated the relative contributions of isolation by distance and environmental adaptation to two contrasting patterns of genetic divergence: EAST-WEST (discrete) and NORTH-SOUTH (continuous). After the importance of environmental adaptation was demonstrated in the EAST-WEST divergence, I then examined the allopatric versus sympatric portions of the species range in order to infer the contributing environmental factors.

3.3.1 Contribution of environment versus geography to population structure

Many studies have investigated the evolutionary processes that drive population differentiation (Hübner *et al.* 2009; McCormack *et al.* 2010; Nakazato *et al.* 2008; Pease *et al.* 2009). While most examples focus on either isolation by distance or environmental

adaptation, my study is one of the first to jointly estimate the relative influence of these two forces on multivariate genetic differentiation at a species-wide level, and to identify distinct patterns at different levels of population structure (also see Cushman *et al.* 2006; Freedman *et al.* 2010; Pease *et al.* 2009). Here, I used neutral molecular markers to represent the pattern of genomic background divergence and used this estimated divergence as a surrogate for the historical accumulation of reproductive isolation. Therefore, my goal in this study is to use the degree of reproductive isolation as response variable and estimate the effect from environmental adaptation, using isolation by distance as background control. This is in contrast to many other studies, which controlled for population structure when searching for phenotype-environment correlation (Keller *et al.* 2009; Keller & Taylor 2008), gene-environment association (Coop *et al.* 2010; Eckert *et al.* 2010a; Freedman *et al.* 2010; Hancock *et al.* 2010), or gene-phenotype association both in the whole-genome (Yu *et al.* 2006) and the single gene level (Korves *et al.* 2007; Samis *et al.* 2008). Specifically, using a multiple regression framework, I tested the contribution from isolation by distance and environmental adaptation at the two hierarchical levels of genetic differentiation and found heterogeneous effects from the two contributing factors across the species range. While isolation by distance alone is sufficient to explain the moderate and continuous NORTH-SOUTH divergence, environmental variables show larger contribution than geographical factors in the discrete divergence between EAST and WEST. Thus, when environmental adaptation is involved, it may create or maintain higher genetic divergence than isolation by distance alone.

In this study, I incorporated genetic principal component analyses (PCA) and discriminant function analyses (DFA) of multivariate geographical and environmental data sets into a multiple regression framework. This regression-based approach enables

the quantitative estimation of genetic variation explained by environmental and geographic factors and their interaction effects, which could not be correctly estimated by partial Mantel test and its derivatives (Balkenhol *et al.* 2009; Legendre & Fortin 2010; Manel *et al.* 2003). Similar regression-based approaches have examined the contributions of environment and geography to genetic variation (e.g., Sork *et al.* 2010), and the dimensions of environmental variables were usually reduced via PCA rather than DFA, and multiple PCA axes were often used. Instead, I examined factors contributing to each of the two hierarchical levels of population structure, and therefore, I chose DFA in order to identify the axis best distinguishing the environmental differences between genetic groups in the hierarchical level being investigated. In addition, my study may be the first to demonstrate the interaction effect between geography and environment in shaping natural genetic variation: In *Boechera stricta*, the significant GEO*ENV interaction effect in Table 2 is further confirmed by the finding that different environmental variables contribute to the EAST vs. WEST divergence in sympatric and allopatric regions (Table 3 and 4).

The possibility that environmental factors contribute to the NORTH-SOUTH divergence pattern in *B. stricta* cannot be ruled out, however. Indeed, several studies have found phenotypic divergence and local adaptation among populations along latitudinal gradients (Arthur *et al.* 2008; Colautti *et al.* 2009; Hopkins *et al.* 2008; Leinonen *et al.* 2009; Mitchell-Olds *et al.* 2007; Montague *et al.* 2008; Stinchcombe *et al.* 2004). As shown by several examples (Hübner *et al.* 2009; Platt *et al.* 2010), when both environmental variables and axes of genetic differentiation are highly correlated with geography, it is difficult to statistically identify the causal factors. This is analogous to the well-known issue of population structure in genome-wide association studies (Bergelson & Roux 2010; Marchini *et al.* 2004). Like association studies, which control

false positives by incorporating population structure into the model (Yu *et al.* 2006), here I employ a similar approach by using isolation by distance as my null model (Novembre *et al.* 2008; Novembre & Stephens 2008; Platt *et al.* 2010; Sharbel *et al.* 2000) and then examine the effect of environmental variables on genetic differentiation while controlling for geographical factors. The importance of performing such controls is illustrated by a recent study (McCormack *et al.* 2010), in which, contrary to previous results not accounting for geographical effects, no niche divergence was detected between taxa after such controls were implemented. Similarly, another study (Zellmer & Knowles 2009) used landscape data from three different time periods to model concurrent genetic differentiation among frog populations, and after controlling the effect from each other, they found only contemporary landscape features, rather than historical ones, significantly predict genetic differentiation. My approach is conservative, since I infer the existence of environmental adaptation only when environment factors explain significant genetic variation in addition to what is already accounted for by geography. If the effects of geography and environment cannot be separated due to their strong correlation, I conservatively attribute genetic differentiation patterns to isolation by distance. Thus, in some circumstances a strong correlation between environment and geography may obscure causal influences of natural selection due to environmental factors.

Nevertheless, even if the NORTH-SOUTH divergence in *B. stricta* is under natural selection from undetected clinal environmental factors, such selection may not cause obvious immigrant or hybrid inviability between adjacent local populations. Under such clinal pattern, although obvious local adaptation may be detected between distant populations (Etterson 2004; Leinonen *et al.* 2009), there may be little environmental difference among nearby populations. For example, if day length mediates local

adaptation between NORTH and SOUTH genetic groups, the limited variation in day length between neighboring populations will cause little reduction in gene flow. This clinal pattern is in sharp contrast to the EAST-WEST divergence, where two genetically distant populations reside in environmentally distinct locations separated only by a few kilometers. Indeed, given the predominant role of isolation by distance in the NORTH-SOUTH divergence of *Boechera stricta* and in *Arabidopsis thaliana*, a close relative having similar breeding system (Platt *et al.* 2010; Sharbel *et al.* 2000), my finding that environmental selection played a large role in the discrete EAST-WEST divergence pattern further illustrates the importance of environmental selection in facilitating or maintaining genetic divergence.

3.3.2 Identifying sources of environmental selection

After the importance of local environment was demonstrated in the EAST-WEST divergence, I examined possible environmental factors underlying this divergence pattern to further confirm the role of environmental variables and the GEO*ENV interaction effect in shaping genetic variation in *B. stricta*. If natural selection by environmental differences were driving phenotypic differentiation during historical allopatry and maintaining reproductive isolation after secondary contact, local genotypes should be consistently associated with predictable environmental conditions. I found similar underlying mechanisms influencing genetic differentiation in allopatric and sympatric regions (Table 3 and 4). In the allopatric region, WEST genotypes occur in habitats with higher winter snowfall, which provides greater water availability in summer. In the sympatric area, WEST genotypes occur in riparian sites near streams, where they may experience higher and more consistent levels of soil moisture. In contrast, EAST genotypes occur on high elevation mountain slopes where ephemeral moisture is

supplied by rainfall and snowmelt in spring and early summer. Therefore, during historical allopatry, climatic differences likely drove the phenotypic divergence between the two genetic groups. Upon secondary contact, this trait divergence causes the two genetic groups to occur in distinct habitats based on topography, because climatic variation in the contact zone is low relative to the species range across western North America. In addition, the importance of controlling for geographical factors is again emphasized. While most variables are significant predictors of local EAST-WEST genotypes in simple logistic regression (data not shown), the putatively most important factors would be identified only when the effect of geography (latitude or longitude) is controlled in multiple logistic regression (Table 4).

The possibility cannot be totally ruled out, however, that other correlated factors (such as local fauna or other plant competitors) contribute to local adaptation of EAST and WEST genotypes, rather than direct effects of water availability. Nevertheless, the importance of soil moisture is supported by preliminary greenhouse and field observations (Lee and Mitchell-Olds, unpublished data). Phenotypic differentiation is significant in a common greenhouse environment, where EAST genotypes show higher tolerance of drought. Also, observations in the field suggest that in their native moist riparian sites, WEST genotypes have greater fruit production than EAST genotypes, possibly due to the longer flowering duration and larger vegetative size. In contrast, slower flowering of WEST genotypes makes them more susceptible to the late summer drought typical of EASTERN habitat on montane slopes. In addition to reciprocal immigrant inferiority (Nosil *et al.* 2005), their difference in flowering time may also reduce the chance of hybridization, causing assortative mating. Although the genome-wide neutral genetic divergence between EAST and WEST may have arisen by genetic drift during historical allopatry, natural selection can be the force currently maintaining

such differentiation in the sympatric zone, given the lack of intrinsic hybrid incompatibility.

Recently, methods have been developed to predict species distribution based on inferred environments at collection sites (Phillips *et al.* 2006; Thomassen *et al.* 2010). However, my results show that even if the same underlying factor (water availability) determines the distribution of EAST and WEST lineages in *B. stricta*, distinct environmental variables ('winter precipitation' or 'distance to nearest stream') may represent this underlying factor in different geographical regions. Therefore, in this study I do not attempt to predict the distribution of these genetic groups. In addition, the lack of a 'distance to the nearest stream' data layer with the resolution in meters may compromise the accuracy and statistical power of such modeling methods. I suggest that future studies involving environmental niche modeling should incorporate understanding of the biology and ecology of the target species before applying a universal model to continental-scale distributions.

3.3.3 Conclusion

This study jointly estimates the relative contribution of isolation by distance versus environmental adaptation to genetic divergence across a species range. In *B. stricta*, the EAST-WEST axis of genetic differentiation, incorporating the joint influences of isolation by distance and environmental adaptation, explains more species-wide genetic variation than the NORTH-SOUTH genetic axis, where only the effect of isolation by distance is significant. In addition, my inference of environmental adaptation contributing to EAST-WEST divergence also is supported by preliminary observations from laboratory and field. In summary, this research emphasizes the role of ecological factors in the creation and maintenance of genetic differentiation.

3.4 Data availability

Data are deposited at Dryad doi:10.5061/dryad.6rs51

4. Complex trait divergence contributes to environmental niche differentiation in ecological speciation of *Boechera stricta*

Natural selection and neutral processes are two major forces contributing to genetic differentiation and reproductive isolation among lineages (Slatkin 1987). Ecological factors may contribute to genetic divergence via differential local adaptation, which reduces immigrant or hybrid fitness and causes reproductive isolation. This process, termed 'ecological speciation' (Sobel *et al.* 2010) or 'isolation by adaptation' (Nosil *et al.* 2008), is an area of active research. If the trait under selection or the source of selection is clear, this may provide starting points for investigation; examples include salt tolerance in *Mimulus guttatus* (Lowry *et al.* 2008) and host plant adaptation in insects (Funk *et al.* 2011; Via *et al.* 2000), among others. However, in many species the trait under selection or the source of selection is unclear.

One possible solution comes from niche modeling and landscape genetics (Manel *et al.* 2003; Storfer *et al.* 2010), which allows the identification of specific environmental factors correlated with genetic differentiation. Often, however, investigations do not advance beyond correlational inference, and the traits under selection remain ambiguous even after possible environmental causes of natural selection are identified statistically. The scarcity of empirical tests of niche modeling predictions may in part reflect the difficulty of conducting manipulative experiments in many species. Nevertheless, verification of correlational inferences requires empirical evidence.

Boechera stricta is a short-lived perennial mustard native to the Rocky Mountains in North America and is an emerging model for ecological genetics (Prasad *et al.* 2012; Rushworth *et al.* 2011). In a previous study (Lee & Mitchell-Olds 2011), I identified two subspecies of *B. stricta* ("EAST" versus "WEST"), which show clear differentiation for

neutral molecular markers, as well as for ecologically important traits (below). Crosses between these subspecies generate fertile recombinant inbred lines, which sometimes show subtle hybrid breakdown (Anderson *et al.* 2011a). Among *B. stricta* populations in the western United States, the primary axis of genetic differentiation is between these EASTERN versus WESTERN subspecies, and the EASTERN subspecies can be subdivided along a NORTHERN to SOUTHERN continuum. While the genetic differentiation between NORTHERN and SOUTHERN groups primarily reflects isolation by distance, the divergence between EASTERN and WESTERN subspecies suggests environmental adaptation, independently from the effects of geographic distance (Lee & Mitchell-Olds 2011). Further analysis showed that local water availability is the most important factor explaining the habitat segregation between the two groups, and WESTERN genotypes mostly inhabit environments with more constant and abundant water supply. Given that fertile hybrid genotypes exist in the field and can be generated in the greenhouse, intrinsic hybrid inviability or infertility may not be the main form of reproductive isolation between these two subspecies. Therefore, the EAST-WEST divergence pattern may represent a case of incipient ecological speciation (isolation by adaptation), where the amount of gene flow in the secondary contact zone is reduced by differential local adaptation. I hypothesized that local water availability may be an important selective agent decreasing the fitness of immigrants or hybrids, causing reproductive isolation and genetic differentiation (Lee & Mitchell-Olds 2011).

In this study, I test the prediction that EASTERN and WESTERN genotypes are diverged in some traits associated with local water availability. Specifically, the EASTERN subspecies should exhibit phenotypes adaptive in their drier native environments, while the WESTERN subspecies should have phenotypes conferring higher fitness in wet riparian environments. By estimating the trait divergence from 24

accessions in the greenhouse and comparing their univariate and multivariate Q_{ST} to the empirical distribution of SNP F_{ST}, I show that the two genetic groups mainly utilize morphology and phenology, but not physiology, for their adaptation to differential water availability.

4.1 Materials and methods

4.1.1 Plant material

Throughout this study I will use the terms EAST and EASTERN interchangeably, and likewise for WEST and WESTERN. I focus my study on the vicinity of the EAST-WEST contact zone in Idaho, USA (Lee & Mitchell-Olds 2011) because this is the region where differential local adaptation is most likely to oppose gene flow. I chose 24 core populations representing the four combinations of 'EAST vs. WEST subspecies' and 'allopatric vs. sympatric geographical zones' (Figure 5). I randomly sampled one genotype from each population because *Boechera stricta* has low genetic variation within local populations (Song *et al.* 2006). The 24 genotypes incorporate most of the genetic and geographical variation around the contact zone (Appendix Figure S1).

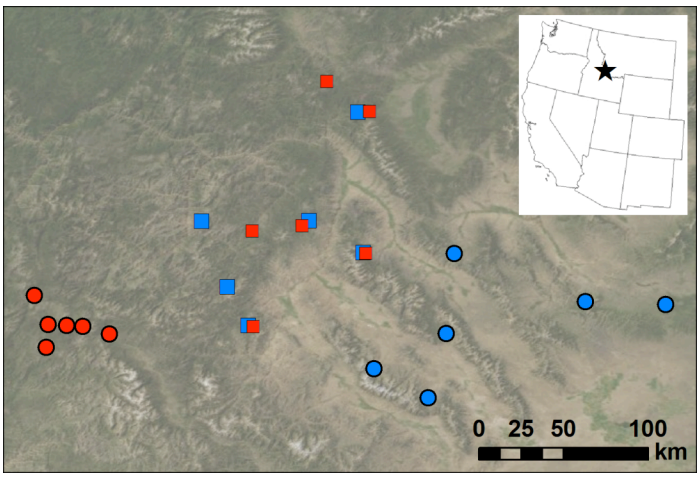


Figure 5: Collection sites of 24 genotypes used in this study. The region is denoted as a black star on the state boundary map. Blue circles – allopatric EAST.

Blue squares – sympatric EAST. Red circles – allopatric WEST. Red squares – sympatric WEST.

Each genotype was grown in the greenhouse for at least one generation to reduce maternal effects. Because *Boechera stricta* has a high self-fertilization rate in natural environments (Song *et al.* 2006), selfed seeds obtained from each genotype can be used as replicates in the three following independent experiments. In addition, this selfed family design has been shown to be better than a half-sib crossing design when estimating trait Q_{ST} in highly selfing species (Goudet & Buchi 2006). Seeds in all experiments were stratified in 4 degrees C for four weeks and planted in Ray Leach SC10 'Cone-tainers' (21 cm in depth and 3.8 cm in diameter, Stuewe & Sons Inc., Tangent, OR, USA). Following my standard procedures for growing *B. stricta* in the greenhouse, the lower 80% of each Cone-tainer was filled with Fafard 4P Mix soil (Conrad Fafard, Agawam, MA, USA), and the top 20% was filled with Sunshine MVP soil (Sun Gro Horticulture, Vancouver, BC, Canada). All experiments were conducted in the same room of the Duke University Greenhouse, with 16-hour day length (6 am to 10 pm), daytime temperature of 65 to 70 degrees F, and nighttime temperature of 55 to 60 degrees F. Because most traits were non-normally distributed, measured traits (Appendix Table S5) were log transformed and standardized to improve normality and provide a more accurate estimate of variance components.

4.1.2 Experiment 1. Short-term drought manipulation and phnology

A total of 576 individuals were arranged into 12 randomized complete blocks. The 48 individuals within each block were composed of the 24 core genotypes, with two individuals from each genotype subjected to different water regime treatments (well-watered or drought). My planting approach imposed water-regime treatments within each block, thereby avoiding a split-plot design. A one-week short-term drought

treatment was imposed on three-month-old rosettes. During the treatment period, roots of well-watered plants were flooded with four inches of water for 30 minutes every day, and drought-treated plants remained un-watered for the week. Instantaneous water-use efficiency (WUE), calculated by dividing carbon fixation rate by water transpiration rate, was recorded on whole plants using a modified system and protocol (Tonsor & Scheiner 2007) based on a Li-Cor LI-6400 apparatus (Li-Cor, Lincoln, Nebraska, USA). At the end of the one-week treatment, each plant was put in a separate cuvette, and from each cuvette, five measurements were taken with a 10-second interval once the concentration of CO_2 had stabilized. The mean of five measurements from each plant was used in further analysis. Measurements were made between 9 am to 5 pm with about 400 μmol mol^{-1} CO_2 and 26% relative humidity in the surrounding environment. I am able to process all plants within each block in the same day, and the seven-day drought treatment for each block was initiated in different dates. Therefore, plants in the drought treatment had experienced dry conditions for exactly seven days at the time of WUE measurement. In addition, the light intensity (photosynthetically active radiation, PAR) was recorded real-time in each cuvette as a covariate for photosynthetic rate.

Statistical analyses were performed with mixed model ANOVA fitted by REML in JMP 8 (SAS, Cary, NC). Subspecies (EAST/WEST), treatment (water/drought), geography (allopatric/sympatric), and their two-way (subspecies-by-geography, subspecies-by-treatment, geography-by-treatment) and three-way (subspecies-by-geography-by-treatment) interactions were used as fixed effects. Random effects include blocks and genotypes (nested within subspecies-by-geography). The interaction effect between treatment and genotype explained virtually no variance and was therefore not included in the model. In addition, the log-transformed light intensity in each Li-Cor chamber and the time of day were used as fixed effect covariates. To investigate the

trait divergence between subspecies under a specific water regime, I also performed statistical analyses separately for the drought and watering treatment, with all factors involving treatment removed. If instantaneous WUE has diverged in response to different local water availability between habitats of the two subspecies, I predict that EASTERN genotypes should have higher overall or treatment-specific WUE.

After Li-Cor measurement, all plants were returned to normal watering for two additional weeks before vernalization. Plants were vernalized in 4 degree C for six weeks under short day condition (12 hour daylight). All plants remained in normal watering conditions after vernalization. I monitored the plants every day and recorded bolting time and the starting and ending dates of flowering. The end of the flowering period is defined as the day after which no flower appeared for ten days. On the day of first flowering, width, height, leaf number, rosette number (total number of main and side rosettes), and stalk number were also recorded. After flowering finished, I also measured the diameter of the main flowering stalk, height of the stalk, and average reproductive internode length (stalk length containing reproductive branches / [number of reproductive branches - 1]).

Statistical models for phenology and morphological traits were similar to the model for physiology traits, except that light and time-of-day covariates were not used. Because prior analyses found no effect of the short-term water-regime treatment on phenology and morphology traits, the effects involving water regime treatment were also excluded from the statistical model. Adapted to their native montane environment with ephemeral water supply, I predicted that EASTERN genotypes should show typical traits of drought escape (Mckay *et al*. 2003), including faster bolting and flowering time, shorter flowering duration, and smaller plant size when flowering. On the other hand, the WESTERN subspecies should have overall delayed phenology and larger size at

reproduction to maximize the reproductive output in their native environment with abundant and persistent water supply. Since the relationship between stalk morphology and local water availability is not yet clear, I make no prediction for this trait.

4.1.3 Experiment 2. Long-term drought manipulation

In this experiment, another 1152 individuals were planted in 24 randomized blocks. Individuals within each block were arranged in the same way as Experiment 1, allowing a within-block watering treatment. The well-watered treatment was the same as experiment 1, but the plants under drought were watered once per week. The treatment was imposed on one-month-old rosettes, and one leaf from each individual was collected after eight weeks of drought treatment. For each genotype in a treatment, leaves from four blocks were pooled together, resulting in 288 samples for carbon stable isotope analysis, with 6 replicates of 24 genotypes and 2 treatments. Leaves were dried in 37 degrees C for one week and homogenized into powder in liquid nitrogen. $\Delta^{13}C$, the parameter associated with long-term water use efficiency (Farquhar *et al.* 1989), was measured in the Duke Environmental Stable Isotope Laboratory.

The statistical model was similar to the model for instantaneous water use efficiency in experiment 1, except that block, light intensity, and time-of-day effects were not included. The leaf samples were submitted for carbon isotope analyses in three batches of 96-well plates, and therefore batch was used as a random effect in the model (following the recommendation of Bolker *et al.* 2009). As in experiment 1, I predict that EASTERN genotypes should have higher overall or treatment-specific WUE if this trait has responded to different local environments.

4.1.4 Experiment 3. Vegetative-phase morphology without drought treatment

In this experiment, five plants from each of the 24 genotypes were grown in a completely randomized design under well-watered conditions for two months. By modeling a rosette as a cone, I calculated rosette volume (cm^3) as:

$$\pi r^2 h / 3,$$

where r is the radius and h is the height of the rosette. Alternatively, rosette volume could be modeled as a cylinder ($\pi r^2 h$), which would not affect my estimation of P-value or Q_{ST} since the volume of a cone and a cylinder only differ by a constant. All leaves were collected from each plant and scanned on a white background. Total leaf area (cm^2) was estimated by calculating the number of non-white pixels in the picture (with a resolution of 200 dpi, or 40,000 pixels per square inch). Rosette leaf packing was calculated as total leaf area divided by rosette volume. In addition, leaf fresh weight was measured at the time of harvest, and dry weight was measured after drying leaves at 65 degree C for one week. Rosette water content and water proportion were also calculated, along with unit-leaf-area fresh weight, dry weight, and water weight. Since all plants were harvested at the same age, the measured whole-rosette dry weight is proportional to the growth rate of each plant. Throughout this study I will use the terms whole-rosette dry weight and plant growth rate interchangeably.

From the scanned image I chose one fully developed leaf from each individual for leaf shape analysis. A custom Perl script was used to generate lines separating the longer axis of a leaf into ten sections of equal length. Twenty landmarks were picked in ImageJ (Abramoff *et al.* 2004) from the intersection between these lines and the leaf perimeter. Another custom Perl script was used to rotate and scale the landmark points to a standardized length for each leaf. Half of the width across the nine line-boundary

intersections of a standardized leaf was used for the final analysis. Therefore, the nine leaf shape parameters (Y1 to Y9, Figure 9) represent the width/length ratio across nine internal segments of a leaf. The statistical model was identical to the model for morphological traits (without treatment) in experiment 1, except that there is no block random effect in this experiment.

For rosette morphology, I predict that WESTERN genotypes would have higher rosette fresh weight, dry weight, and total leaf area, reflecting a non-conservative water use strategy to obtain maximum biomass before reproduction. On the other hand, the EASTERN subspecies may have higher leaf water content and lower overall growth rate, reflecting a life history strategy for water conservation. In addition, the EASTERN subspecies may have higher leaf packing (total leaf area per unit rosette volume) to reduce leaf water loss (McKay *et al.* 2001). Finally, the thermoregulation of leaves is critical to plants. During exposure to sunlight leaves may decrease their temperature via convection and transpiration. Small and narrow leaves have small boundary layers and are more efficient in heat convection, while large and wide leaves are more efficient in thermoregulation via transpiration (Nicotra *et al.* 2011). Therefore, I expect that genotypes from water-limited environments (EAST) would have narrow leaves, while the riparian WESTERN genotypes would have wider leaves, reflecting different strategies of foliar thermoregulation in response to different local water availability. In addition, wider leaves of WESTERN genotypes may also contribute to rapid biomass accumulation before reproduction.

4.1.5 Principal component analysis

To summarize and visualize the trait differentiation among genotypes, I performed principal component analysis (PCA) with function prcomp in R

(http://www.r-project.org/), using least square means of the 24 genotypes from the univariate mixed model ANOVA described above. I further separated all measured traits into five categories (physiology, phenology, morphology-stalk, morphology-rosette, and morphology-leaf) and performed PCA within each category. Notice that PCA was calculated from genotypic means rather than individual-level data because micro-environmental effects may influence the pattern of major PC axes in individual-level PCA.

4.1.6 Calculation of univariate and multivariate Q_{ST}

Q_{ST} calculates the proportion of heritable trait variation that exists among populations. If a trait is under divergent selection, Q_{ST} may be higher than F_{ST}, the proportion of neutral molecular variation among populations (Whitlock 2008). To calculate the variance components of subspecies and genotypes, I used subspecies and genotype nested within subspecies as random effects. For traits measured in experiment 1, block was also used as a random effect. Geographical effects were not included in this model because they lack significant effects for nearly all traits. Q_{ST} was calculated as:

$$V_{Subsp} \; / \; (V_{Subsp} + V_{Genotype}),$$

where V_{Subsp} and $V_{Genotype}$ are the variance components of subspecies and genotype-within-subspecies, respectively. Notice this differs from the typical Q_{ST} formula in that I did not multiply the within-subspecies variance component ($V_{Genotype}$) by two in the denominator. Like *Arabidopsis thaliana*, *Boechera stricta* is a highly selfing species and therefore can be modeled as haploid for these calculations (Whitlock 2008). In addition, since *B. stricta* has low genetic variation within local populations, my experimental design does not involve multiple genotypes from the same local population. The trait

'instantaneous water use efficiency under drought' had zero heritability, and therefore I do not calculate its Q_{ST}.

The multivariate trait Q_{ST} was calculated separately for four trait categories (phenology, morphology-stalk, morphology-rosette, and morphology-leaf). Within each trait category, the individual-level phenotypes of multiple traits were first scaled to zero mean and unit variance and then analyzed in a discriminant function analysis (DFA), with subspecies as the grouping variable. DFA identifies a linear combination of traits that maximizes the variation between and minimizes the variation within subspecies, providing a rotation of axes to the direction of greatest divergence between groups. The DFA score of each individual was then considered as a new univariate trait, and the Q_{ST} of this 'composite trait' was calculated with the same random effects model above (refer to Appendix Figure S2 for a detailed explanation of composite trait). I did not calculate the multivariate Q_{ST} of physiological traits under dry and wet treatments because the calculation requires traits from the same individual plants within the same experiment. To investigate the relationship between univariate traits and the composite traits, I estimated Pearson's correlation coefficient between each univariate trait and the DFA score from the same trait category.

4.1.7 Empirical SNP F_{ST} distribution

When comparing Q_{ST} with F_{ST}, recent opinion has called for the use of SNPs rather than microsatellite markers, because the high mutation rate of microsatellites may increase the within-population molecular variation and thus falsely decrease F_{ST} (Edelaar & Björklund 2011; Edelaar *et al.* 2011). In addition, Whitlock (2008) emphasized that Q_{ST} should be compared to genome-wide F_{ST} distribution, not to mean F_{ST}. To generate the empirical distribution of SNP F_{ST}, I used the method developed by

Andolfatto et al. (2011a). Genomic DNA of 18 genotypes (a subset of the core 24 in this study) was digested using the Sau3AI restriction enzyme, and a barcoded library was prepared with modified Illumina adaptors (Andolfatto *et al.* 2011a). The library was sequenced in one lane of HiSeq 2000 (Illumina, San Diego, CA, USA) with paired-end 100 bp reads. This was the first trial of this method for *Boechera stricta*, and I only obtained ~33 million read-pairs, which proved sufficient for the current study. I applied a stringent quality filtering, retaining a sequence pair only if all bases in both reads have sequencing error rate <= 10^{-5}. Among the 33 million pairs, 26.6 million passed the quality filtering and had unambiguous barcode sequences.

The LTM genotype, one of the 24 core genotypes used in this study, has been sequenced with the Roche 454 platform by the Department of Energy Joint Genome Institute and with Sanger BAC end-sequences by HudsonAlpha Institute for Biotechnology. From these data, I assembled a draft genome with Newbler software (454 Life Sciences, Branford, CT, USA) using default parameters. The draft genome after length filtering is about 170 Mb, ~80% of the estimated *B. stricta* 216 Mb genome. About 21 million Illumina HiSeq read-pairs from the 18 *B. stricta* accessions were successfully mapped to the LTM draft genome with BWA (Li & Durbin 2009) using default parameters, and genotypes were called with SAMtools (Li *et al.* 2009) with default parameters. In every SNP, the genotype of a plant accession was considered missing if the sequencing depth is less than 6x, and a SNP was retained only when the proportion of missing plant accessions is < 25%. Together with the LTM reference genome, this data set contains 23,379 SNP from 11 WEST and 8 EAST genotypes. The F_{ST} of each SNP was estimated with the package HIERFSTAT (Goudet 2005) in R. With about 23.5 thousand SNPs, the expected distance between neighboring SNPs is roughly 9 kb. Since SNPs in close linkage may not evolve independently and the linkage disequilibrium (LD) in *B.*

stricta decays in about 10 kb (Song *et al.* 2009), I compared the F_{ST} distribution from all 23.5 thousand SNPs to the average distribution from 1,000 re-sampled data sets where SNPs have lower LD due to their wider separation in the genome. Each data set contains 5,000 randomly re-sampled SNPs, with the expected mean distance between SNPs as 43 kb. I then obtained the average distribution from those 1,000 distributions and obtained the 101 percentiles (0% to 100% with 1% intervals) from this average distribution. There is a strong correlation (Pearson's correlation coefficient $r \approx 1.0$) between the percentiles from the average 5,000-SNP distribution and the percentiles from the 23.5-thousand-SNP distribution. Therefore in this study I used the original F_{ST} distribution with all SNPs for F_{ST}-Q_{ST} comparison.

Since *B. stricta* is a primarily self-fertilizing species and has high microsatellite homozygosity (Lee & Mitchell-Olds 2011; Song *et al.* 2006), some SNPs with apparently high heterozygosity may represent duplicated genomic regions. Indeed, the distribution of SNP heterozygosity is highly skewed, with the median at zero (all homozygous) and upper 5% tail at about 0.5 (half of the accessions are heterozygous). Excluding SNPs with heterozygosity > 0.5 only slightly increases the mean F_{ST} from 0.237 to 0.245, but the upper 5% or 10% F_{ST} tail used for F_{ST}-Q_{ST} comparison remains unchanged.

4.2 Results

4.2.1 No significant divergence in eco-physiological traits between subspecies

In this study, I performed two differential watering treatment experiments, one with one-week (experiment 1) and the other with eight-week (experiment 2) drought treatments. Although I found significant effects for genotype under long-term drought, for light intensity under short-term drought, and for drought treatment in both

experiments, I did not observe any significant effects involving subspecies, geography, or their interaction effects with treatment (Table 5).

Table 5: Mixed model ANOVA results of water use efficiency in short-term and long-term drought experiment.

Factor[a]	Effect type[b]	Instantaneous WUE		Long-term $\Delta^{13}C$	
		F-value	*P-value*	*F-value*	*P-value*
Subsp	Fixed	0.32	0.580	3.13	0.092
Geo	Fixed	1.23	0.281	1.75	0.201
Trt	Fixed	13.21	< 0.001*	101.24	< 0.001*
Subsp*Geo	Fixed	2.10	0.163	2.73	0.114
Subsp*Trt	Fixed	2.03	0.155	1.11	0.292
Geo*Trt	Fixed	0.80	0.371	0.41	0.521
Subsp*Geo*Trt	Fixed	1.79	0.181	0.05	0.821
Time of day	Fixed	0.49	0.482	-	-
Light intensity	Fixed	11.56	< 0.001*	-	-
Geno(Subsp,Geo)	Random	-	0.556	-	< 0.001*
Block or batch	Random	-	< 0.001*	-	0.009*

a. Subsp – subspecies; Geo – geography; Trt – treatment; Geno(Subsp,Geo) – genotype nested within subspecies and geography.
b. The degree of freedom is 1 for all effects

4.2.2 Trait divergence between EAST and WEST subspecies

Figure 6 shows the PCA result of all traits together and for five subsets of traits (physiology, phenology, morphology-stalk, morphology-rosette, and morphology-leaf). In all trait categories except physiology (Figure 6B), PC1 separates the two subspecies, signifying the substantial trait divergence between subspecies.

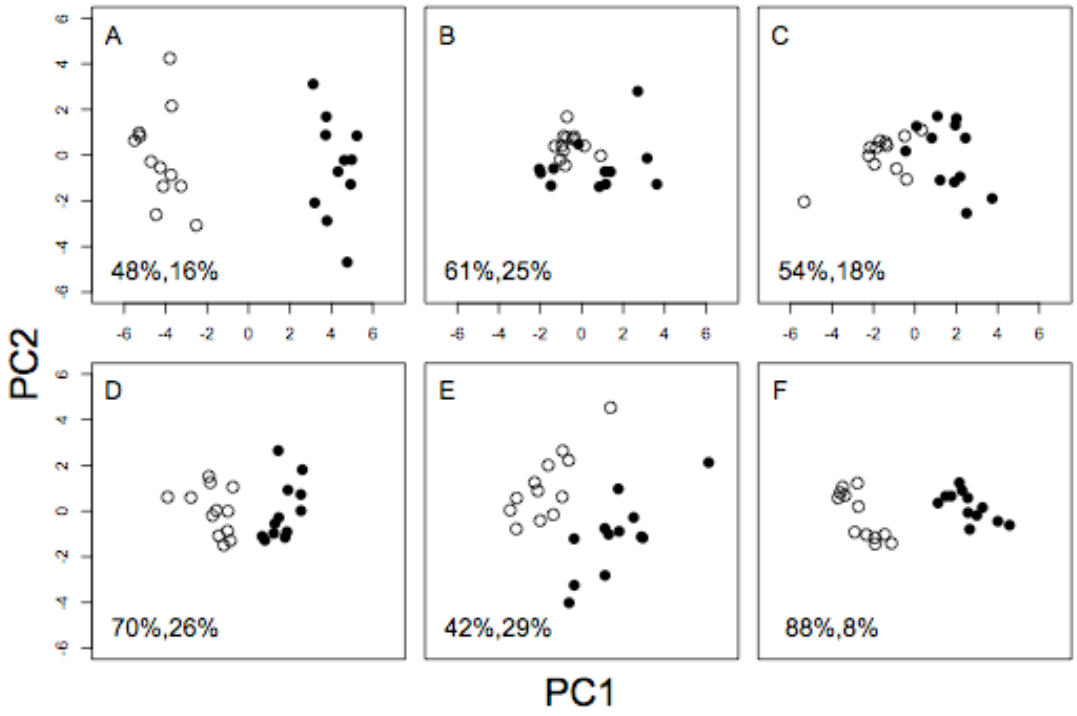


Figure 6: Principal components of genotype-level trait values. EAST genotypes - closed circles. WEST genotypes - open circles. A - all traits. B – four physiology traits. C – eight phenology traits. D – five stalk morphology traits. E – thirteen rosette morphology traits. F – nine leaf shape traits. Refer to Table 2 for the traits within each category.

To specifically examine which traits show significant EAST-WEST divergence, I performed mixed model ANOVA for each trait. Consistent with the trend from PCA, many non-physiological traits show significant divergence between subspecies after sequential Bonferroni correction within each trait category (Appendix Table S5 and Figure 7A). In addition, the direction of trait divergence is mostly consistent with my previous niche modeling prediction. Specifically, the WESTERN subspecies has faster growth rate (higher biomass and larger total leaf area at the time of leaf harvest), overall delayed phenology (slower bolting time, delayed flowering time, and longer flowering duration), and larger photosynthetic organ size (larger total leaf area and broader leaves), allowing them to attain higher overall biomass and reproductive output in their

native riparian habitat. On the other hand, the EASTERN subspecies has a slower growth rate, overall accelerated phenology, and narrower and more succulent leaves (higher water weight but not dry weight per unit leaf area), consistent with the escape from late-summer drought in their native montane habitat. Results from the 19 genotypes with SNP data (Appendix Table S6 and Appendix Figure S3) are highly consistent with the results from all 24 genotypes.

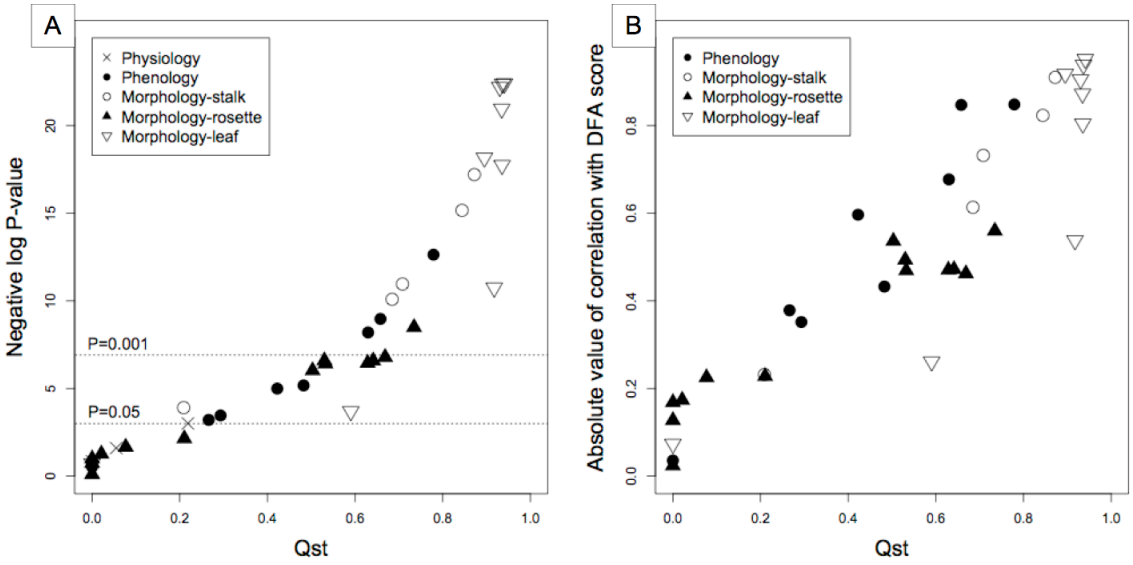


Figure 7: Relationship between trait Q_{ST} and (A) negative log *P*-value of subspecies effect in ANOVA (B) absolute value of correlation with discriminant function analysis (DFA) score from each trait category. Traits with high Q_{ST} generally have low *P*-values (high negative log *P*) and high correlation with DFA score. Consistent with Figure 2, many morphological and phenological traits are highly diverged. Shown are data from all 24 genotypes.

4.2.3 Comparing F_{ST} to univariate and multivariate Q_{ST}

In general, Q_{ST} of most traits corresponds to the *P*-values for subspecies divergence in ANOVA (Figure 7A), and traits with small *P*-values also have large Q_{ST} values. Because many traits were chosen to test divergent selection between the two subspecies, I only compared trait Q_{ST} to the upper tail of genome-wide distribution of SNP F_{ST}.

Figure 8 shows the F_{ST} distribution from 23,379 SNPs across the *B. stricta* genome, with the 5% cutoff at 0.88. Leaf shape parameters Y3 to Y9 have higher Q_{ST} than this F_{ST} cutoff, suggesting divergent selection on leaf shape between the two subspecies (Appendix Table S5 and Figure 7A). Figure 9 shows the average leaf shape of the two subspecies from all samples standardized for leaf length. Given the high amount of variation explained by PC1 of these leaf shape parameters (88%, Figure 6F), these parameters mostly represent the width/length ratio of a leaf. Clearly, the width/length ratio of the blade portion of a leaf is highly diverged between the two subspecies. In addition, some other traits have higher Q_{ST} than the 10% F_{ST} tail (0.75), including flowering height, main stalk height, and internode length between reproductive branches (Appendix Table S5). The adaptive significance of the three height-related traits, however, is not yet clear. Q_{ST} values obtained from the 19 genotypes with SNP data have only minor numerical difference from the 24 genotypes (Appendix Table S6). Specifically, in the 19-genotype data set two additional traits (rosette dry weight and rosette leaf area) have higher Q_{ST} than the 10% F_{ST} tail. Together with the higher leaf width/length ratio, this higher growth rate and larger photosynthetic organ size of the WEST subspecies may contribute to enhanced biomass accumulation before reproduction, which may be adaptive in its native environment with abundant water supply. On the other hand, the slower growth rate and narrower leaves of EAST subspecies may facilitate more water conservation.

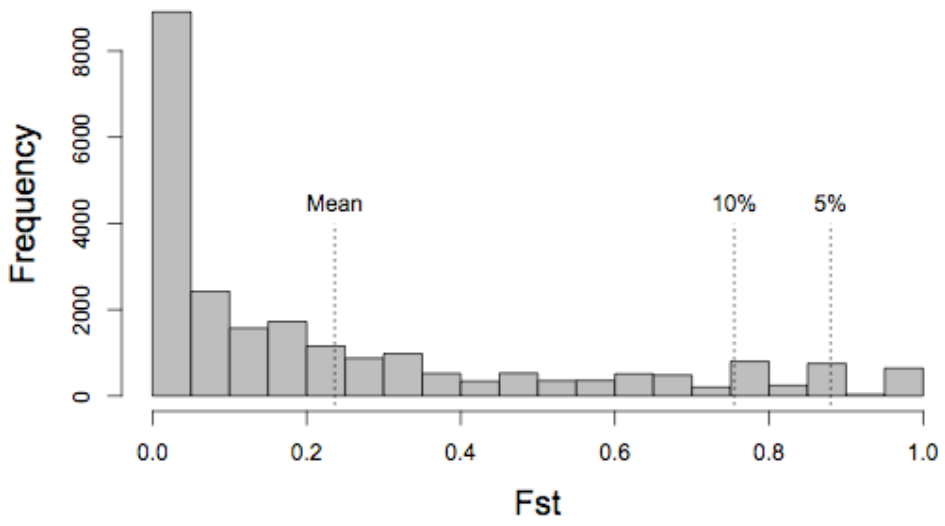


Figure 8: Empirical SNP F_{ST} distribution between 11 WEST and 8 EAST genotypes.

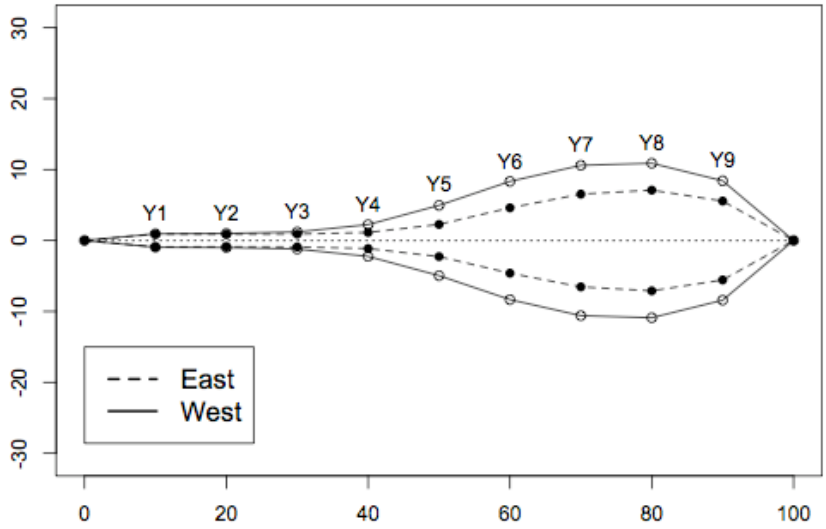


Figure 9: Average leaf shape of EAST and WEST genotypes (n = 60 from each subspecies). EASTERN leaf - closed circles connected by dashed line. WESTERN leaf - open circles connected by solid line. For every leaf, landscape points were rotated and scaled to obtain equal length among all leaves (a standardized length of 100 units across the horizontal axis), and points Y1 to Y9 separate the central leaf axis (dotted line) into ten sections of equal length. The Y coordinates of Y1 to Y9 were used in the statistical analyses.

Because traits within each category may be correlated, and natural selection may simultaneously act on multiple traits, I employed a multivariate version of Q_{ST}, looking at the divergence of all traits within each category. Within each trait category, the use of discriminant function analysis (DFA) between two subspecies generates a 'composite trait' with highest degree of between-subspecies divergence. This approach asks: what combination of traits shows greatest divergence between subspecies, and what is the Q_{ST} for this direction of maximum genetic divergence? The biological meaning of each composite trait can be inferred by examining the direction of individual trait divergence ('Higher' column in Appendix Table S5) and the sign of their correlation to DFA score ('DFA-cor' column in Appendix Table S5). In each trait category, small DFA values represent typical EASTERN traits (Phenology: accelerated reproductive time, smaller reproductive size, and more branches when flowering. Stalk morphology: thinner and taller flowering stalk with longer internodes. Rosette morphology: less rosette weight and total leaf area, but higher unit-leaf-area fresh weight and water weight. Leaf morphology: smaller width/length ratio.)

As shown in Table 6, all four categories have their composite trait Q_{ST} near or above the 5% F_{ST} cutoff of 0.88. The marginally significant ($P = 0.061$, Table 6) phenological multivariate Q_{ST}, for example, may reflect simultaneous natural selection on multiple phenological traits to accelerate reproduction of EASTERN genotypes. Similar pattern exists when only 19 genotypes were analyzed (Appendix Table S7). In addition, univariate traits with high Q_{ST} generally show higher correlations with the composite trait (Appendix Table S5 and Figure 7B). These data suggest that aspects of phenology as a whole may be under divergent selection, although to a lesser extent than morphological traits.

Table 6: Divergence of the 'composite trait' for each trait category. For DFA scores from each trait category, this table shows the *P*-value of the subspecies effect in univariate ANOVA, the Q_{ST}, and the empirical *P*-value of Q_{ST} compared to genome-wide distribution of SNP F_{ST} (Figure 4). Data are from all 24 genotypes.

Trait category	ANOVA *P*	Q_{ST}	*P* vs. F_{ST}
Phenology	< 0.001	0.87	0.061
Morphology – stalk	< 0.001	0.89	0.042
Morphology – rosette	< 0.001	0.97	0.027
Morphology – leaf	< 0.001	0.96	0.027

4.3 Discussion

Isolation by adaptation and ecological speciation result from differential local adaptation, where natural selection in distinct environments favors different organismal phenotypes. Reproductive isolation among populations may result from natural selection against immigrants or hybrids with deleterious phenotypes in local environments. Therefore, to understand how ecological factors affect genetic differentiation, one must investigate 1) the source of natural selection, 2) the traits under disruptive selection, and 3) whether the direction of trait divergence is concordant with local environments (Kawecki & Ebert 2004). However, few studies have investigated all three aspects of differential local adaptation. On one hand, many famous examples in ecological and evolutionary genetics investigated traits or genes under selection (Barrett & Hoekstra 2011; Mitchell-Olds *et al.* 2007), but sometimes little is known about the ecological causes of phenotypic change. For example, despite more than 80 years of study and clear empirical evidence of strong selection, the cause of natural selection on bony armor plates in three-spined sticklebacks remains ambiguous (MacColl 2011). On the other hand, the source of selection can be inferred via niche modeling in landscape genetics, but subsequent experimental verification is still needed.

My previous landscape genetics study (Lee & Mitchell-Olds 2011) suggested that local water availability may be an important selection force underlying ecological speciation between EAST and WEST subspecies of *Boechera stricta*. Using 24 genotypes in several large-scale greenhouse experiments, in this study I investigate: 1) whether water-regime associated traits have diverged between the two subspecies, and 2) whether the direction of trait divergence corresponds to their native environments. As in other studies, I employed Q_{ST}-F_{ST} comparison in controlled environments to identify possible traits under divergent selection. The alternative approach would be estimating the correlation between traits and fitness in the native environments. However, ideally such an experiment would be performed with a cross (instead of natural accessions) to minimize historical linkage disequilibrium among traits. These efforts are ongoing in my laboratory.

4.3.1 Trait divergence corresponds to niche modelingpredictions

Many traits have diverged significantly between these subspecies, especially for phenological and morphological traits (Appendix Table S5). Although there is significant variation among genotypes for water use efficiency estimated from $\Delta^{13}C$, the lack of subspecies or subspecies-by-treatment effect suggests that local adaptation between the two subspecies is not based on physiological traits for differential water usage (Table 5). Alternatively, this phenotype might be significantly different in the field environment, given the possibility of genotype-by-environment interaction.

Most of the significantly diverged traits show a direction of divergence conforming to my previous predictions (Appendix Table S5). In phenology, EASTERN genotypes flower significantly faster and for shorter duration, which are typical traits of drought escape (Mckay *et al.* 2003). Escaping from drought during the reproductive stage

is probably important for this species, since my preliminary greenhouse observations show that reproductive organs (flowers and fruits) are more susceptible to drought stress than vegetative organs, as is commonly found in crop plants (Bernier *et al.* 2008; Messmer *et al.* 2011). Although I did not identify univariate phenological traits with high Q_{ST}, the high multivariate Q_{ST} suggests divergent selection on overall phenology. In addition, the divergence in phenology may further decrease gene flow between subspecies. Since the decreased gene flow would increase genome-wide genetic divergence, this effect may make the F_{ST}-Q_{ST} comparison conservative for moderately diverged traits (such as phenology itself), but has little effect on the identification of highly diverged ecologically important traits (such as vegetative morphology, below).

In vegetative morphology, the leaves of WESTERN genotypes are more mesophytic (broader, thinner, and with larger surface area), which may facilitate the higher growth rate and higher biomass observed in my controlled environment. On the other hand, leaves of EASTERN genotypes are narrower, smaller, and more succulent (with higher water content per unit leaf area), reflecting a more xerophytic morphology which may enable water conservation. WESTERN genotypes' faster growth rate and delayed phenology result in higher biomass accumulation before the onset of reproduction, which may be advantageous in their native riparian habitats where the length of growing season is not strongly constrained by water availability. In addition, the high Q_{ST} of leaf shape parameters (width/length ratio) may be caused by their dual functions in photosynthesis and thermoregulation, both of which are related to local water availability.

The significant divergence and high Q_{ST} of some stalk morphology traits, however, may not reflect natural selection from local water availability. For example, EASTERN genotypes have taller central reproductive stalks and longer reproductive

internodes. I hypothesize that taller fruiting structures enable seeds to disperse further (which may be adaptive in complex or successional environments) and that longer internodes between fruits might reduce the risk that multiple fruits be attacked by an insect herbivore. However, detailed studies are needed to identify the real agent of selection on these traits.

At first glance, my results may seem mixed: some water-related traits (physiology) do not diverge significantly, and some traits (stalk morphology) diverge without obvious reason. The strong divergence in stalk morphology may indicate that my previous niche modeling study did not identify all factors contributing to the EAST-WEST divergence. On the other hand, for water availability to cause ecological speciation, not all water-related traits have to diverge significantly, and ecological speciation could be caused by divergent selection on a few traits (Sobel *et al.* 2010). Indeed, among all traits that are predicted to be water-related and observed to be significantly divergent, all but one trait show the direction of divergence conforming to my prediction. This exception is leaf packing (leaf area per unit rosette volume). In theory, leaf packing should be higher in drought-adapted genotypes, where similar amounts of total leaf area are packed into smaller rosette volume to minimize leaf water loss. Given the similar rosette volume between the two subspecies, I think the high leaf packing in the WEST subspecies may be a by-product of its larger total leaf area, a water-related trait under strong divergent selection.

4.3.2 Lack of physiological differentiation

Previous studies have shown that leaf morphology, such as specific leaf area (leaf area per unit dry weight) and leaf water content, can influence water use efficiency (Condon *et al.* 2004; Hoffmann *et al.* 2005; Nautiyal *et al.* 2002). In this study I found

significant subspecies differentiation in some leaf morphology traits, but physiological traits (instantaneous and long-term water use efficiency) did not differ significantly between subspecies. At first glance my result may seem contradictory to previous studies. However, my results show that the two subspecies lack significant differentiation in the two important morphological traits that influence water use efficiency (rosette water proportion and rosette dry weight / area, Appendix Table S5), and this is consistent with the lack of physiological differentiation between subspecies. In addition, the whole-rosette-level physiology is a balance between individual-leaf-level physiology and rosette structure. As discussed above, rosette leaf packing is the only significantly diverged trait that contradicts my prediction. The higher leaf packing in WESTERN genotypes may decrease rosette water loss from convection and offset the higher evaporation rate from the mesophytic WESTERN leaves (and vice versa for EASTERN genotypes), leading to non-significant EAST-WEST physiological differentiation. Another influencing factor may be that my experimental conditions are imperfect models of natural environments.

In addition to rosette-level water use efficiency, a recent study has shown that inflorescences have higher water use efficiency than rosettes in *A. thaliana* (Earley *et al.* 2009). It is possible that similar patterns may exist in *B. stricta*, and there may be different water use efficiency between EASTERN and WESTERN inflorescences given my observed difference in stalk morphology. Future experiments are needed to examine this possibility.

4.3.3 Lack of geographic effects

Previous analysis of molecular polymorphism suggests that the current geographical distribution of these subspecies represents secondary contact after

historical allopatry (Lee & Mitchell-Olds 2011). From the previous niche modeling result, I proposed a possible relationship between trait divergence and reproductive isolation: during the allopatric phase, the two subspecies diverged in traits associated with local water regime. After secondary contact, these diverged traits caused differential local adaptation in distinct environments, and therefore immigrants or hybrids had reduced fitness, contributing to reproductive isolation. This hypothesis predicts that, within each subspecies, the traits associated with local water regime would not differ between sympatric and allopatric regions. Consistent with this hypothesis, I found no evidence for water-regime-associated traits with significant geography or geography-by-subspecies interaction effects.

Between two taxa, reinforcement in speciation refers to the situation where sympatric populations have higher pre-mating reproductive isolation than allopatric populations (Coyne & Orr 2004), which avoids the costs of producing unfit hybrids. For a reproductive trait, reinforcement is inferred when the trait divergence is higher in sympatric than in allopatric regions. I do not observe this pattern in phenology traits, and this is consistent with the observed high hybrid viability from artificial crosses and the highly-selfing reproductive system in this species. On the other hand, if hybridization homogenized trait distributions, then trait divergence would be lower in sympatric than in allopatric regions. I find no evidence for this pattern, either. In fact, the observation that some traits have higher Q_{ST} than neutral F_{ST} shows that, instead of being reduced by hybridization, the trait divergence has been maintained by divergent selection between heterogeneous environments.

4.3.4 Comparing F_{ST} with multivariate Q_{ST}

Three methods could be used to analyze trait divergence among genetic groups: 1) estimating subspecies effects (*P*-value) in ANOVA, 2) comparing Q_{ST} to the confidence interval of mean F_{ST}, and 3) comparing Q_{ST} with the genome-wide distribution of F_{ST}. Although the second method is the most widely used for F_{ST}-Q_{ST} comparison, recent opinions advise against this practice (Whitlock 2008). In accordance with recent suggestions (Edelaar & Björklund 2011; Edelaar *et al.* 2011; Whitlock 2008), I compared trait Q_{ST} to the genome-wide distribution of SNP $F_{ST.}$

The high divergence (mean F_{ST} = 0.24) between EAST and WEST subspecies, however, sets a high threshold for detecting significant Q_{ST}, and I only find a few univariate traits (leaf shape parameters, in particular) with Q_{ST} above the 5% F_{ST} cutoff (Appendix Table S5). In addition, judging from the frequency of SNPs with F_{ST} higher than 0.75 (Figure 8) and field evidence that many traits and QTL experience natural selection in this species (Anderson *et al.* 2011a; Anderson *et al.* 2013; Anderson *et al.* 2012; Prasad *et al.* 2012), my genome-wide F_{ST} distribution also may contain SNPs linked with genomic regions under divergent selection. Since Q_{ST} should be compared to the distribution of neutral F_{ST} (Whitlock 2008), my results are likely conservative. I therefore designed a measure of multivariate Q_{ST} to investigate the joint divergence of multiple traits. Q_{ST} allows researchers to search for signatures of natural selection on individual traits, while its population genetics analogs (F_{ST} and related parameters) facilitate the search for single target genes under selection. Recently, population geneticists have emphasized that adaptation may occur by slight allele frequency changes at many genes (polygenic adaptation), and each locus may show little signature of natural selection (Pritchard *et al.* 2010). Similarly, natural selection often acts on

combinations of traits (Blows 2007), causing only moderate increase in the Q_{ST} of univariate traits. Thus, I present a simple measure of selection on multiple traits, using the Q_{ST} of a new composite trait from discriminant function analyses (DFA) between these subspecies. This composite trait represents the axis of maximum divergence in the multivariate trait space (Appendix Figure S2). Indeed, my results show that the multivariate Q_{ST} is close to the 5% tail of F_{ST} distribution, as expected when multiple traits are simultaneous targets of divergent selection (Chenoweth *et al*. 2008).

Because the DFA approach (by definition) maximizes the among-group variation and minimizes the within-group variation, is this multivariate Q_{ST} somehow unrepresentative or biased? This is not a concern for several reasons. First, most quantitative traits are multivariate, embedded in combinations of other traits (Houle *et al*. 2010). Therefore, a DFA composite trait is biologically meaningful – the trait (which I am unable to identify *a priori*, such as overall phenology) that is under the strongest divergent selection (Appendix Figure S2). Second, this procedure is simply a rotation of axes, hence the statistical concept of bias does not apply. Identifying this direction of greatest divergence is an important evolutionary question, which is not related to statistical bias. Third, DFA is closely related to MANOVA. Although MANOVA may give lower *P*-values than univariate ANOVA, this does not imply that MANOVA has biased the *P*-value downwards, and MANOVA is still a standard practice in biology. Similarly, there is no reason to think that Q_{ST} of DFA score would be biased upwards. Fourth, similar concepts have been proposed by several authors. Lande (1979), when regressing fitness onto multiple traits, suggested '…constructing a selection index or discriminant function where each character is weighted by the force of directional selection on it…', and therefore 'Calculation of the minimum selective mortality is thus reduced to a consideration of truncation selection on the index, a one-dimensional

problem…' Blows (2007) proposed a similar idea: 'This immediately suggests that the presence of linear selection can be most effectively tested for by considering the significance of selection on the univariate discriminant function…' Both suggestions by definition maximize the variation of fitness explained by traits, but this does not introduce bias. Finally, although one might apply DFA to SNP polymorphisms, this approach would be unlikely to represent the neutral null distribution needed for F_{ST}-Q_{ST} comparison.

Although other measures of multivariate Q_{ST} has been proposed based on decomposing covariance matrices (Kremer *et al.* 1997; Martin *et al.* 2008; Ovaskainen *et al.* 2011), my method has two differences: 1) Estimating the covariance component matrix may be time-consuming and unstable when the number of groups or subspecies is low. My composite-trait method avoids this complication. 2) The DFA composite trait is biologically meaningful – it is analogous to the most diverged combination of traits between two subspecies (Blows 2007; Lande 1979).

4.3.5 Conclusion

Differential local adaptation forms the basis of ecological speciation and isolation by adaptation. To understand the process of ecological speciation, one must investigate the source of natural selection and the traits under selection, whose interactions shape the patterns of differential local adaptation. In a previous study (Lee & Mitchell-Olds 2011), I showed that local water regime may be the selective force underlying ecological speciation between two genetically diverged subspecies of *Boechera stricta*. In this study I have identified possible traits experiencing this disruptive selection, and the direction of trait divergence mostly corresponds to niche modeling predictions. On the other hand, I have also identified several traits that are highly

diverged without obvious water-related functionality. This suggests that some important selection forces were not identified in my previous niche modeling study (Lee & Mitchell-Olds 2011). In summary, this study identifies traits contributing to incipient ecological speciation in *B. stricta* and demonstrates the importance of experimental verification of inferences from niche modeling approaches. Furthermore, this evidence for differentiation of ecologically important traits provides the starting point for genetic dissection and evolutionary interpretation of trait variation contributing to ecological speciation.

4.4 Data availability

Data are deposited at Dryad doi:10.5061/dryad.rh0mv

5. Quantitative trait loci mapping identifies genomic region controlling ecological speciation of *Boechera stricta*

The study of ecological speciation emphasizes the role of ecological factors in generating contrasting selection forces in the native environments of diverging lineages. Under ecological speciation with occasional gene flow, it is expected that genomic regions (quantitative trait loci, QTL) underlying ecological speciation will control ecologically important traits, will contribute to fitness difference in the field, and will show high divergence compared to the rest of genome.

Based on the existence of genetic tradeoffs for fitness in different environments, two distinct patterns may describe the effects of QTL controlling fitness in reciprocal transplant experiments (Anderson *et al.* 2012; Colautti *et al.* 2012): In antagonistic pleiotropy, both alleles of a QTL exhibit local adaptation in their respective native sites and are maladaptive in the other environments, i.e., reciprocal change in rank fitness. In conditional neutrality, while one allele is advantageous in its native site, in the other environment this QTL has no fitness effect. Empirical evidence has identified both patterns (Anderson *et al.* 2012; Hall *et al.* 2010). Nevertheless, more examples are needed to understand the relative importance of antagonistic pleiotropy and conditional neutrality in ecological speciation, the establishment of reproductive isolation via local adaptation.

The pattern and effect of 'speciation QTL' may differ according to the geographic scale of speciation or the mating system of organisms, and different strategies may be required to study speciation loci in each case. Conceptually, parapatric or sympatric speciation with continuous gene flow in obligate outcrossing organisms may be more likely to show speciation QTL with antagonistic pleiotropy effects, because antagonistic pleiotropy is more effective in maintaining genetic variation despite ongoing gene flow in other regions of

the genome, and conditionally neutral QTL may be fixed across all populations in the absence of fitness tradeoffs. Accordingly, the reverse genetic approach of whole-genome scanning for highly diverged regions (Ekblom & Galindo 2011; Feder *et al.* 2012) might be successful in this case. On the other hand, loci with conditionally neutral effects on fitness may have higher probability to be observed in cases of secondary contact after historical allopatry. During the allopatric stage, different lineages may separately evolve and adapt to distinct environments, fixing alleles in different genes that are locally advantageous but not necessarily maladaptive in the other environment. After secondary contact, reproductive isolation within the contact zone may still be maintained if natural selection is strong enough to eliminate immigrant individuals before hybridization occurs, especially for primarily self-fertilizing organisms.

Boechera stricta is an emerging model organism for evolutionary genetics (Rushworth *et al.* 2011). This species contains two distinct genetic groups (subspecies) with a contact zone in the Northern Rocky Mountains. In previous studies, I have shown that environmental adaptation contributes to the genetic differentiation between subspecies, and local water availability appears to be the most important environmental variable differentiating preferred habitats (Lee & Mitchell-Olds 2011). While the EAST subspecies mostly occur in high elevation montane habitats with low and ephemeral water availability, the WEST subspecies mostly occurs in low elevation riparian sites where soil water supply is more abundant and persistent. Further greenhouse experiments have shown that the two subspecies differ in traits associated with adaptation to different water availability. Comparing Q_{ST} (the proportion of quantitative genetic variation distributed between subspecies) versus F_{ST} (the proportion of neutral genetic variation occurring between subspecies), I found that Q_{ST} is significantly higher than F_{ST} for some ecologically important traits, suggesting that trait

divergence between subspecies reflects adaptive responses to environmental differences (Lee & Mitchell-Olds 2013). While the two subspecies do not differ significantly in short-term or long-term water use efficiency, EAST genotypes have overall traits that are more suitable for escaping or resisting drought (Mckay *et al.* 2003; Nicotra *et al.* 2011): faster phenology to escape drought, narrower leaves for more efficient heat convection, and more succulent leaf structure to prevent water loss by transpiration. While the EAST genotypes display traits for drought adaptation, alternative trait values in WEST genotypes are also hypothesized to increase fecundity in the benign WEST habitats with greater water availability.

The different types of environments and traits for these two subspecies suggest distinct selective forces or fitness components may be important in the native sites of each subspecies: I hypothesize that plants in the drier EAST environments may be more likely to experience selection on survival, and the benign WEST environments may be more likely to be under selection for fecundity. In this chapter, 1) I examined a cross from one EAST and one WEST genotype and measure different fitness components in both environments. 2) I also measured many traits in different environments and 3) performed quantitative trait loci (QTL) mapping to identify important genomic regions controlling adaptive traits and local adaptation between the two subspecies. 4) In addition, since the EAST-WEST distribution pattern suggests secondary contact after historical allopatry (Lee & Mitchell-Olds 2011), I also test whether fitness QTL exhibit patterns of antagonistic pleiotropy or conditional neutrality.

5.1 Materials and methods

5.1.1 Plant materials, phenotypic measurements, and trait analyses

The cross used for QTL mapping was developed from two parents in the EAST-WEST contact zone: one in Parker Meadow (Parker, EAST subspecies, 44°37′ N, 114°31′ W) and one in Ruby Creek (Ruby, WEST subspecies, 45°33′ N, 113°46′ W). The F1 hybrid was self-fertilized to produce F2 plants, and subsequent generations were propagated by self-fertilization and single-seed descent to create 153 independent genetic lines (families). In each line, multiple F4 progeny from the same F3 plant were used in a randomized complete block design, and the phenotypic least-square means (LSMEANS) were calculated to represent the genotypic value for their F3 parent. Each block consists of one F4 plant from each of the 153 lines and multiple Parker and Ruby individuals.

The Duke greenhouse experiment consists of 12 blocks. Seeds were stratified in 4° C for four weeks and planted in 'Cone-tainers' (Ray Leach SC10, Stuewe & Sons Inc., Tan-gent, OR, USA), with soil composition and greenhouse conditions as previously described (Lee & Mitchell-Olds 2013). When rosettes were 11-week old, all leaves from three-blocks of plants were harvested for rosette- and leaf-morphology measurements as described (Lee & Mitchell-Olds 2013). At 12-weeks of age, the remaining nine blocks were vernalized in 4° C for 6 weeks, then returned to the same greenhouse conditions for phenology and fitness measurements. All traits were measured in the same way as previous described (Lee & Mitchell-Olds 2013), except: 1) no physiological traits were measured; 2) leaf width/length ratio was used instead of leaf shape morphometrics because the leaf-shape landscape points were highly correlated (Lee & Mitchell-Olds 2013).

Using the same experimental design, a total of 12 blocks were used in the field experiment, with six blocks planted in the EAST and six in the WEST garden. Due to logistic constraints, I was unable to transplant these experiments to the exact locations where parents were collected. Instead, Jackass Meadow (JAM, 44°58′ N, 114°5′ W) and Alder Creek (ALD, 44°47′ N, 114°15′ W) are used as the EAST and WEST gardens, respectively. The JAM garden (elevation 2680 m) is located on a mountain slope, and the ALD garden (elevation 1980 m) is located at a riparian plain. Both gardens are within the EAST-WEST contact zone, and local environment and plant genotype correspond to typical EAST and WEST subspecies (Lee & Mitchell-Olds 2011). Following previous procedures (Anderson *et al.* 2011a; Anderson *et al.* 2012), plants were grown in the greenhouse to 10-weeks old before transplantation in fall 2011 and allowed to over-winter under natural vernalization conditions.

In summer 2012, each garden was visited every seven to ten days throughout the entire growing season, and plant stage was recorded as: missing (.), dead (X), rosette (R), bolting (B), flower-only (FO), flower-silique stage 1 (FS1 – more flowers than fruits), flower-silique stage 2 (FS2 – more fruits than flowers), and siliques-only (SO). The plant stage from each census was transformed to a quantitative trait for QTL mapping, where R = 1, B = 2, FO = 3, FS1 = 4, FS2 = 5, SO = 6, and missing or dead were not included. For census when the flower and fruit numbers were not counted, the flower-silique stages were collectively coded as 4.5. This is essentially a data transformation from an ordinal to continuous scale of measurement, summarizing the phenotypic variation in phenology. The proportion of leaf area damaged by insect herbivores was recorded in mid-summer, and plant fecundity in the end of summer was defined as the number of fruits (fecundity fruit) and the number of fruits multiplied by the length of a randomly chosen fruit of average length (fecundity seed).

For QTL mapping, all individual-level measurements were transformed to family-level LSMEANS in JMP 8 (SAS, Cary, NC, USA). Due to the highly skewed distribution of most traits, all characters, except binomial traits or plant stages, were log-transformed at the individual level. For greenhouse measurements, all measurements were made by Cheng-Ruei Lee, and block and genotype were considered as random effects. For field measurements, observer, block, and genotype were used as random effects while vegetation cover around each plant, the plant width before transplantation in fall 2011, and the square of plant width were used as covariates. Using 'initial plant width before transplantation' as a covariate controls for plant growth conditions during the 10-week period in Duke greenhouse, and therefore the LSMEANS can better reflect plant growth conditions in the field environment. Overall, a total of 85 traits in eight trait categories were measured (Appendix Table S8), including 25 traits in ALD, 26 in JAM, and 34 in the greenhouse (GH). Traits were excluded from further analyses if the heritability was less than 1%.

To estimate the relative effect of different episodes of selection on overall fitness output in the year, I conducted multiple regression using family LSMEANS. In each field garden, two analyses were performed separately for fruit number or seed number (approximated by fruit number multiplied by average fruit length):

FITNESS = SURVIVAL + BOLT + FECUNDITY_BOLTED + INTERACTIONS,

where FITNESS is the mean family-level fruit or seed number calculated from all plants (including individuals in all plant stages except missing), SURVIVAL is survival probability in each family, BOLT is the probability of bolting for individuals that survived, FECUNDITY_BOLTED is the number of fruits or seeds for individuals that bolted, and INTERACTIONS include all possible interaction terms of the three fitness components. Two parameter estimates are used to estimate the relative importance of

each episode of selection on overall fitness: 1) Regression slopes were compared among predictor variables. For the regression slopes to be comparable, all response variables in the regression model were divided by their mean to have mean at one, and all predictor variables were standardized to have mean of zero and standard deviation of one. 2) The proportional contribution of each predictor to the response variable is calculated from the decrease of r^2 when a variable is removed from the full multiple regression model. In the JAM garden, FECUNDITY_BOLTED had zero heritability. Therefore in JAM garden FECUNDITY_BOLTED was not used in the regression model, and this fitness component therefore had no contribution to the variation in overall fitness.

In addition to 85 univariate traits, I also calculated a 'composite trait' for each trait category (survival, fruit fecundity, seed fecundity, phenology, leaf morphology, rosette morphology, and stalk morphology in each garden and combined, Appendix Table S9). The composite trait was defined as the projection of family trait values on the vector connecting two parental means, and this new composite trait reflects the direction of parental divergence in the groups of traits in the same trait category. The composite trait value denotes how close a family is to each parent: larger values have overall traits more similar to the EAST parent, and lower values are more similar to the WEST parent. For example, a higher value in the phenology composite trait denotes faster flowering, smaller size and more branching when flowering, and lower probability of retaining active tip buds at the end of season – an overall faster phenology pattern typical of the EAST parent. In contrast, a lower value denotes slower phenology, which is more similar to the WEST parent. The relative contribution of each univariate trait to its composite trait can be estimated by correlation coefficients (Appendix Table S8).

5.1.2 Genotyping by sequencing

To genotype the cross, I employed an updated genotyping by sequencing (GBS) method derived from Andolfatto *et al.* (2011a). In each family, DNA was extracted (Qiagen DNeasy Plant Mini Kit) from at least ten pooled F4 individuals to represent the genotype of their F3 parent. Different from the original protocol (Andolfatto *et al.* 2011a), I used a new adaptor design which is compatible with TruSeq adaptors and indexes while allowing paired-end sequencing (Andolfatto, personal communication, Appendix Table S10). The combination of 48 unique barcodes with four different TruSeq indexes allowed multiplexing of 192 samples (153 families, 19 samples for Parker and 20 for Ruby parent). The library was sequenced in one Illumina HiSeq-2000 lane by the Duke Genome Sequencing & Analysis Core Resource, where ~249 million reads with unambiguous barcodes were obtained. Read pairs were assigned to genotypes and two parents by custom Perl code, and low-quality bases in the end of reads were trimmed by DynamicTrim (Cox *et al.* 2010).

I was unable to use the software from Andolfatto *et al.* (2011a) because a high quality reference genome sequence was not available at the time of these analyses. Following previous procedures (Lee & Mitchell-Olds 2013), all reads were mapped to *Boechera stricta* draft genomic scaffolds (Joint Genome Institute, version 2013 Feb. 11) with BWA (Li & Durbin 2009), and genotypes were called with SAMtools (Li *et al.* 2009). From ~712,000 raw SNPs (where any difference exists among the families, parents, and reference genome, including genotype-calling error) generated from SAMtools, my SNP-filtering script identified 1,690 high-quality SNPs where: 1) both parents are homozygous, have sequencing depth >= 4x, and have different alleles; 2) at least 70% of all families have sequencing depth >= 6x, where a genotype call with depth

< 6x is treated as missing data. By expectation, the selfed F3 generation has genotype frequency of 1/4 (~38 families) for heterozygotes and 3/8 (~57 families) for each homozygous genotype. Therefore, to prevent serious segregation distortion from affecting linkage map estimation and QTL mapping, SNPs with less than 25 families in any of the three genotypes were excluded, leaving 1,069 SNPs for further analyses.

To remove erroneous genotype calls and impute missing genotypes, the linkage map and genotype matrix were inferred with the following procedure:

1) I regard two recombination events within a 5-cM interval in the same copy of chromosome as unlikely: Given one recombination breakpoint generated by the F1 parent, the probability that another recombination event is observed within 5 cM in the F2 parent (with 50% heterozygosity, which decreases the chance of observing a recombination event by half) is roughly 2.5%. A preliminary analysis from the Joint Genome Institute shows that 5 cM roughly equals 1 Mb in physical length (Hellsten, unpublished). Therefore, my custom Perl script first scans for genotyping error along the same scaffold. Within a family, if two recombination events were inferred within a 1 Mb interval, genotype calls between the two recombination breakpoints were assigned as missing data.

2) From the filtered data, a linkage map was built by MSTMap (Wu *et al.* 2008), and seven linkage groups were obtained. All scaffolds were blasted to the ancestral chromosomal blocks of Brassicaceae (Schranz *et al.* 2007), and a SNP was manually removed if it was physically located on the ancestral block from the wrong chromosome or if it is more than 10 cM away from the two flanking markers in the linkage map.

3) New linkage maps were separately built for each linkage group, and another Perl script was used to remove suspicious genotype calls: if two recombination events

happened within a 5 cM interval on a chromosome for this new linkage map, genotype calls flanked by the two recombination breakpoints were assigned as missing data.

4) Another updated linkage map was then built from the filtered data set, and missing data were imputed based on genotype calls in the same family when: a) the two flanking SNPs with data have the same allele, unless this missing genotype is more than 30 cM away from both available markers; b) for missing data in the end of chromosomes, the allele is assigned the same as the nearest available SNP, unless it is more than 10 cM away. In short, a missing genotype is only imputed when the chance of recombination in the interval is low.

5) The final linkage map was built from this filtered and imputed data set, and this genotype matrix was used for QTL mapping.

5.1.3 Quantitative trait locus (QTL) mapping

All phenotypic measurements and DNA extractions were performed from multiple F4 plants, and the trait LSMEANS and pooled genotype of their F3 parent were used for QTL mapping.

To first investigate if there are any QTL controlling measured traits, I conducted multivariate least square interval mapping (MLSIM) on all traits in each trait category (Appendix Table S8) of each garden (Anderson *et al.* 2011a). In short, the genotype scores are calculated with 1 cM step size for the interval between neighboring markers. For each genomic location, multivariate ANOVA (MANOVA) is conducted with all traits in the same trait category as response variables and genotype scores of the target genomic location as predictor variables. QTL were added into the model with stepwise forward addition: the QTL with highest effect was first identified, and controlling for the previous QTL, the remaining genomic region with highest effect was then identified

and kept in the model. The steps were continued until no further QTL was significant. QTL significance was determined by comparing marker effect to genome-wide permutation distributions.

QTL mapping of all univariate and composite traits was conducted with the composite interval mapping algorithm in QTL Cartographer version 1.17 (Basten *et al.* 2005). For each trait, a stepwise multiple regression (program SRmapqtl) with forward and backward regression significance levels as 0.05 was first conducted to identify significant markers. The five significant markers with highest-effects, if available, were used as controlling cofactors in composite interval mapping (program Zmapqtl), and the empirical genome-wide significance threshold was generated by 1,000 permutations (Churchill & Doerge 1994). Following default setting, the walking speed within marker intervals is 2 cM, and a cofactor is temporarily ignored if it is within 10 cM of a genomic location being tested.

In each natural environment, fitness QTL conferring local advantage were identified. However, none of the QTL were statistically significant in both gardens. To test the effect of fitness QTL identified in one environment on the corresponding fitness components in the other field garden, one-way ANOVA was performed with family mean estimated from standard ANCOVA, using family as fixed effect, block and observer as random effects, and local vegetation density around each plant, rosette width before shipping, and the square of rosette width before shipping as covariates. To further test whether the pattern conforms to true conditional neutrality or possible antagonistic pleiotropy with low statistical power in the other garden, statistical power was estimated using 'design of experiments' in JMP 8.

5.2 Results and Discussion

5.2.1 Quantitative traits, heritability, and fitness components

In general, traits measured in the greenhouse have higher heritability than in field gardens (Appendix Table S8). Morphology has similar or higher heritability than phenology, and fitness components often have low or zero heritability. Except for plant stage (which was measured on different census dates), exactly the same traits were measured in the EAST (JAM) and WEST (ALD) gardens, facilitating the comparison between sites. For fitness components, while the EAST garden has higher heritability than the WEST garden for survival, the WEST garden has higher heritability for fecundity components of fitness. Indeed, the only fecundity traits in the EAST garden that has non-zero heritability is fecundity from all plants, which also is influenced by variation in survival. This may indicate that different selective forces or genetic mechanisms are responsible for local adaptation in the native sites of each subspecies. I further estimated the proportional contribution of 'survival', 'bolted in summer', and 'fecundity of bolted plants' to the overall fitness at the family level (Table 7). While survival is the most important contributing factor (~ 30%) of overall fitness in JAM, fecundity of bolted plants dominates (~ 50%) overall fitness in ALD. This observation is consistent with my previous results on the population genetics, niche modeling, and quantitative genetics both subspecies (Lee & Mitchell-Olds 2011; Lee & Mitchell-Olds 2013): EAST environments are mostly high-elevation mountain slopes with limited water availability where survival may be a major selective force, and EAST genotypes mostly show accelerated phenology and xerophytic morphology to avoid or survive drought. On the other hand, WEST environments are mostly low-elevation riparian sites with more consistent water availability, where fecundity may be a major selective force, and WEST

genotypes mostly show delayed phenology and mesophytic morphology to increase fecundity.

5.2.2 Linkage map

Seven unambiguous linkage maps were constructed (Figure 10), and the order of scaffolds along chromosomes is consistent with the ancestral blocks from Schranz et al (2007). Although in theory the GBS protocol sampled SNPs randomly from the genome, the marker density is non-homogeneous on the linkage map. This could be in part due to the uneven recombination rate across chromosomes, the uneven SNP distribution between parents, or the existence of highly repetitive genomic regions.

Table 7: Relative contribution of survival, bolting, and fecundity fitness components to the variation of overall fitness at the family level.

	JAM		ALD	
	Fruit fitness [b]	Seed fitness	Fruit fitness	Seed fitness
Survival (S)	0.060 (0.33)***	0.052 (0.35)***	0.016 (0.02)***	0.013 (0.02)***
Bolted in summer (B)	0.029 (0.07)***	0.025 (0.08)***	0.025 (0.08)***	0.023 (0.10)***
Fecundity of bolted plants (F) [b,c]	-	-	0.069 (0.51)***	0.054 (0.46)***
S*B [d]	0.017 (0.02)*	0.012 (0.02)*	0.002 (<0.01)	0.002 (<0.01)
S*F	-	-	0.006 (<0.01)***	0.005 (<0.01)***
B*F	-	-	-0.001 (<0.01)	-0.001 (<0.01)
S*B*F	-	-	-0.002 (<0.01)	-0.002 (<0.01)

a. Shown are the regression slopes and proportional variation explained (r^2, in parenthesis) by each fitness component.
b. Within each garden, shown are the fitness components measured as fruit or seed number (approximated by fruit number * average fruit length).
c. In the JAM garden, the heritabilities of fruit or seed fecundity of bolted plants equal zero, leading to identical values for all family LSMEANS. This factor is therefore not used in the regression and percent contribution coded as missing.
d. Any small but non-zero proportion of contribution is indicated as < 0.01
* $P <= 0.05$; ** $P <= 0.01$; *** $P <= 0.001$

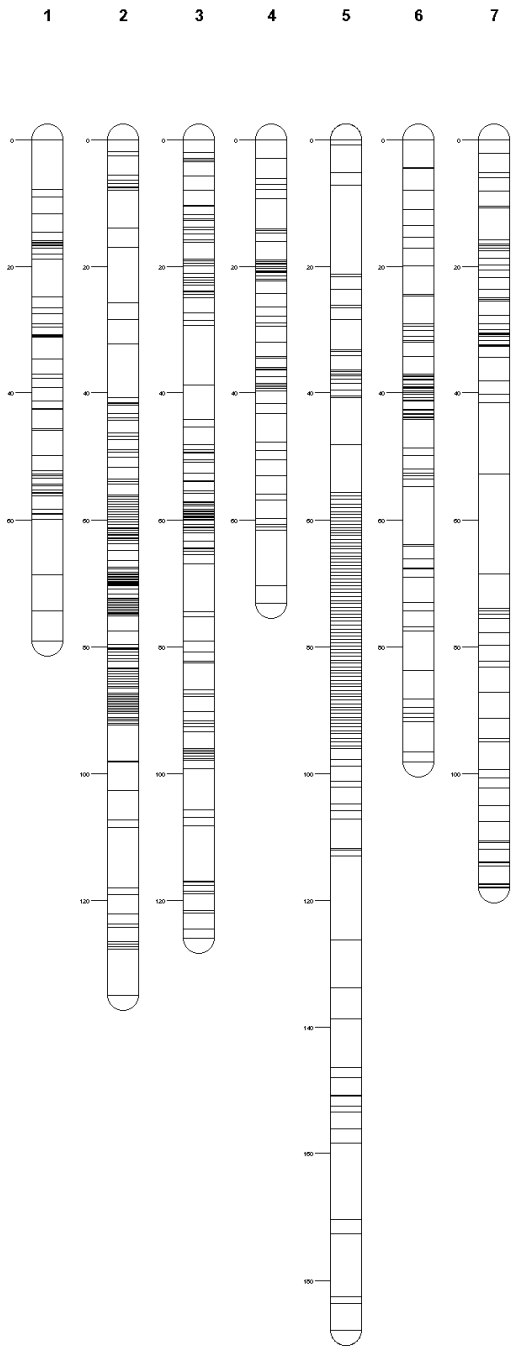
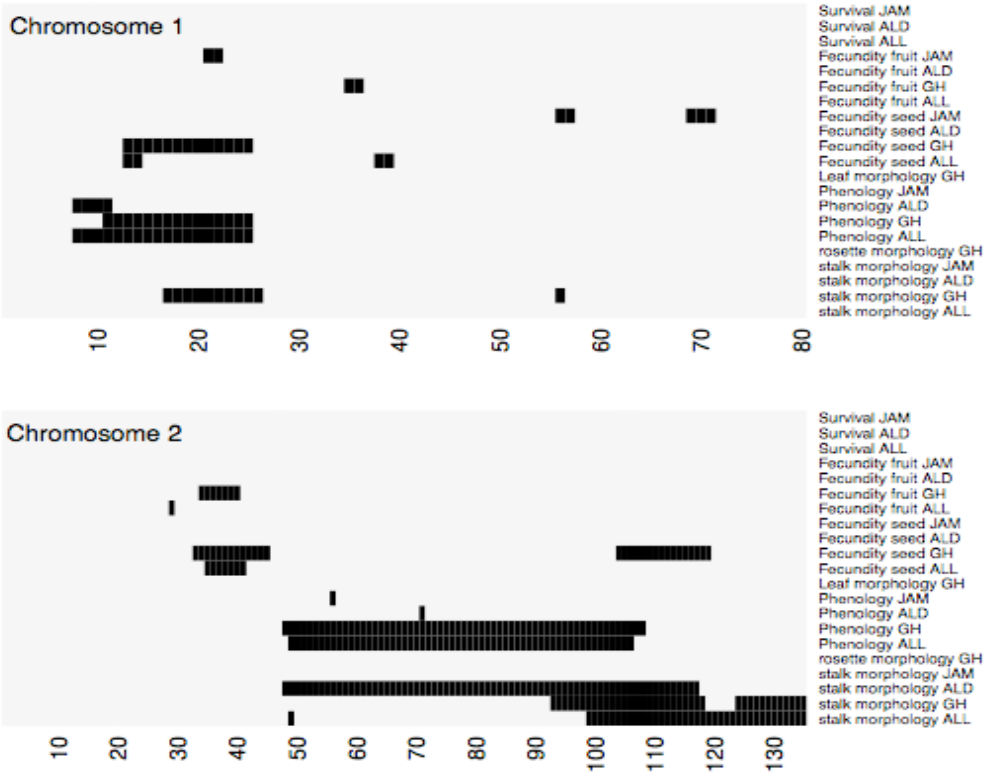
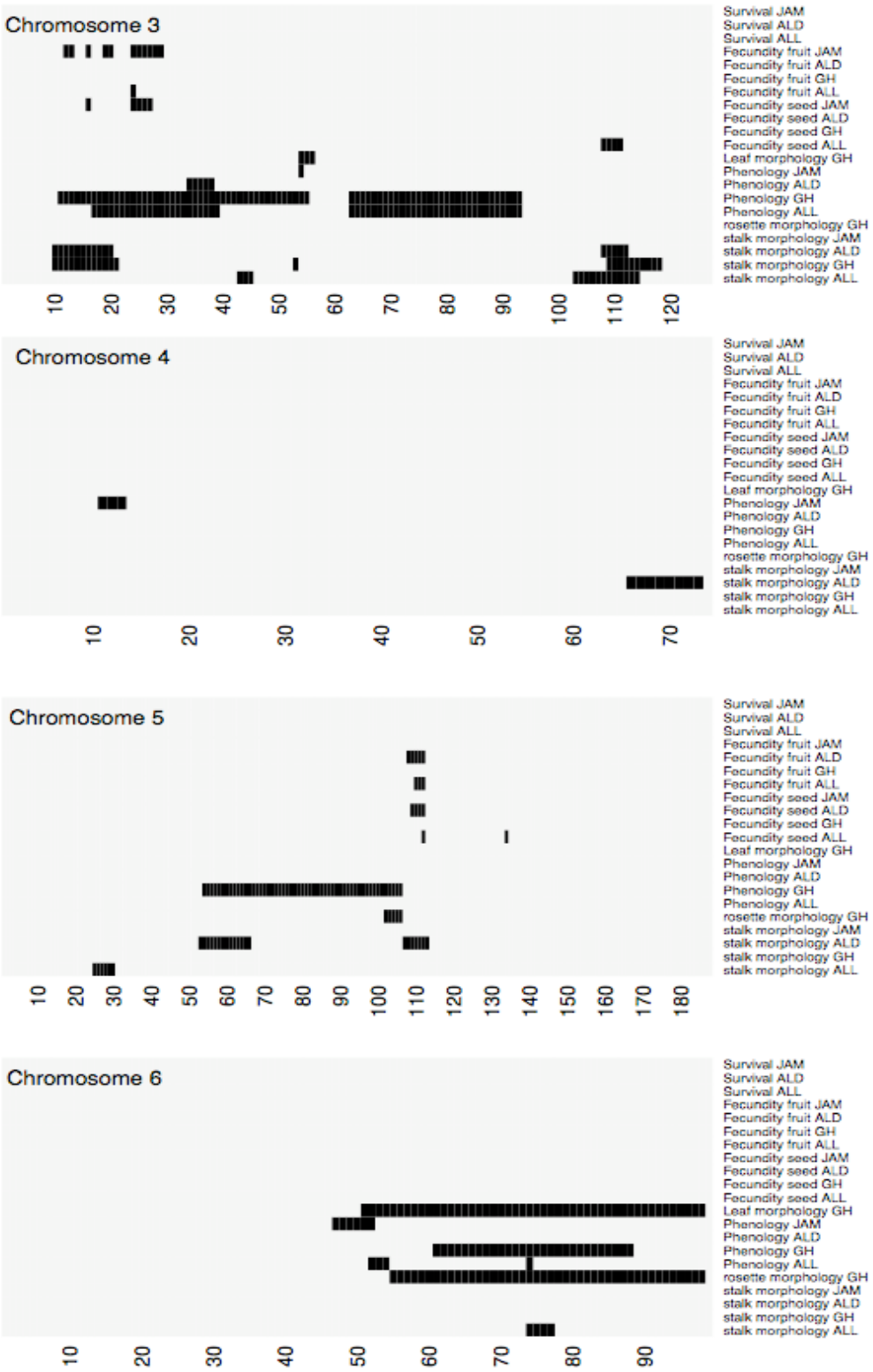


Figure 10: Linkage map of *Boechera stricta*. Horizontal lines on each linkage group represent genetic markers.

5.2.3 Quantitative trait loci for important traits

Multivariate least square interval mapping (MLSIM) identified many genomic regions controlling different trait categories (Figure 11). Three genomic regions are of considerable importance: 1) Chromosome 5, ~110 cM controls fruit and seed fitness components in the ALD (WEST) garden. This region also controls stalk morphology in ALD and is only a few cM away from a QTL controlling rosette morphology in the greenhouse. 2) Chromosome 6, 60-90 cM controls leaf morphology, rosette morphology, and phenology in the greenhouse. 3) Chromosome 7, 40-70 cM is a major QTL controlling stalk morphology and phenology. These three multivariate QTL also have large effects on individual univariate traits, and their effects are described in detail below.




Chromosome 3
Survival JAM
Survival ALD
Survival ALL
Fecundity fruit JAM
Fecundity fruit ALD
Fecundity fruit GH
Fecundity fruit ALL
Fecundity seed JAM
Fecundity seed ALD
Fecundity seed GH
Fecundity seed ALL
Leaf morphology GH
Phenology JAM
Phenology ALD
Phenology GH
Phenology ALL
rosette morphology GH
stalk morphology JAM
stalk morphology ALD
stalk morphology GH
stalk morphology ALL
10 20 30 40 50 60 70 80 90 100 110 120
Chromosome 4
10 20 30 40 50 60 70
Chromosome 5
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Chromosome 6
10 20 30 40 50 60 70 80 90

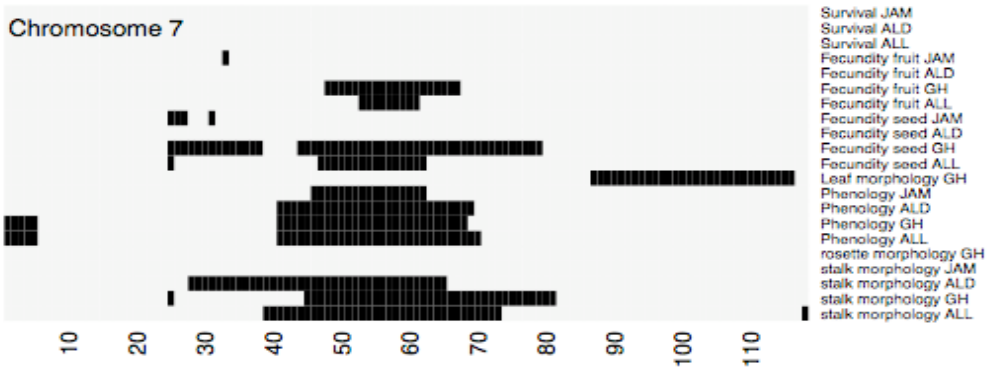


Figure 11: Multivariate least square interval mapping (MLSIM) result for each trait category. Around a QTL peak, the region where the statistical value is higher than the permutation significance threshold is marked in black.

For univariate QTL mapping, many separate QTL were identified (Appendix Table S11 and Figure 12), among which four major QTL have large effects, controlling ~20% or more of genetic variation for several traits. Interestingly, the four large-effect QTL either control fitness, or traits previously shown to be under natural selection in the field (Anderson *et al.* 2011a), or traits with high Q_{ST} (Lee & Mitchell-Olds 2013), and these four QTL have additive allelic effects consistent with patterns of parental subspecies divergence and local fitness advantage: 1) Chromosome 5, 110-120 cM confers local advantage (fecundity of fruiting plants) in the ALD garden; 2) Chromosome 6, 55-70 cM controls high-Q_{ST} traits such as leaf width/length ratio and succulence; 3) Chromosome 6, 85-95 cM controls leaf succulence and confers local advantage (fecundity of plants that survived the previous winter) in ALD garden; 4) Chromosome 7, 40-70 cM is a major phenology and stalk morphology QTL in the greenhouse.

All QTL for field fitness components show patterns of conditional neutrality, but I found no evidence for antagonistic pleiotropy – the QTL only have fitness effects in one field site but no significant effect in the other. Consistent with the trait-level

analyses in Table 7, although the EAST garden has two significant survival QTL and one overall fitness QTL, the WEST garden has no survival QTL and four QTL for various fecundity components of fitness. In the EAST garden, all three fitness QTL indicate adaptation to local conditions, with the local EAST allele conferring higher fitness. In the WEST garden, two QTL show adaptation to local conditions, and two additional QTL show higher fitness for the foreign EAST alleles. These two locally maladaptive QTL, however, have smaller effects than a locally adaptive QTL on chromosome 5, 110-120 cM (controlling ~20% of genetic variation in fitness). Therefore the overall effect of fitness QTL in the WEST garden still confers higher fitness for the local WEST parent.

For QTL controlling fitness components in the field, none showed statistical significance in both field gardens. Therefore, we did not find evidence of antagonistic pleiotropy for the field fitness QTL. One-way ANOVA analysis for the effect of fitness QTL identified in one field garden on the corresponding fitness component in the other garden (Table 8) shows lack of fitness effect of all fitness QTL in the other environment, consistent with the lack of clear antagonistic pleiotropy effect. These analyses, however, show only low to moderate statistical power ranging from (10% - 40%; Table 8), and therefore it is unclear whether the pattern shows true conditional neutrality or antagonistic pleiotropy with low statistical power in one environment.

Several QTL contribute to fitness in the greenhouse environment, and different parental alleles confer higher fitness in different QTL. Noticeably, there is no overlap between greenhouse and any field fitness QTL, despite the abundant water supply in the WEST garden and the greenhouse. This suggests that other environmental factors besides water availability control local adaptation in the WEST garden. It is also possible that I do not have enough power in the field experiments.

Table 8: One-way ANOVA and power analysis of fitness QTL identified in one field environment on the corresponding fitness components in the other garden

QTL peak	Garden of origin [a]	Fitness component [c]	Garden tested [b]	ANOVA *F*	ANOVA *P*	Power (%)
CH6cM90	WEST	Winter-survived plant fruit number	EAST	0.45	0.64	11.36
CH6cM90	WEST	Winter-survived plant fitness	EAST	0.43	0.65	11.15
CH5cM109	WEST	Fruited plant fruit number	EAST	0.33	0.72	12.46
CH5cM109	WEST	Fruited plant fitness	EAST	1.22	0.30	36.78
CH4cM49	EAST	Winter survival	WEST	0.61	0.55	13.97
CH4cM49	EAST	Overall survival	WEST	1.51	0.22	41.11
CH6cM17	EAST	Winter survival	WEST	0.53	0.59	15.33

a. The field environment where the QTL was identified
b. The other field environment where the corresponding fitness component was used for ANOVA and power analysis
c. Both fruit number and approximated seed number (fitness) were used

I was only able to measure leaf and rosette morphology in the greenhouse, and the major-effect QTL of these traits often co-localize. For leaf morphology, most QTL directions are consistent with the previous study of parental divergence, where the WEST allele confers greater width and width/length ratio. Many QTL of varying effects control rosette morphology traits, and the effects of most QTL are consistent with parental divergence, where the EAST allele confers smaller rosette size, weight, leaf area, and leaf packing, but higher rosette fresh weight and water weight per unit leaf area (more succulent). Of considerable importance is a QTL on chromosome 6, 55-70 cM. This QTL controls many leaf and rosette morphology traits that have the highest Q_{ST} among all traits measured between subspecies (Lee & Mitchell-Olds 2013). Therefore, this may be a candidate genomic region responsible for adaptive divergence between the subspecies. This QTL, however, does not control any other traits or fitness components in the field. It is possible that my field experiments do not capture all necessary selection forces or spatial/temporal environmental variation responsible for the subspecies-level adaptive

divergence. For example, due to logistics and time constraints, all plants were grown in the greenhouse for 10 weeks before transplantation. The field environmental selection during this 10-week period (where rosettes and leaves were developing and most likely to be under environmental selection) is therefore missing from my experiment. Another QTL 20 cM downstream, on the other hand, controls both leaf succulence in the greenhouse and plant fecundity in the ALD garden.

No stalk morphology QTL was identified in the JAM garden. Four QTL were identified in the ALD garden, two of which also control stalk morphology in the greenhouse. In the greenhouse, the EAST alleles generally confer taller but thinner stalks and longer internodes, consistent with previous subspecies-level comparison (Lee & Mitchell-Olds 2013). A genomic region in chromosome 7, 40-70 cM simultaneously controls many stalk morphology and phenology traits. Interestingly, this genomic region has opposite effects on final stalk height between greenhouse and WEST garden: the EAST allele has higher final stalk height in the greenhouse but lower in the WEST garden.

For phenology, I identified QTL with effects across all gardens, as well as those having effects only in specific gardens, and almost all QTL have the same direction across all environments, with EAST alleles showing faster phenology, more branching when flowering, and more rapid completion of development. A major phenology QTL in chromosome 7, 40-70 cM controls phenology in all three environments and stalk morphology in the greenhouse.

The QTL for composite traits are mostly consistent with their univariate trait components, and in most cases the EAST allele confers trait direction more similar to the EAST parent.

5.2.4 Co-localization of fitness and trait QTL

Some fitness QTL overlap with trait QTL in various environments, although it is not known whether the same underlying genes control both fitness and other quantitative traits. The most notable examples are phenology QTL. Chromosome 1, 40-50 cM controls phenology and fecundity in the greenhouse; Chromosome 3, 20-45 cM controls phenology in all three environments and fecundity in both field gardens; Chromosome 7, 40-70 cM is a major phenology QTL in all three gardens and a stalk morphology QTL in the greenhouse, and is also controls fitness in the greenhouse. In all three genomic regions where phenology and fitness QTL overlap, the QTL have consistent effects, with the EAST allele conferring faster phenology and higher fitness. For this QTL, the rapidly developing EAST allele appears to be advantageous whenever it controls fitness components, even in the WEST garden. Although previous study has shown that phenology, especially flowering time, is an important selective agent in *Boechera stricta* (Anderson *et al.* 2011a), here I do not find statistically significant evidence that phenology QTL contribute to differential local adaptation between EAST and WEST subspecies.

Chromosome 5, 110-120 cM is a major fitness QTL in ALD, with the WEST allele conferring higher fitness in the WEST environment. In MLSIM, this QTL controls stalk morphology in ALD and rosette morphology in the greenhouse (Figure 11). For univariate traits, it influences the number of reproductive branches in ALD, rosette number in the greenhouse, and the stalk length with reproductive branches in the greenhouse. These traits, however, are not among the traits with highest Q_{ST} from my previous study (Lee & Mitchell-Olds 2013), and it is not clear whether those traits are adaptive in the ALD garden. On the other hand, it is possible that the multivariate trait components

controlled by this QTL represent the linear combination of important rosette morphology traits such as leaf succulence and rosette packing, and its effect is not large enough be to identified in univariate trait mapping.

Chromosome 6, 85-95 cM controls two important traits with high Q_{ST} (Lee & Mitchell-Olds 2013), and the direction of allelic divergence is consistent with the subspecies-level expectation, where the EAST allele has higher fresh weight and water weight per unit leaf area, showing a more succulent and xerophytic vegetative morphology. The WEST allele of this QTL is also locally advantageous in the WEST garden. Noticeably, the fitness component controlled by this QTL is 'winter-survival plant fitness' in the ALD garden, and it is possible that this QTL (and leaf succulence) only control fitness in the summer growing season without effects on winter survival. The over-winter survival in JAM garden is not high (46%; other plants were identified as dead [33%], and 21% were buried by landslide and were counted as missing). Consequently, I found zero heritability for 'winter-survival plant fitness' in JAM, presumably due to the lack of statistical power. As a consequence, I was unable to map QTL for the same fitness component in JAM garden, and therefore it is not clear whether this QTL, which controls high-Q_{ST} traits and fitness in the field, is an example of true conditional neutrality or an antagonistically pleiotropic QTL suffering from lack of power in the JAM garden.

In summary, only a few cases of colocalization between QTL controlling fitness and high-Q_{ST} traits were identified. For large-effect fitness QTL without trait effect, it is possible that other important traits for local adaptation were not measured, such as the overall resource allocation to roots or the root system architecture. In addition, the methods of field experimentation may contribute to the lack of fitness effects in QTL with large trait effect. In the natural environment, *B. stricta* is a short-lived perennial

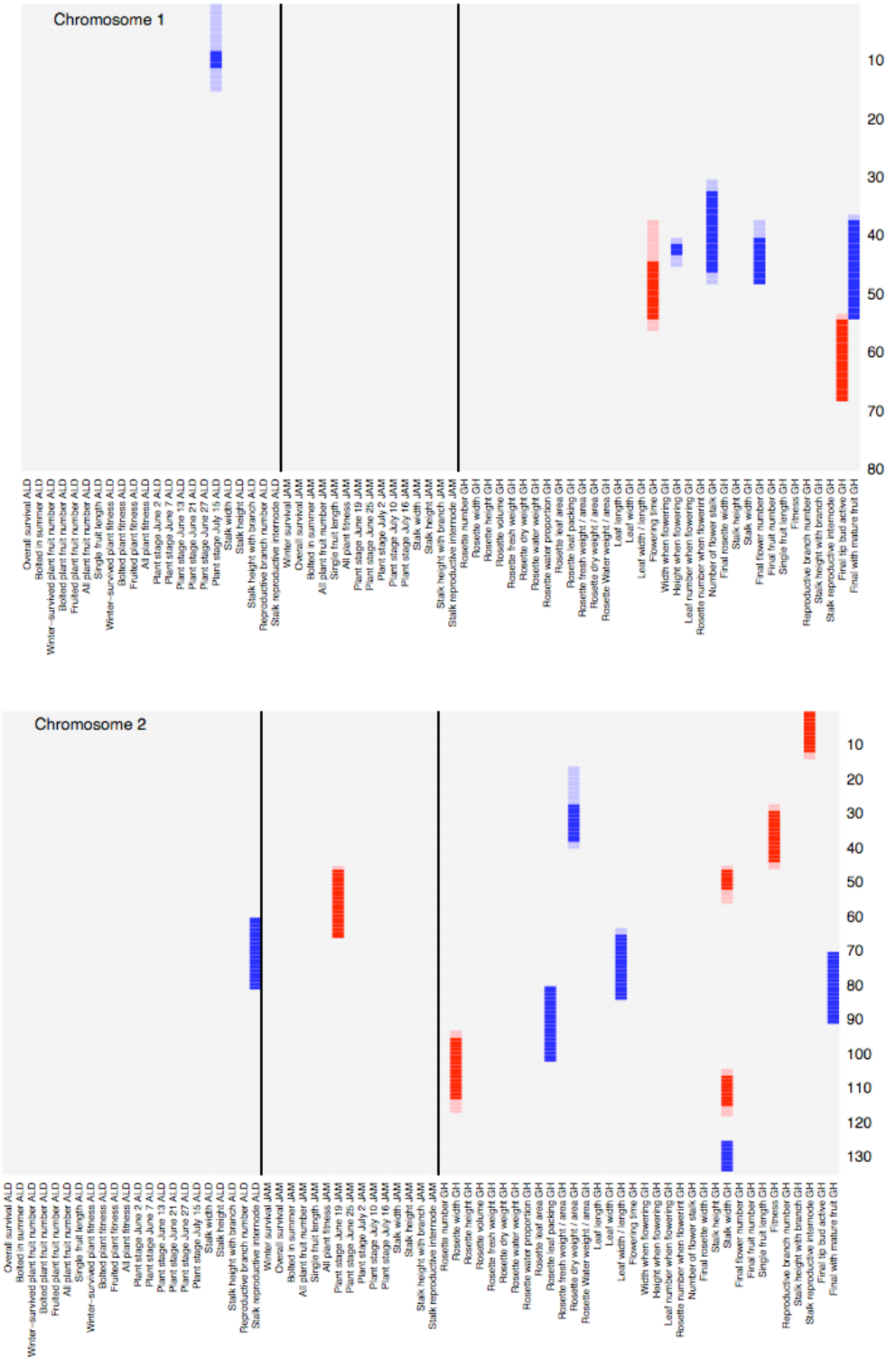
spending multiple years as rosette. In my experiment, I transplanted fully-grown rosettes in fall and measured fitness output in the next summer, and therefore the experimental plants only experienced natural selection during the late rosette stage and the reproductive stage. Since many of the high-Q_{ST} traits belong to leaf and rosette morphology, the lack of fitness effect in trait QTL may be due to the logistic constraints of transplanting young seedlings to the field environment and measuring fitness in the early rosette stage.

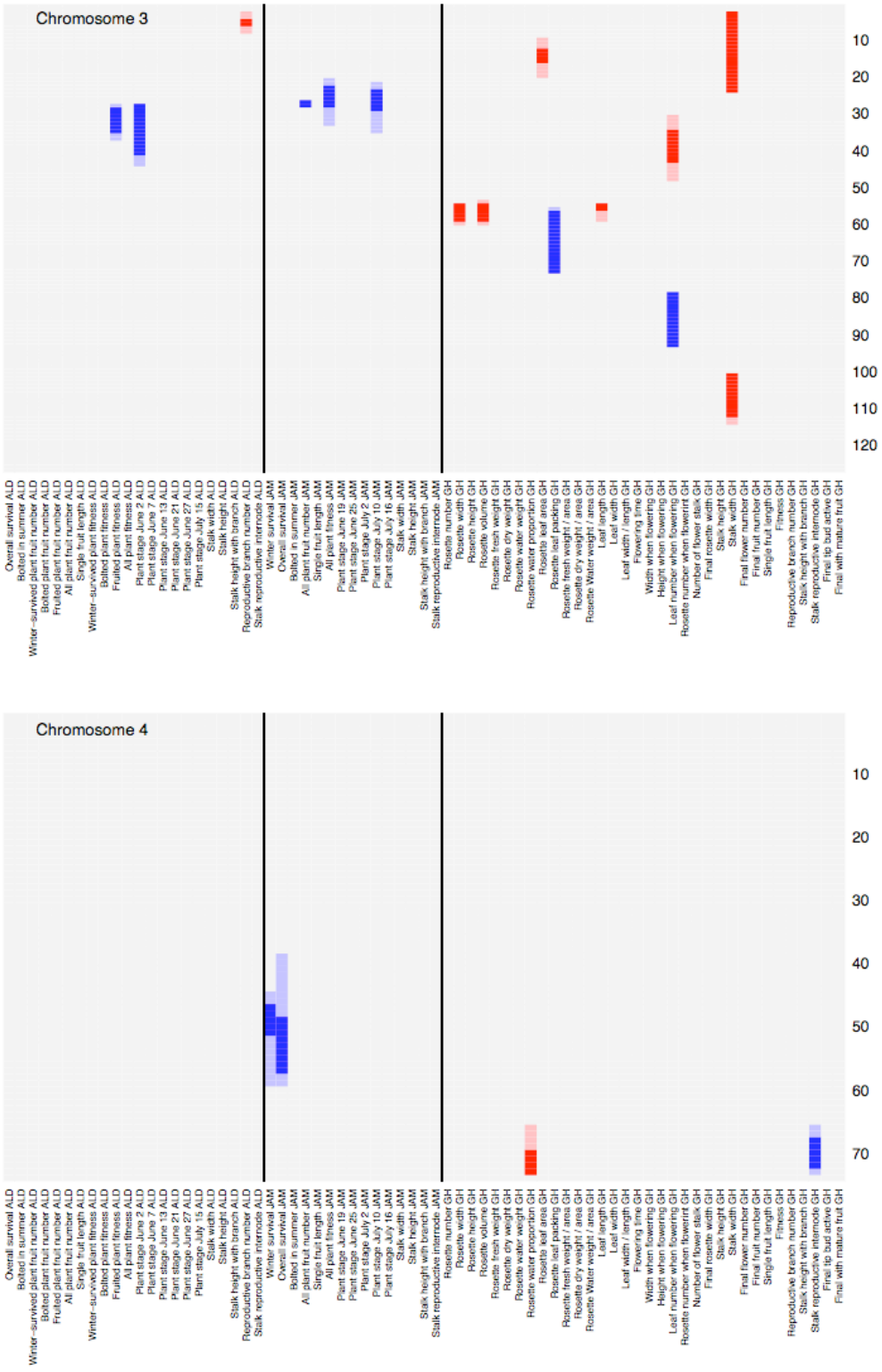
5.2.5 Conclusion

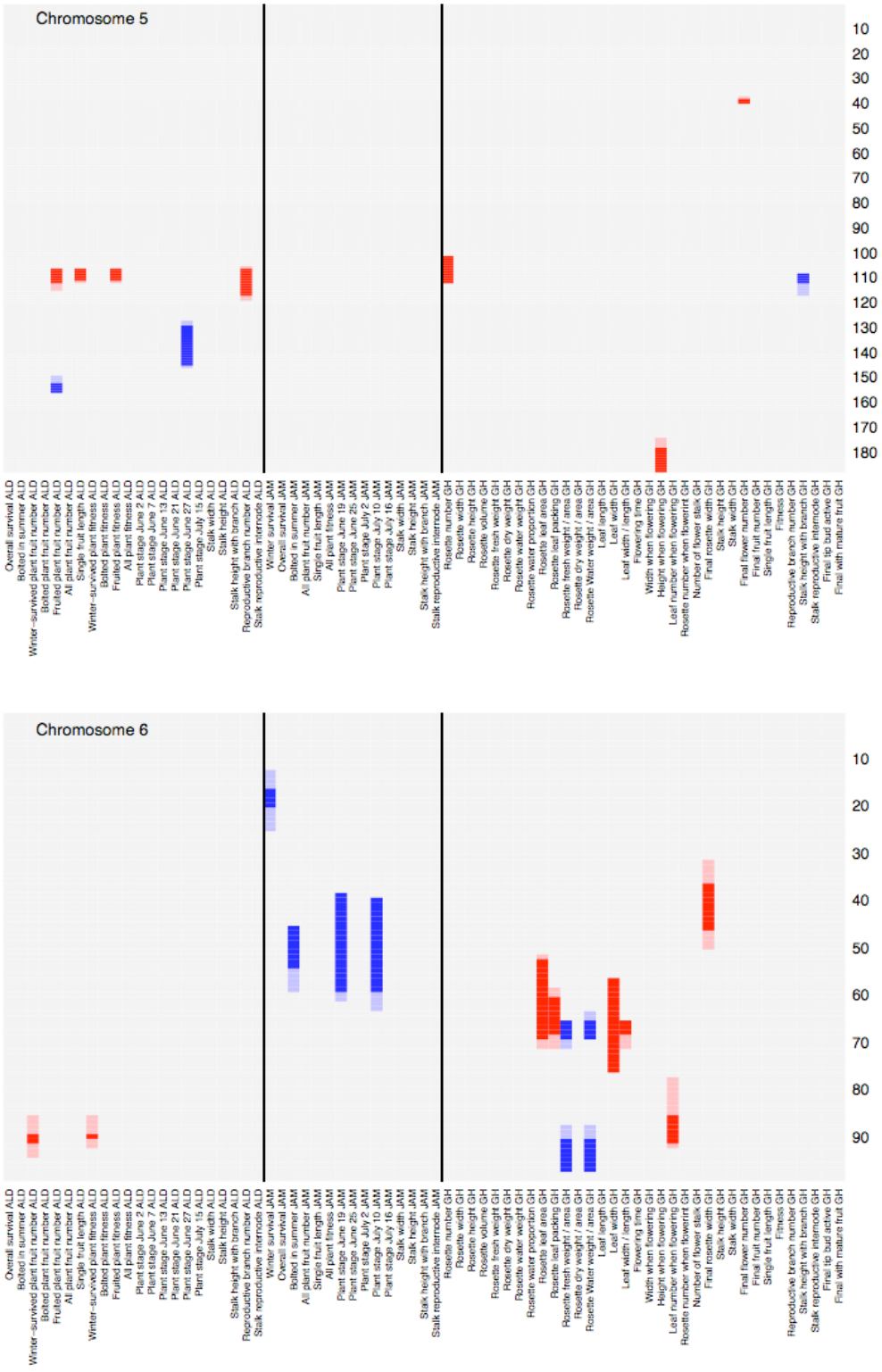
Ecological speciation refers to the speciation process where distinct natural environments cause reproductive isolation by selecting against unfit immigrants or hybrids, and the pattern or effects of loci controlling local fitness may differ depending on the mode of speciation (sympatric, parapatric, or allopatric) and the breeding system of organisms (outcrossing or self-fertilizing). My study shows different water regime, types of natural selection, trait response, and QTL underlying the local adaptation between EAST and WEST subspecies of *Boechera stricta*: In the harsh EASTERN native environment with drought stress, survival is the major force of natural selection, and the EAST subspecies employed life history strategies for drought adaptation to maximize survival. In the benign WESTERN native environment with abundant water, fecundity is the major determinant of lifetime fitness, and the WEST subspecies employed strategies that increase fecundity. Therefore, different life history strategies have evolved independently between subspecies during the allopatric stage of speciation. This pattern of speciation and adaptive divergence, together with the lower chance of hybridization during secondary contact due to the predominantly self-fertilizing breeding system,

suggest different loci may be responsible for local fitness in the EAST or WEST environment.

Consistent with expectation, I do not identify clear patterns of antagonistic pleiotropy on fitness QTL: those QTL control fitness in only one of the two field environments. This pattern, however, can be due to the low to moderate statistical power in the field environments, and therefore I am unable to distinguish whether the observed patterns are true conditional neutrality or possible antagonistic pleiotropy with low power. Nevertheless, conditional neutrality is not unexpected given the pattern of secondary contact after historical allopatry in *B. stricta*. With the high self-fertilization rate of *B. stricta*, it is possible that the most important factor limiting EAST-WEST gene flow is natural selection against unfit immigrants rather than unfit hybrids, and immigrant allele may be eliminated by natural selection, which acts on the immigrant genome as a whole, before hybridization could occur. In this situation, the combined effect of many conditionally neutral QTL may contribute to ecological speciation by preventing the successful immigration between natural subspecies habitats.


Chromosome 1
Chromosome 2
10
20
30
40
50
60
70
80
90
100
110
120
130
Overall survival ALD
Bolted in summer ALD
Winter−survived plant fruit number ALD
Bolted plant fruit number ALD
Fruited plant fruit number ALD
All plant fruit number ALD
Single fruit length ALD
Winter−survived plant fitness ALD
Bolted plant fitness ALD
Fruited plant fitness ALD
All plant fitness ALD
Plant stage June 2 ALD
Plant stage June 7 ALD
Plant stage June 13 ALD
Plant stage June 21 ALD
Plant stage June 27 ALD
Plant stage July 15 ALD
Stalk width ALD
Stalk height ALD
Reproductive branch number ALD
Stalk reproductive internode ALD
Winter survival JAM
Overall survival JAM
Bolted in summer JAM
All plant fruit number JAM
Single fruit length JAM
All plant fitness JAM
Plant stage June 19 JAM
Plant stage June 25 JAM
Plant stage July 2 JAM
Plant stage July 10 JAM
Plant stage July 16 JAM
Stalk width JAM
Stalk height JAM
Stalk height with branch JAM
Stalk reproductive internode JAM
Rosette number GH
Rosette width GH
Rosette height GH
Rosette volume GH
Rosette fresh weight GH
Rosette dry weight GH
Rosette water weight GH
Rosette water proportion GH
Rosette leaf area GH
Rosette leaf packing GH
Rosette fresh weight / area GH
Rosette dry weight / area GH
Rosette Water weight / area GH
Leaf length GH
Leaf width GH
Leaf width / length GH
Flowering time GH
Width when flowering GH
Height when flowering GH
Leaf number when flowering GH
Rosette number when flowerint GH
Number of flower stalk GH
Final rosette width GH
Stalk height GH
Stalk width GH
Final flower number GH
Final fruit number GH
Single fruit length GH
Fitness GH
Reproductive branch number GH
Stalk height with branch GH
Stalk reproductive internode GH
Final tip bud active GH
Final with mature fruit GH


Chromosome 3
10
20
30
40
50
60
70
80
90
100
110
120
Overall survival ALD
Bolted in summer ALD
Winter−survived plant fruit number ALD
Bolted plant fruit number ALD
Fruited plant fruit number ALD
All plant fruit number ALD
Single fruit length ALD
Winter−survived plant fitness ALD
Bolted plant fitness ALD
Fruited plant fitness ALD
All plant fitness ALD
Plant stage June 2 ALD
Plant stage June 7 ALD
Plant stage June 13 ALD
Plant stage June 21 ALD
Plant stage June 27 ALD
Plant stage July 15 ALD
Stalk width ALD
Stalk height ALD
Stalk height with branch ALD
Reproductive branch number ALD
Stalk reproductive internode ALD
Winter survival JAM
Overall survival JAM
Bolted in summer JAM
All plant fruit number JAM
Single fruit length JAM
All plant fitness JAM
Plant stage June 19 JAM
Plant stage June 25 JAM
Plant stage July 2 JAM
Plant stage July 10 JAM
Plant stage July 15 JAM
Stalk width JAM
Stalk height JAM
Stalk height with branch JAM
Stalk reproductive internode JAM
Rosette number GH
Rosette width GH
Rosette height GH
Rosette volume GH
Rosette fresh weight GH
Rosette dry weight GH
Rosette water weight GH
Rosette water proportion GH
Rosette leaf area GH
Rosette leaf packing GH
Rosette fresh weight / area GH
Rosette dry weight / area GH
Rosette Water weight / area GH
Leaf length GH
Leaf width GH
Leaf width / length GH
Flowering Time GH
Width when flowering GH
Height when flowering GH
Leaf number when flowering GH
Rosette number when flowerint GH
Number of flower stalk GH
Final rosette width GH
Stalk height GH
Stalk width GH
Final flower number GH
Final fruit number GH
Single fruit length GH
Fitness GH
Reproductive branch number GH
Stalk height with branch GH
Stalk reproductive internode GH
Final tip bud active GH
Final with mature fruit GH
Chromosome 4
10
20
30
40
50
60
70
Overall survival ALD
Bolted in summer ALD
Winter−survived plant fruit number ALD
Bolted plant fruit number ALD
Fruited plant fruit number ALD
All plant fruit number ALD
Single fruit length ALD
Winter−survived plant fitness ALD
Bolted plant fitness ALD
Fruited plant fitness ALD
All plant fitness ALD
Plant stage June 2 ALD
Plant stage June 7 ALD
Plant stage June 13 ALD
Plant stage June 21 ALD
Plant stage June 27 ALD
Plant stage July 15 ALD
Stalk width ALD
Stalk height ALD
Stalk height with branch ALD
Reproductive branch number ALD
Stalk reproductive internode ALD
Winter survival JAM
Overall survival JAM
Bolted in summer JAM
All plant fruit number JAM
Single fruit length JAM
All plant fitness JAM
Plant stage June 19 JAM
Plant stage June 25 JAM
Plant stage July 2 JAM
Plant stage July 10 JAM
Plant stage July 15 JAM
Stalk width JAM
Stalk height JAM
Stalk height with branch JAM
Stalk reproductive internode JAM
Rosette number GH
Rosette width GH
Rosette height GH
Rosette volume GH
Rosette fresh weight GH
Rosette dry weight GH
Rosette water weight GH
Rosette water proportion GH
Rosette leaf area GH
Rosette leaf packing GH
Rosette fresh weight / area GH
Rosette dry weight / area GH
Rosette Water weight / area GH
Leaf length GH
Leaf width GH
Leaf width / length GH
Flowering Time GH
Width when flowering GH
Height when flowering GH
Leaf number when flowering GH
Rosette number when flowerint GH
Number of flower stalk GH
Final rosette width GH
Stalk height GH
Stalk width GH
Final flower number GH
Final fruit number GH
Single fruit length GH
Fitness GH
Reproductive branch number GH
Stalk height with branch GH
Stalk reproductive internode GH
Final tip bud active GH
Final with mature fruit GH


Chromosome 5
Chromosome 6
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
Overall survival ALD
Bolted in summer ALD
Winter−survived plant fruit number ALD
Bolted plant fruit number ALD
Fruited plant fruit number ALD
All plant fruit number ALD
Single fruit length ALD
Winter−survived plant fitness ALD
Bolted plant fitness ALD
Fruited plant fitness ALD
All plant fitness ALD
Plant stage June 2 ALD
Plant stage June 7 ALD
Plant stage June 13 ALD
Plant stage June 21 ALD
Plant stage June 27 ALD
Plant stage July 15 ALD
Stalk width ALD
Stalk height ALD
Stalk height with branch ALD
Reproductive branch number ALD
Stalk reproductive internode ALD
Winter survival JAM
Overall survival JAM
Bolted in summer JAM
All plant fruit number JAM
Single fruit length JAM
All plant fitness JAM
Plant stage June 19 JAM
Plant stage June 25 JAM
Plant stage July 2 JAM
Plant stage July 10 JAM
Plant stage July 15 JAM
Stalk width JAM
Stalk height JAM
Stalk height with branch JAM
Stalk reproductive internode JAM
Rosette number GH
Rosette width GH
Rosette height GH
Rosette volume GH
Rosette fresh weight GH
Rosette dry weight GH
Rosette water weight GH
Rosette water proportion GH
Rosette leaf area GH
Rosette leaf packing GH
Rosette fresh weight / area GH
Rosette dry weight / area GH
Rosette Water weight / area GH
Leaf length GH
Leaf width GH
Leaf width / length GH
Flowering Time GH
Width when flowering GH
Height when flowering GH
Leaf number when flowering GH
Rosette number when flowering GH
Number of flower stalk GH
Final rosette width GH
Stalk height GH
Stalk width GH
Final flower number GH
Final fruit number GH
Single fruit length GH
Fitness GH
Reproductive branch number GH
Stalk height with branch GH
Stalk reproductive internode GH
Final tip bud active GH
Final with mature fruit GH

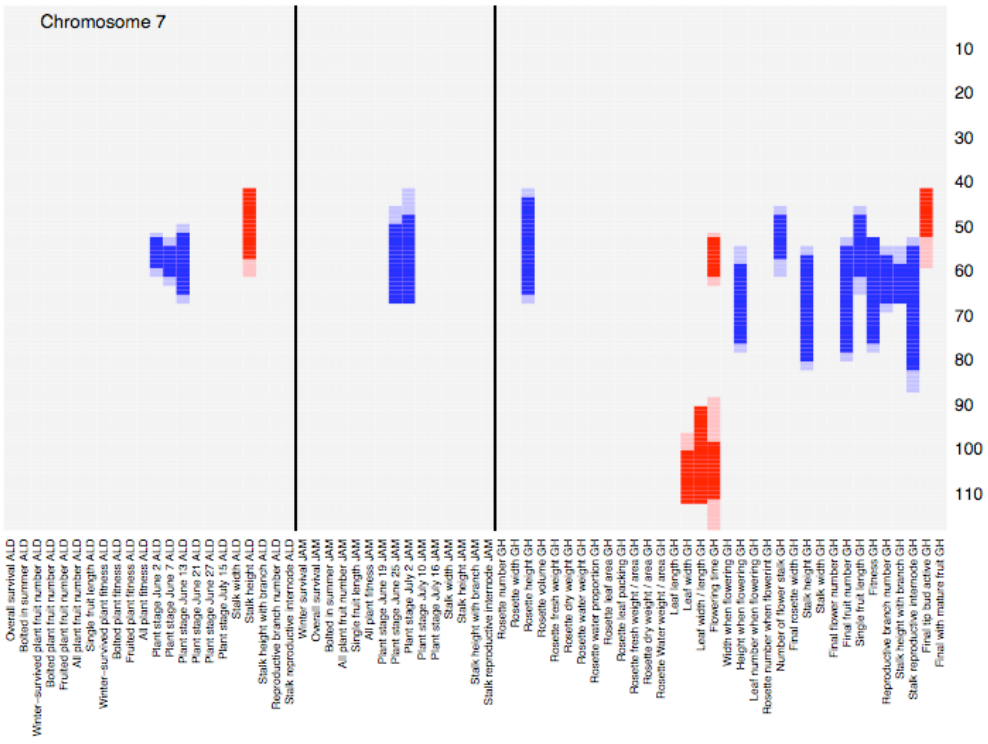


Figure 12: Quantitative trait loci (QTL) of univariate traits in three environments on seven *Boechera stricta* chromosomes. Each graph represents chromosome 1 to 7 in order. Within each graph, columns are univariate traits where three environments are separated by two vertical black lines, and rows are centi-Morgan on the linkage map. QTL and confidence intervals are presented as colored bars, where blue means the Parker (EAST subspecies) allele has higher trait value and red means the Ruby (WEST subspecies) allele has higher trait value. Darker red or blue region represents 1-LOD confidence interval, and lighter red or blue region represents 2-LOD confidence interval.

Appendix A. Supplementary tables

Table S 1: Predictor variable used in Chapter 2

Variable [a]	Type [b]	Classification [c]
Environmental relevance	Continuous	ENV
Duplication status	Categorical	DUP
Chromosome	Categorical	PHY
Recombination rate	Continuous	PHY
Chromosome position	Continuous	PHY
GC content	Continuous	PHY
5′ UTR length	Continuous	PHY
3′ UTR length	Continuous	PHY
Coding sequence length	Continuous	PHY
Intron number	Continuous	PHY
Average intron length	Continuous	PHY
dSM	Continuous	FUN
Expression level	Continuous	FUN
Tissue specificity	Continuous	FUN
Fop	Continuous	FUN
Multifunctionality	Continuous	FUN

a. Except for environmental relevance, chromosome, and coding sequence length, most variables are adopted from: Yang and Gaut. 2011. Factors that Contribute to Variation in Evolutionary Rate among Arabidopsis Genes. Mol Biol Evol 28(8):2359-2369.
b. Indicates whether the variable is used as a continuous or categorical variable in the statistical model
c. The four major groups of predictor variables used in this study: ENV – environment, DUP – duplication status, PHY – physical property, FUN – functional constraint.

Table S 2: Twenty environmental variables used to estimate the environmental relevance of each gene in Chapter 2

Name	Description	Category
Alt	Altitude	Altitude
BIO1	Annual Mean Temperature	Temperature
BIO2	Mean Diurnal Range	Temperature Variation
BIO3	Isothermality	Temperature Variation
BIO4	Temperature Seasonality	Temperature Variation
BIO5	Max Temperature of Warmest Month	Temperature
BIO6	Min Temperature of Coldest Month	Temperature
BIO7	Temperature Annual Range	Temperature Variation
BIO8	Mean Temperature of Wettest Quarter	Temp*Prec Interaction
BIO9	Mean Temperature of Driest Quarter	Temp*Prec Interaction
BIO10	Mean Temperature of Warmest Quarter	Temperature
BIO11	Mean Temperature of Coldest Quarter	Temperature
BIO12	Annual Precipitation	Precipitation
BIO13	Precipitation of Wettest Month	Precipitation
BIO14	Precipitation of Driest Month	Precipitation
BIO15	Precipitation Seasonality	Precipitation Variation
BIO16	Precipitation of Wettest Quarter	Precipitation
BIO17	Precipitation of Driest Quarter	Precipitation
BIO18	Precipitation of Warmest Quarter	Temp*Prec Interaction
BIO19	Precipitation of Coldest Quarter	Temp*Prec Interaction

Table S 3: Seventeen microsatellite loci and their primer sequences used in Chapter 3

Loci name	Forward primer [a]	Reverse primer
BF-11	TCCTCCATTGTAGAGCAGAGC	CCATTGCTTAAACCCTAAACC
BF-3	TTTTTAGACAGTAGTGGCTGTGAG	ACTTCGTTCCAGGCTCGTC
BF-20	TTCTCGGGAAAGTAATGAGGAG	GCAAATCTGACCAATGCAAG
BF-9	AAACACATTCCCGTCAGCTC	TTGATTGAATCCTGCGTTTG
c8	TTCCGGGTATCATTCCTAG	GTTGTAAGTTCTTTCTCAG
Bdru-266	TTTAATTTGTGCGTTTGATCC	CAAAATCGCAGAATGAGAGG
BF-15	CAGCATCTCCTTTGGGTTTG	ACTTGCTCCTTTGCATGACC
BF-19	ACCGCATTGGTGTTGTGTC	ATAACGGACGCGACCAAAG
d3	GGTTATGTGAGAGTTAAG	ATTGTTGAATGCAACAGG
Bdru-1220	TCTATGCAAACAGCAAATCG	TTCTTCTACGAAACATTCCTTGC
ca72	AATCCCAGTAACCAAACACACA	CCCAGTCTAACCACGACCAC
ICE3	GACTAATCATCACCGACTCAGCCAC	ATTCTTCTTCACTTTTCTTGATCCCG
Bdru-878	GGGAACCTTCATGTCCAAAG	TGCCTTTTCCGTTTTTCTAATC
ICE11	TTTCAAGTTGAGAAGTGGAGTG	AAGAATTAGGCAAGAGTTTAGTGG
e9	AGGAAAGGACAAAAGACATG	GCTTCCATGGAAGGAGACCC
ICE14	TCGAGGTGCTTTCTGAGGTT	TACCTCACCCTTTTGACCCA
a3	AGCTTTGTTTGCAATGGAG	GTGAGAATAATATTGACC

a. Only original primer sequences are shown. M13 sequences (CACGACGTTGTAAAACGAC) were added to the 5′ end of forward primers.

Table S 4: Environmental variables used in Chapter 3

Variable name	Variable category	Explanation	Source
BIO1	ENV-CLIM	Annual mean temperature	WorldClim
BIO4	ENV-CLIM	Temperature seasonality	WorldClim
BIO12	ENV-CLIM	Annual precipitation	WorldClim
BIO15	ENV-CLIM	Precipitation seasonality	WorldClim
BIO18	ENV-CLIM	Precipitation of warmest quarter	WorldClim
BIO19	ENV-CLIM	Precipitation of coldest quarter	WorldClim
ALT	ENV-TOPO	Altitude	WorldClim
CTI	ENV-TOPO	Compound topographic index	USGS-Hydro1k
SLOPE	ENV-TOPO	Slope	USGS-Hydro1k
NDS	ENV-TOPO	Distance to the nearest stream	Google Earth
LAT	GEO	Latitude	GPS
LON	GEO	Longitude	GPS

Table S 5: Trait divergence between subspecies. For all traits in 24 genotypes, shown are trait, category, *P*-value for subspecies in ANOVA, the subspecies with higher trait value, Q_{ST}, *P*-value of Q_{ST} compared to empirical F_{ST} distribution, and the correlation with discriminant function analysis (DFA) score from each trait category.

Trait	Category	*P*-value [a]	Higher [b]	Q_{ST}	Q_{ST} *P* [c]	DFA-cor [d]
Instantaneous WUE dry	Physiology	0.256	EAST	-	-	-
Instantaneous WUE wet	Physiology	0.429	WEST	0.00	1.000	-
Long-term $\Delta^{13}C$ dry	Physiology	0.050	WEST	0.22	0.358	-
Long-term $\Delta^{13}C$ wet	Physiology	0.202	WEST	0.05	0.616	-
Bolting time	Phenology	**< 0.001**	WEST	0.66	0.129	0.85
Flowering time	Phenology	**0.008**	WEST	0.42	0.217	0.60
Flowering duration	Phenology	**< 0.001**	WEST	0.63	0.142	0.68
Flowering rosette width	Phenology	0.591	EAST	0.00	1.000	-0.04
Flowering height	Phenology	**< 0.001**	EAST	0.78	0.076	-0.85
Flowering leaf number	Phenology	0.032	WEST	0.29	0.290	0.35
Flowering rosette number	Phenology	0.040	EAST	0.27	0.309	-0.38
Flowering stalk number	Phenology	**0.006**	EAST	0.48	0.195	-0.43
Main stalk diameter	Morph-stalk	**< 0.001**	WEST	0.71	0.110	0.73
Reproductive branch number	Morph-stalk	**0.020**	WEST	0.21	0.369	0.23
Main stalk height	Morph-stalk	**< 0.001**	EAST	0.84	0.061	-0.82
Stalk height containing branch	Morph-stalk	**< 0.001**	EAST	0.68	0.120	-0.61
Internode between branch	Morph-stalk	**< 0.001**	EAST	0.87	0.061	-0.91
Rosette width	Morph-rosette	0.189	WEST	0.08	0.570	0.23
Rosette height	Morph-rosette	0.901	EAST	0.00	1.000	-0.02
Rosette volume	Morph-rosette	0.369	WEST	0.00	1.000	0.17
Rosette leaf number	Morph-rosette	0.117	EAST	0.21	0.369	-0.23
Rosette fresh weight	Morph-rosette	**0.001**	WEST	0.64	0.138	0.47
Rosette dry weight	Morph-rosette	**0.001**	WEST	0.67	0.123	0.46
Rosette water weight	Morph-rosette	**0.002**	WEST	0.63	0.142	0.47
Rosette water proportion	Morph-rosette	0.475	EAST	0.00	1.000	-0.13

Rosette leaf area	Morph-rosette	**< 0.001**	WEST	0.73	0.107	0.56
Rosette leaf packing	Morph-rosette	**0.002**	WEST	0.50	0.185	0.54
Rosette fresh weight / area	Morph-rosette	**0.002**	EAST	0.53	0.174	-0.47
Rosette dry weight / area	Morph-rosette	0.282	EAST	0.02	0.718	-0.17
Rosette water weight / area	Morph-rosette	**0.001**	EAST	0.53	0.175	-0.49
Leaf shape Y1	Morph-leaf	0.492	WEST	0.00	1.000	0.07
Leaf shape Y2	Morph-leaf	**0.025**	WEST	0.59	0.158	0.26
Leaf shape Y3	Morph-leaf	**< 0.001**	WEST	0.92	**0.029**	0.54
Leaf shape Y4	Morph-leaf	**< 0.001**	WEST	0.94	**0.029**	0.80
Leaf shape Y5	Morph-leaf	**< 0.001**	WEST	0.93	**0.029**	0.87
Leaf shape Y6	Morph-leaf	**< 0.001**	WEST	0.93	**0.029**	0.90
Leaf shape Y7	Morph-leaf	**< 0.001**	WEST	0.94	**0.029**	0.94
Leaf shape Y8	Morph-leaf	**< 0.001**	WEST	0.94	**0.029**	0.95
Leaf shape Y9	Morph-leaf	**< 0.001**	WEST	0.90	**0.036**	0.92

a. Boldface denotes significant *P*-values after sequential Bonferroni correction within each trait category.
b. This column denotes which subspecies has the higher trait value.
c. Empirical *P*-values, based on the proportion of SNPs with F_{ST} higher than or equal to Q_{ST} for this trait. Boldface denotes values less than 0.05.
d. Pearson's correlation coefficient between trait value and the DFA score of each trait category. DFA is not available for physiology traits.

Table S 6: Trait divergence between subspecies. For all traits in 19 genotypes, shown are trait, category, *P*-value for subspecies in ANOVA, the subspecies with higher trait value, Q_{ST}, *P*-value of Q_{ST} compared to empirical F_{ST} distribution, and the correlation with discriminant function analysis (DFA) score from each trait category.

Trait	Category	*P*-value [a]	Higher [b]	Q_{ST}	Q_{ST} *P* [c]	DFA-cor [d]
Instantaneous WUE dry	Physiology	0.346	EAST	-	-	-
Instantaneous WUE wet	Physiology	0.670	WEST	0.00	1.000	-
Long-term $\Delta^{13}C$ dry	Physiology	0.038	WEST	0.33	0.270	-
Long-term $\Delta^{13}C$ wet	Physiology	0.071	WEST	0.26	0.314	-
Bolting time	Phenology	**0.001**	WEST	0.66	0.128	0.84
Flowering time	Phenology	0.065	WEST	0.29	0.290	0.44
Flowering duration	Phenology	**< 0.001**	WEST	0.75	0.103	0.63
Flowering rosette width	Phenology	0.983	EAST	0.00	1.000	0.09
Flowering height	Phenology	**< 0.001**	EAST	0.74	0.107	-0.83
Flowering leaf number	Phenology	0.072	WEST	0.29	0.315	0.28
Flowering rosette number	Phenology	0.096	EAST	0.06	0.614	-0.29
Flowering stalk number	Phenology	**0.010**	EAST	0.31	0.276	-0.39
Main stalk diameter	Morph-stalk	**< 0.001**	WEST	0.66	0.123	0.65
Reproductive branch number	Morph-stalk	**0.044**	WEST	0.18	0.406	0.22
Main stalk height	Morph-stalk	**< 0.001**	EAST	0.82	0.062	-0.78
Stalk height containing branch	Morph-stalk	**0.002**	EAST	0.59	0.157	-0.55
Internode between branch	Morph-stalk	**< 0.001**	EAST	0.88	**0.048**	-0.87
Rosette width	Morph-rosette	0.149	WEST	0.11	0.501	0.27
Rosette height	Morph-rosette	0.618	EAST	0.00	1.000	-0.12
Rosette volume	Morph-rosette	0.400	WEST	0.00	1.000	0.17
Rosette leaf number	Morph-rosette	0.743	EAST	0.00	1.000	-0.03
Rosette fresh weight	Morph-rosette	**0.001**	WEST	0.72	0.107	0.53
Rosette dry weight	Morph-rosette	**< 0.001**	WEST	0.77	0.088	0.52
Rosette water weight	Morph-rosette	**0.001**	WEST	0.71	0.109	0.52
Rosette water proportion	Morph-rosette	0.468	EAST	0.00	1.000	-0.16

Rosette leaf area	Morph-rosette	**< 0.001**	WEST	0.79	0.071	0.60
Rosette leaf packing	Morph-rosette	**0.004**	WEST	0.56	0.167	0.60
Rosette fresh weight / area	Morph-rosette	0.025	EAST	0.42	0.216	-0.40
Rosette dry weight / area	Morph-rosette	0.631	EAST	0.00	1.000	-0.10
Rosette water weight / area	Morph-rosette	0.019	EAST	0.42	0.215	-0.43
Leaf shape Y1	Morph-leaf	0.392	WEST	0.00	1.000	0.11
Leaf shape Y2	Morph-leaf	0.042	WEST	0.50	0.185	0.30
Leaf shape Y3	Morph-leaf	**< 0.001**	WEST	0.87	0.061	0.53
Leaf shape Y4	Morph-leaf	**< 0.001**	WEST	0.94	**0.029**	0.80
Leaf shape Y5	Morph-leaf	**< 0.001**	WEST	0.93	**0.029**	0.87
Leaf shape Y6	Morph-leaf	**< 0.001**	WEST	0.93	**0.029**	0.91
Leaf shape Y7	Morph-leaf	**< 0.001**	WEST	0.93	**0.029**	0.94
Leaf shape Y8	Morph-leaf	**< 0.001**	WEST	0.94	**0.028**	0.95
Leaf shape Y9	Morph-leaf	**< 0.001**	WEST	0.92	**0.029**	0.93

a. Boldface denotes significant *P*-values after sequential Bonferroni correction within each trait category.
b. This column denotes which subspecies has the higher trait value.
c. Empirical *P*-values, based on the proportion of polymorphic SNPs with F_{ST} higher than or equal to Q_{ST} for this trait. Boldface denotes values less than 0.05.
d. Pearson's correlation coefficient between trait value and the DFA score of each trait category. DFA is not available for physiology traits.

Table S 7: Divergence of the 'composite trait' from each trait category. This table shows the data from 19 genotypes. For DFA scores from each trait category, this table shows the subspecies effect *P*-value in univariate ANOVA, the Q_{ST}, and the empirical *P* value of Q_{ST} compared to genome-wide distribution of SNP F_{ST}.

Trait category	ANOVA *P*	Q_{ST}	*P* vs. F_{ST}
Phenology	< 0.001	0.87	0.061
Morphology – stalk	< 0.001	0.91	0.029
Morphology – rosette	< 0.001	0.97	0.028
Morphology – leaf	< 0.001	0.97	0.028

Table S 8: List of all univariate traits in Chapter 5

Univariate trait	Garden	Category	Heritability	Correlation with composite trait
ALD Winter-survived plant fruit number	ALD	Fecundity fruit	0.07	-0.77
ALD Bolted plant fruit number	ALD	Fecundity fruit	0.05	-0.80
ALD Fruited plant fruit number	ALD	Fecundity fruit	0.05	-0.97
ALD All plant fruit number	ALD	Fecundity fruit	0.07	-0.76
ALD Single fruit length	ALD	Fecundity seed	0.10	0.89
ALD Winter-survived plant fitness	ALD	Fecundity seed	0.05	0.72
ALD Bolted plant fitness	ALD	Fecundity seed	0.04	0.70
ALD Fruited plant fitness	ALD	Fecundity seed	0.04	0.82
ALD All plant fitness	ALD	Fecundity seed	0.05	0.74
ALD Insect herbivory	ALD	Herbivory	< 0.01	
ALD Bolted in summer	ALD	Phenology	0.10	0.67
ALD Plant stage June 2	ALD	Phenology	0.32	0.86
ALD Plant stage June 7	ALD	Phenology	0.22	0.92
ALD Plant stage June 13	ALD	Phenology	0.11	0.94
ALD Plant stage June 21	ALD	Phenology	0.09	0.94
ALD Plant stage June 27	ALD	Phenology	0.06	0.89
ALD Plant stage July 15	ALD	Phenology	0.04	0.79
ALD Stalk width	ALD	Stalk morphology	0.05	-0.40
ALD Stalk height	ALD	Stalk morphology	0.13	0.11
ALD Stalk height with branch	ALD	Stalk morphology	0.15	0.49
ALD Reproductive branch number	ALD	Stalk morphology	0.05	-0.19
ALD Stalk reproductive internode	ALD	Stalk morphology	0.25	0.93
ALD Winter survival	ALD	Survival	0.00	
ALD Summer survival	ALD	Survival	< 0.01	
ALD Overall survival	ALD	Survival	0.02	

JAM Winter-survived plant fruit number	JAM	Fecundity fruit	0.00	
JAM Bolted plant fruit number	JAM	Fecundity fruit	0.00	
JAM Fruited plant fruit number	JAM	Fecundity fruit	0.00	
JAM All plant fruit number	JAM	Fecundity fruit	0.02	1
JAM Single fruit length	JAM	Fecundity seed	0.05	0.87
JAM Winter-survived plant fitness	JAM	Fecundity seed	0.00	
JAM Bolted plant fitness	JAM	Fecundity seed	0.00	
JAM Fruited plant fitness	JAM	Fecundity seed	0.00	
JAM All plant fitness	JAM	Fecundity seed	0.02	0.83
JAM Insect herbivory	JAM	Herbivory	0.00	
JAM Bolted in summer	JAM	Phenology	0.04	0.77
JAM Plant stage June 19	JAM	Phenology	0.08	0.81
JAM Plant stage June 25	JAM	Phenology	0.13	0.90
JAM Plant stage July 2	JAM	Phenology	0.11	0.93
JAM Plant stage July 10	JAM	Phenology	0.04	0.92
JAM Plant stage July 16	JAM	Phenology	0.01	0.89
JAM Plant stage July 23	JAM	Phenology	0.00	
JAM Plant stage July 31	JAM	Phenology	0.00	
JAM Stalk width	JAM	Stalk morphology	0.20	0.38
JAM Stalk height	JAM	Stalk morphology	0.11	0.74
JAM Stalk height with branch	JAM	Stalk morphology	0.10	0.95
JAM Reproductive branch number	JAM	Stalk morphology	0.00	
JAM Stalk reproductive internode	JAM	Stalk morphology	0.17	0.91
JAM Winter survival	JAM	Survival	0.04	0.96
JAM Summer survival	JAM	Survival	0.00	
JAM Overall survival	JAM	Survival	0.04	0.93
GH Final flower number	GH	Fecundity fruit	0.15	1.00
GH Final fruit number	GH	Fecundity fruit	0.27	0.39
GH Single fruit length	GH	Fecundity seed	0.24	0.89
GH Fitness	GH	Fecundity seed	0.24	0.24
GH Leaf length	GH	Leaf morphology	0.44	-0.49

GH Leaf width	GH	Leaf morphology	0.61	-0.99
GH Leaf width / length	GH	Leaf morphology	0.55	-0.82
GH Flowering time	GH	Phenology	0.36	-0.81
GH Width when flowering	GH	Phenology	0.28	-0.30
GH Height when flowering	GH	Phenology	0.50	0.37
GH Leaf number when flowering	GH	Phenology	0.30	-0.25
GH Rosette number when flowerint	GH	Phenology	0.21	0.43
GH Number of flower stalk	GH	Phenology	0.17	0.66
GH Final tip bud active	GH	Phenology	0.11	-0.56
GH Final with mature fruit	GH	Phenology	0.35	0.75
GH Rosette number	GH	Rosette morphology	0.32	-0.26
GH Rosette width	GH	Rosette morphology	0.40	-0.66
GH Rosette height	GH	Rosette morphology	0.37	-0.60
GH Rosette volume	GH	Rosette morphology	0.46	-0.73
GH Rosette fresh weight	GH	Rosette morphology	0.49	-0.91
GH Rosette dry weight	GH	Rosette morphology	0.46	-0.89
GH Rosette water weight	GH	Rosette morphology	0.49	-0.91
GH Rosette water proportion	GH	Rosette morphology	0.39	0.13
GH Rosette leaf area	GH	Rosette morphology	0.54	-0.97
GH Rosette leaf packing	GH	Rosette morphology	0.40	-0.35
GH Rosette fresh weight / area	GH	Rosette morphology	0.56	0.31
GH Rosette dry weight / area	GH	Rosette morphology	0.43	0.11
GH Rosette Water weight / area	GH	Rosette morphology	0.56	0.33
GH Final rosette width	GH	Rosette morphology	0.24	-0.36
GH Stalk height	GH	Stalk morphology	0.22	0.44
GH Stalk width	GH	Stalk morphology	0.36	-0.30
GH Reproductive branch number	GH	Stalk morphology	0.22	0.27
GH Stalk height with branch	GH	Stalk morphology	0.22	0.79
GH Stalk reproductive internode	GH	Stalk morphology	0.31	0.96

Table S 9: List of all composite traits used in Chapter 5

Composite trait	Number of traits[a]	Eigenvalue[b]	Proportion variation explained[c]
JAM Survival	2	1.77	0.89
ALD Survival	2	1.12	0.56
ALL Survival	4	1.46	0.37
JAM Fecundity fruit	1	1.00	1.00
ALD Fecundity fruit	4	1.52	0.38
GH Fecundity fruit	2	1.06	0.53
ALL Fecundity fruit	7	1.22	0.17
JAM Fecundity seed	2	1.39	0.70
ALD Fecundity seed	5	1.75	0.35
GH Fecundity seed	2	0.49	0.24
ALL Fecundity seed	9	1.92	0.21
GH Leaf morphology	3	1.89	0.63
JAM Phenology	6	4.43	0.74
ALD Phenology	7	4.94	0.71
GH Phenology	8	2.08	0.26
ALL Phenology	21	6.14	0.29
GH Rosette morphology	14	4.50	0.32
JAM Stalk morphology	4	2.14	0.53
ALD Stalk morphology	5	1.19	0.24
GH Stalk morphology	5	1.65	0.33
ALL Stalk morphology	14	2.93	0.21

a. Number of univariate traits included in each composite trait
b. Eigenvalue equals the variance of each composite trait
c. The proportion of total variation among the univariate traits explained by this composite trait

Table S 10: Adaptor oligos and PCR primers in the modified Andolfatto *et. al.* (2011) protocol

Name	Full sequence[a]
B1_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTCGG
B2_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATT
B3_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATACT
B4_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGCTGT
B5_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCTCT
B6_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCTCC
B7_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTGCT
B8_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTATAGT
B9_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCAT
B10_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAATCC
B11_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCAGTT
B12_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTGAA
B13_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGCGA
B14_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACTTA
B15_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTCCT
B16_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAATA
B17_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGAAT
B18_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGTA
B19_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCCTCG
B20_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGAC
B21_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGACT
B22_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTAGC
B23_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGACTA
B24_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGAGA
B25_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATCTA
B26_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAAGCT

B27_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTCTCA
B28_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTCCT
B29_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCAA
B30_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTTA
B31_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGGAA
B32_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAATGA
B33_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGGTAA
B34_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTCAA
B35_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTGCA
B36_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCATA
B37_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTCACT
B38_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTCGA
B39_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTCGA
B40_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATATAC
B41_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTAGTA
B42_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGCTT
B43_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGTCT
B44_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAATAT
B45_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCATGG
B46_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGGAA
B47_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCGAC
B48_FC2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGCGA
B1_FC1	p-TACCGAGCAGATCGGAAGAGCACACGTCT
B2_FC1	p-TAAATCAGAGATCGGAAGAGCACACGTCT
B3_FC1	p-TAAGTATTAGATCGGAAGAGCACACGTCT
B4_FC1	p-TAACAGCGAGATCGGAAGAGCACACGTCT
B5_FC1	p-TAAGAGACAGATCGGAAGAGCACACGTCT
B6_FC1	p-TAGGAGATAGATCGGAAGAGCACACGTCT
B7_FC1	p-TAAGCAGAAGATCGGAAGAGCACACGTCT
B8_FC1	p-TAACTATAAGATCGGAAGAGCACACGTCT

B9_FC1	p-TAATGCTGAGATCGGAAGAGCACACGTCT
B10_FC1	p-TAGGATTAAGATCGGAAGAGCACACGTCT
B11_FC1	p-TAAACTGCAGATCGGAAGAGCACACGTCT
B12_FC1	p-TATTCAGGAGATCGGAAGAGCACACGTCT
B13_FC1	p-TATCGCTTAGATCGGAAGAGCACACGTCT
B14_FC1	p-TATAAGTTAGATCGGAAGAGCACACGTCT
B15_FC1	p-TAAGGAACAGATCGGAAGAGCACACGTCT
B16_FC1	p-TATATTAGAGATCGGAAGAGCACACGTCT
B17_FC1	p-TAATTCGGAGATCGGAAGAGCACACGTCT
B18_FC1	p-TATACAAGAGATCGGAAGAGCACACGTCT
B19_FC1	p-TACGAGGGAGATCGGAAGAGCACACGTCT
B20_FC1	p-TAGTCAGTAGATCGGAAGAGCACACGTCT
B21_FC1	p-TAAGTCACAGATCGGAAGAGCACACGTCT
B22_FC1	p-TAGCTAGTAGATCGGAAGAGCACACGTCT
B23_FC1	p-TATAGTCGAGATCGGAAGAGCACACGTCT
B24_FC1	p-TATCTCACAGATCGGAAGAGCACACGTCT
B25_FC1	p-TATAGATTAGATCGGAAGAGCACACGTCT
B26_FC1	p-TAAGCTTGAGATCGGAAGAGCACACGTCT
B27_FC1	p-TATGAGAGAGATCGGAAGAGCACACGTCT
B28_FC1	p-TAAGGAGTAGATCGGAAGAGCACACGTCT
B29_FC1	p-TATTGATCAGATCGGAAGAGCACACGTCT
B30_FC1	p-TATAAGCCAGATCGGAAGAGCACACGTCT
B31_FC1	p-TATTCCACAGATCGGAAGAGCACACGTCT
B32_FC1	p-TATCATTTAGATCGGAAGAGCACACGTCT
B33_FC1	p-TATTACCCAGATCGGAAGAGCACACGTCT
B34_FC1	p-TATTGACGAGATCGGAAGAGCACACGTCT
B35_FC1	p-TATGCAACAGATCGGAAGAGCACACGTCT
B36_FC1	p-TATATGGAAGATCGGAAGAGCACACGTCT
B37_FC1	p-TAAGTGAGAGATCGGAAGAGCACACGTCT
B38_FC1	p-TATCGAGTAGATCGGAAGAGCACACGTCT

B39_FC1	p-TATCGAACAGATCGGAAGAGCACACGTCT
B40_FC1	p-TAGTATATAGATCGGAAGAGCACACGTCT
B41_FC1	p-TATACTAAAGATCGGAAGAGCACACGTCT
B42_FC1	p-TAAAGCGAAGATCGGAAGAGCACACGTCT
B43_FC1	p-TAAGACGAAGATCGGAAGAGCACACGTCT
B44_FC1	p-TAATATTGAGATCGGAAGAGCACACGTCT
B45_FC1	p-TACCATGAAGATCGGAAGAGCACACGTCT
B46_FC1	p-TATTCCTGAGATCGGAAGAGCACACGTCT
B47_FC1	p-TAGTCGATAGATCGGAAGAGCACACGTCT
B48_FC1	p-TATCGCAAAGATCGGAAGAGCACACGTCT
FC1_PCR_Index1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTC*T
FC1_PCR_Index2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTC*T
FC1_PCR_Index3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTC*T
FC1_PCR_Index4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTC*T
FC2_PCR_primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T

a. For FC1 oligos, the p- stands for 5′ phosphate modification. For PCR primers the asterisk before the last base stands for phosphorothioate modification

Table S 11: List of all QTL, allelic direction, and proportional genetic variation explained in Chapter 5

CH	Range (cM)	Higher parental allele	Trait name	Trait Category	Proportion (%)
6	85-95	Ruby	ALD Winter-survived plant fruit number	Fecundity fruit	9.24
5	110-120	Ruby	ALD Fruited plant fruit number	Fecundity fruit	18.95
5	150-160	Parker	ALD Fruited plant fruit number	Fecundity fruit	10.53
5	110-120	Ruby	ALD Single fruit length	Fecundity seed	11.49
6	85-95	Ruby	ALD Winter-survived plant fitness	Fecundity seed	10.34
3	20-45	Parker	ALD Fruited plant fitness	Fecundity seed	12.31
5	110-120	Ruby	ALD Fruited plant fitness	Fecundity seed	19.61
3	20-45	Parker	ALD Plant stage June 2	Phenology	9.48
7	40-70	Parker	ALD Plant stage June 2	Phenology	45.17
7	40-70	Parker	ALD Plant stage June 7	Phenology	39.18
7	40-70	Parker	ALD Plant stage June 13	Phenology	18.12
5	130-145	Parker	ALD Plant stage June 27	Phenology	12.21
1	0-15	Parker	ALD Plant stage July 15	Phenology	11.63
6	15-30	Ruby	ALD Insect herbivory	Herbivory	10.38
7	40-70	Ruby	ALD Stalk height	Stalk morphology	12.01
3	0-10	Ruby	ALD Reproductive branch number	Stalk morphology	10.45
5	110-120	Ruby	ALD Reproductive branch number	Stalk morphology	16.42
2	60-85	Parker	ALD Stalk reproductive internode	Stalk morphology	11.73
4	45-55	Parker	JAM Winter survival	Survival	9.37
6	15-30	Parker	JAM Winter survival	Survival	9.76
4	45-55	Parker	JAM Overall survival	Survival	9.58
6	40-60	Parker	JAM Bolted in summer	Phenology	17.63
3	20-45	Parker	JAM All plant fruit number	Fecundity fruit	16.1
3	20-45	Parker	JAM All plant fitness	Fecundity seed	12.84

2	45-65	Ruby	JAM Plant stage June 19	Phenology	10.29
6	40-60	Parker	JAM Plant stage June 19	Phenology	15.91
7	40-70	Parker	JAM Plant stage June 25	Phenology	11.99
7	40-70	Parker	JAM Plant stage July 2	Phenology	11.92
3	20-45	Parker	JAM Plant stage July 10	Phenology	11
6	40-60	Parker	JAM Plant stage July 10	Phenology	13.18
5	110-120	Ruby	GH Rosette number	Rosette morphology	19.23
2	95-115	Ruby	GH Rosette width	Rosette morphology	12.04
3	55-65	Ruby	GH Rosette width	Rosette morphology	10.46
7	40-70	Parker	GH Rosette height	Rosette morphology	13.18
3	55-65	Ruby	GH Rosette volume	Rosette morphology	11.35
4	65-75	Ruby	GH Rosette water proportion	Rosette morphology	13.26
3	10-20	Ruby	GH Rosette leaf area	Rosette morphology	11.37
6	55-70	Ruby	GH Rosette leaf area	Rosette morphology	13.33
2	80-105	Parker	GH Rosette leaf packing	Rosette morphology	10.89
3	55-65	Parker	GH Rosette leaf packing	Rosette morphology	12.49
6	55-70	Ruby	GH Rosette leaf packing	Rosette morphology	17.78
6	55-70	Parker	GH Rosette fresh weight / area	Rosette morphology	15.99
6	85-95	Parker	GH Rosette fresh weight / area	Rosette morphology	13.06
2	25-45	Parker	GH Rosette dry weight / area	Rosette morphology	8.95
6	55-70	Parker	GH Rosette Water weight / area	Rosette morphology	21.59
6	85-95	Parker	GH Rosette Water weight / area	Rosette morphology	19.65
3	55-65	Ruby	GH Leaf length	Leaf morphology	13.13
6	55-70	Ruby	GH Leaf width	Leaf morphology	26.9
7	90-110	Ruby	GH Leaf width	Leaf morphology	8.94
2	60-85	Parker	GH Leaf width / length	Leaf morphology	6.78
6	55-70	Ruby	GH Leaf width / length	Leaf morphology	28.22
7	90-110	Ruby	GH Leaf width / length	Leaf morphology	7.82
1	40-50	Ruby	GH Flowering time	Phenology	8.78
7	40-70	Ruby	GH Flowering time	Phenology	34.81
7	90-110	Ruby	GH Flowering time	Phenology	7.09

1	40-50	Parker	GH Height when flowering	Phenology	9.94
5	175-185	Ruby	GH Height when flowering	Phenology	8.47
7	40-70	Parker	GH Height when flowering	Phenology	26.26
3	20-45	Ruby	GH Leaf number when flowering	Phenology	8.03
3	80-90	Parker	GH Leaf number when flowering	Phenology	14.5
6	85-95	Ruby	GH Leaf number when flowering	Phenology	8.43
1	40-50	Parker	GH Number of flower stalk	Phenology	9.33
7	40-70	Parker	GH_ln_Flower_Stalk	Phenology	13.47
6	30-50	Ruby	GH Final rosette width	Rosette morphology	10.58
7	40-70	Parker	GH Stalk height	Stalk morphology	16.91
2	125-135	Parker	GH Stalk width	Stalk morphology	12.95
2	45-65	Ruby	GH Stalk width	Stalk morphology	8.64
2	95-115	Ruby	GH Stalk width	Stalk morphology	11.41
3	100-110	Ruby	GH Stalk width	Stalk morphology	8.46
3	10-20	Ruby	GH Stalk width	Stalk morphology	15.7
1	40-50	Parker	GH Final flower number	Fecundity fruit	10.64
5	35-40	Ruby	GH Final flower number	Fecundity fruit	12.59
7	40-70	Parker	GH Final fruit number	Fecundity fruit	24.07
7	40-70	Parker	GH Single fruit length	Fecundity seed	17.75
2	25-45	Ruby	GH Fitness	Fecundity seed	10.51
7	40-70	Parker	GH Fitness	Fecundity seed	19.16
7	40-70	Parker	GH Reproductive branch number	Stalk morphology	27.62
5	110-120	Parker	GH Stalk height with branch	Stalk morphology	11.25
7	40-70	Parker	GH Stalk height with branch	Stalk morphology	33.52
2	0-15	Ruby	GH Stalk reproductive internode	Stalk morphology	12.34
4	65-75	Parker	GH Stalk reproductive internode	Stalk morphology	8.71
7	40-70	Parker	GH Stalk reproductive internode	Stalk morphology	16.94
1	55-70	Ruby	GH Final tip bud active	Phenology	12.44
7	40-70	Ruby	GH Final tip bud active	Phenology	9.28
1	40-50	Parker	GH Final with mature fruit	Phenology	8.8
2	60-85	Parker	GH Final with mature fruit	Phenology	10.42

4	45-55	Parker	JAM Survival	Composite	11.63
3	20-45	Parker	JAM Fecundity fruit	Composite	16.11
5	105-115	Parker	ALD Fecundity fruit	Composite	17.16
5	150-160	Ruby	ALD Fecundity fruit	Composite	14.12
1	40-55	Parker	GH Fecundity fruit	Composite	12.37
5	105-115	Parker	ALL Fecundity fruit	Composite	13.82
3	20-45	Parker	JAM Fecundity seed	Composite	12.86
5	105-115	Ruby	ALD Fecundity seed	Composite	9.01
2	110-130	Parker	GH Fecundity seed	Composite	11.84
7	30-45	Parker	GH Fecundity seed	Composite	11.45
3	20-45	Parker	ALL Fecundity seed	Composite	25.04
6	60-70	Parker	GH Leaf morphology	Composite	28.78
7	100-120	Parker	GH Leaf morphology	Composite	9.78
3	20-45	Parker	JAM Phenology	Composite	9.48
6	40-60	Parker	JAM Phenology	Composite	12.98
7	50-70	Parker	JAM Phenology	Composite	11.22
7	50-70	Parker	ALD Phenology	Composite	22.73
1	40-55	Parker	GH Phenology	Composite	9.43
2	60-90	Parker	GH phenology	Composite	6.91
3	20-45	Parker	GH Phenology	Composite	9.78
7	100-120	Parker	GH phenology	Composite	6.61
7	50-70	Parker	GH Phenology	Composite	27.07
3	20-45	Parker	ALL Phenology	Composite	9.6
7	50-70	Parker	ALL Phenology	Composite	25.08
6	60-70	Parker	GH Rosette morphology	Composite	17.1
7	30-45	Ruby	GH Rosette morphology	Composite	8.39
2	60-90	Ruby	ALD Stalk morphology	Composite	10.84
2	0-15	Ruby	GH Stalk morphology	Composite	11.42
2	60-90	Parker	GH Stalk morphology	Composite	10.17
7	50-70	Parker	GH Stalk morphology	Composite	23.21
3	110-120	Parker	ALL Stalk morphology	Composite	16.63

3	20-45	Parker	ALL Stalk morphology	Composite	16.6

Appendix B. Supplementary figures

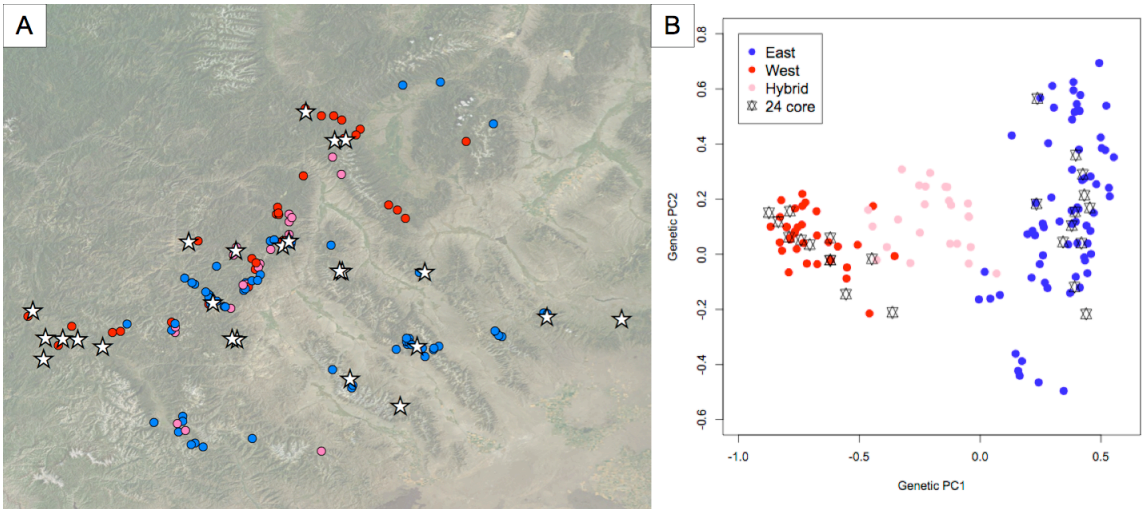


Figure S 1: The 24 genotypes represent most of the (A) geographical and (B) genetic variation among all *Boechera stricta* accessions in my study area (Latitude: 43.50 to 46.00 N, Longitude: 111.00 to 116.00 W). In both panels, white stars represent 24 core genotypes used in this study, blue dots represent EASTERN genotypes, red dots represent WESTERN genotypes, and pink dots represent hybrids. All data are obtained from Lee and Mitchell-Olds (2011). Genetic groups (EAST/WEST/hybrid) were assigned by STRUCTURE.

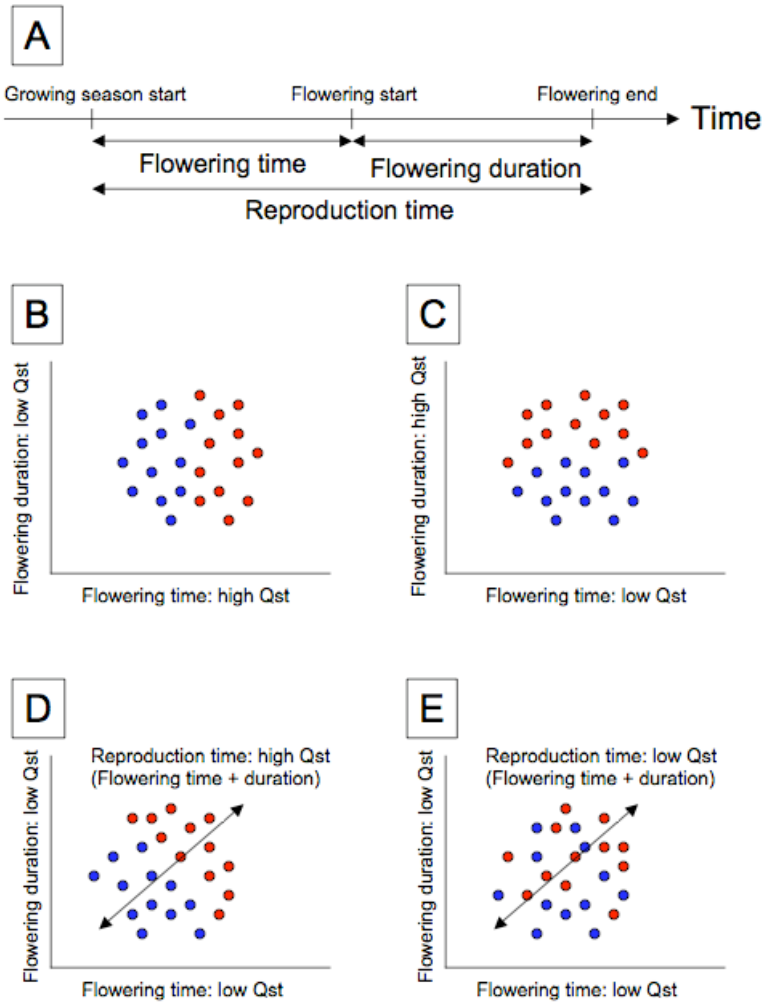


Figure S 2: Example of multivariate trait divergence in phenology, assuming natural selection favors the divergence in 'total reproduction time' between the red and blue population. Each point represents one genotype. (A) This trait, although not directly measured, is a linear combination of flowering time and duration. The two populations may diverge in either flowering time (B), duration (C), or both (D). In examples (B) and (C), the traits under divergent selection could be identified via their high Q_{ST}. In case (D), however, no univariate trait has Qst higher than the significance threshold, and the divergent selection on phenology as a whole might not be identified. Nevertheless, these three examples all have the same amount of divergence in total reproduction time. In case (D), the composite trait under strongest divergent selection (and therefore its Q_{ST}) could be identified via discriminant function analysis or MANOVA between the two populations. Notice that this method only involves a rotation of axis and does not produce an upward bias in multivariate Qst. Finally, in (E) if none of the univariate or multivariate traits has diverged, the multivariate Qst also will be low.

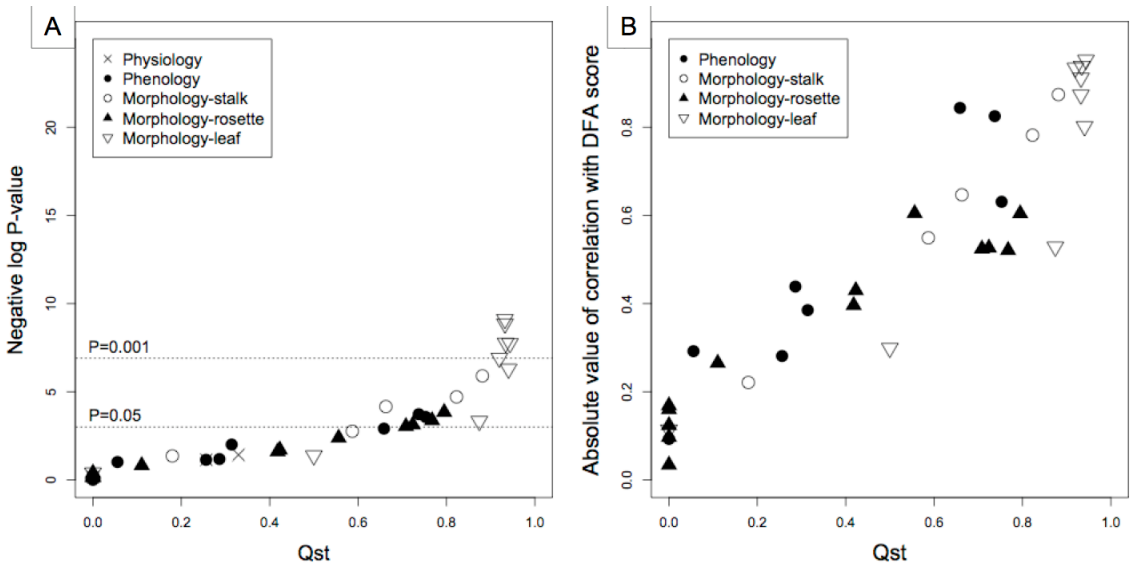


Figure S 3: Relationship between trait Q_{ST} and (A) P value of subspecies effect in ANOVA (B) absolute value of correlation with discriminant function analysis (DFA) score from each trait category. Shown are data from 19 genotypes. All axes and scales are equivalent to Figure 7.

References

Abramoff MD, Magalhaes PJ, Ram SJ (2004) Image processing with ImageJ. *Biophotonics International* **11**, 36-42.

Anderson JT, Lee C-R, Mitchell-Olds T (2011a) Life history QTLs and natural selection on flowering time in *Boechera stricta*, a perennial relative of *Arabidopsis*. *Evolution* **65**, 771-787.

Anderson JT, Lee C-R, Mitchell-Olds T (2013) Strong selection genome-wide enhances fitness tradeoffs across environments and episodes of selection. *Evolution* **In press**.

Anderson JT, Lee C-R, Rushworth C, Colautti RI, Mitchell-Olds T (2012) Genetic tradeoffs and conditional neutrality contribute to local adaptation. *Molecular Ecology* **22**, 699-708.

Anderson JT, Willis JH, Mitchell-Olds T (2011b) Evolutionary genetics of plant adaptation. *Trends Genet* **27**, 258-266.

Andolfatto P, Davison D, Erezyilmaz D, *et al.* (2011a) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research* **21**, 610-617.

Andolfatto P, Wong KM, Bachtrog D (2011b) Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species. *Genome Biol Evol* **3**, 114-128.

Arthur AL, Weeks AR, SgrÒ CM (2008) Investigating latitudinal clines for life history and stress resistance traits in Drosophila simulans from eastern Australia. *Journal of Evolutionary Biology* **21**, 1470-1479.

Atwell S, Huang YS, Vilhjalmsson BJ, *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627-631.

Bachtrog D, Charlesworth B (2002) Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* **416**, 323-326.

Balkenhol N, Waits LP, Dezzani RJ (2009) Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography* **32**, 818-830.

Barrett RD, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet* **12**, 767-780.

Basten CJ, Weir BS, Zeng Z-B (2005) *QTL Cartographer, Version 1.17*, Department of Statistics, North Carolina State University, Raleigh, N.C.

Berardini TZ, Mundodi S, Reiser L, *et al.* (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol.*, pp.104.040071.

Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat Rev Genet* **11**, 867-879.

Bernier J, Atlin GN, Serraj R, Kumar A, Spaner D (2008) Breeding upland rice for drought resistance. *Journal of the Science of Food and Agriculture* **88**, 927-939.

Blows MW (2007) A tale of two matrices: multivariate approaches in evolutionary biology. *J Evol Biol* **20**, 1-8.

Bolker BM, Brooks ME, Clark CJ, *et al.* (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* **24**, 127-135.

Boutin-Ganache I, Raposo M, Raymond M, Deschepper C (2001) M13-Tailed Primers Improve the Readability and Usability of Microsatellite Analyses Performed with Two Different Allele-Sizing Methods. *BioTechniques* **31**, 24-27.

Brunelle A, Whitlock C (2003) Postglacial fire, vegetation, and climate history in the Clearwater Range, Northern Idaho, USA. *Quaternary Research* **60**, 307-318.

Bryc K, Auton A, Nelson MR, *et al.* (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences* **107**, 786-791.

Cao J, Schneeberger K, Ossowski S, *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* **43**, 956-963.

Chenoweth SF, Blows M, Wolf J (2008) Q_{ST} meets the G matrix: the dimensionality of adaptive divergence in multiple correlated quantitative traits. *Evolution* **62**, 1437-1449.

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971.

Colautti RI, Lee C-R, Mitchell-Olds T (2012) Origin, fate, and architecture of ecologically relevant genetic variation. *Current Opinion in Plant Biology* **15**, 199–204.

Colautti RI, Maron JL, Barrett SCH (2009) Common garden comparisons of native and introduced plant populations: latitudinal clines can obscure evolutionary inferences. *Evolutionary Applications* **2**, 187-199.

Colbourne JK, Pfrender ME, Gilbert D, *et al.* (2011) The ecoresponsive genome of Daphnia pulex. *Science* **331**, 555-561.

Comeron JM, Kreitman M, Aguade M (1999) Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. *Genetics* **151**, 239-249.

Condon AG, Richards RA, Rebetzke GJ, Farquhar GD (2004) Breeding for high water-use efficiency. *Journal of Experimental Botany* **55**, 2447-2460.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics* **185**, 1411-1423.

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**, 485.

Coyne JA, Orr HA (2004) *Speciation* Sinauer Associates, Sunderland, MA.

Cushman SA, McKelvey KS, Hayden J, Schwartz MK (2006) Gene flow in complex landscapes: Testing multiple hypotheses with causal modeling. *American Naturalist* **168**, 486-499.

Dyer RJ, Nason JD, Garrick RC (2010) Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. *Molecular Ecology* **19**, 3746-3759.

Earley EJ, Ingland B, Winkler J, Tonsor SJ (2009) Inflorescences contribute more than rosettes to lifetime carbon gain in *Arabidopsis thaliana* (Brassicaceae). *American Journal of Botany* **96**, 786-792.

Eckert AJ, Bower AD, Gonzalez-Martinez SC, *et al.* (2010a) Back to nature: ecological genomics of loblolly pine (Pinus taeda, Pinaceae). *Molecular Ecology* **19**, 3789-3805.

Eckert AJ, van Heerwaarden J, Wegrzyn JL, *et al.* (2010b) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). *Genetics* **185**, 969-982.

Edelaar PIM, Björklund M (2011) If F_{ST} does not measure neutral genetic differentiation, then comparing it with Q_{ST} is misleading. Or is it? *Molecular Ecology* **20**, 1805-1812.

Edelaar PIM, Burraco P, Gomez-Mestre I (2011) Comparisons between Q_{ST} and F_{ST} - how wrong have we been? *Molecular Ecology* **20**, 4830-4839.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1-15.

Engelhardt BE, Stephens M (2010) Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. *PLoS Genet* **6**, e1001117.

Etterson JR (2004) Evolutionary potential of Chamaecrista fasciculata in relation to climate change. II. Genetic architecture of three populations reciprocally planted along an environmental gradient in the great plains. *Evolution* **58**, 1459-1471.

Farquhar GD, Ehleringer JR, Hubick KT (1989) Carbon isotope discrimination and photosynthesis. *Annual Review of Plant Biology* **40**, 503-537.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends Genet* **28**, 342-350.

Filiault DL, Maloof JN (2012) A genome-wide association study identifies variants underlying the Arabidopsis thaliana shade avoidance response. *PLoS Genet* **8**, e1002589.

Flowers JM, Molina J, Rubinstein S, *et al.* (2012) Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Mol Biol Evol*.

Fournier-Level A, Korte A, Cooper MD, *et al.* (2011) A map of local adaptation in Arabidopsis thaliana. *Science* **334**, 86-89.

Freedman AH, Thomassen HA, Buermann W, Smith TB (2010) Genomic signals of diversification along ecological gradients in a tropical lizard. *Molecular Ecology* **19**, 3773-3788.

Funk DJ, Egan SP, Nosil P (2011) Isolation by adaptation in Neochlamisus leaf beetles: host-related selection promotes neutral genomic divergence. *Mol Ecol* **20**, 4671-4682.

Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* **176**, 1635-1651.

Gaut B (2012) Arabidopsis thaliana as a model for the genetics of local adaptation. *Nat Genet* **44**, 115-116.

Gillespie JH, Langley CH (1974) A genera model to account for enzyme variation in natural populations. *Genetics* **76**, 837-848.

Goudet J (2001) FSTAT: A program to estimate and test gene diversities and fixation indices (version 2.9.3). *www2.unil.ch/popgen/softwares/fstat.htm*.

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes* **5**, 184-186.

Goudet J, Buchi L (2006) The effects of dominance, regular inbreeding and sampling design on Q_{ST}, an estimator of population differentiation for quantitative traits. *Genetics* **172**, 1337-1347.

Guillot G, Mortier F, Estoup A (2005) GENELAND: a computer package for landscape genetics. *Molecular Ecology Notes* **5**, 712-715.

Hall MC, Lowry DB, Willis JH (2010) Is local adaptation in Mimulus guttatus caused by trade-offs at individual loci? *Molecular Ecology* **19**, 2739-2753.

Hancock AM, Brachi B, Faure N, *et al.* (2011) Adaptation to climate across the Arabidopsis thaliana genome. *Science* **334**, 83-86.

Hancock AM, Witonsky DB, Ehler E, *et al.* (2010) Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences* **107**, 8924-8930.

Hartl DL, Clark AG (2007) *Principles of population genetics*, 4th edn. Sinauer Associates, Sunderland, MA.

Hedrick PW (1986) Genetic polymorphism in heterogeneous environments: A decade later. *Annual Review of Ecology and Systematics* **17**, 535-566

Hedrick PW (2006) Genetic polymorphism in heterogeneous environments: The age of genomics. *Annual Review of Ecology Evolution and Systematics* **37**, 67-93.

Hershberg R, Petrov DA (2008) Selection on codon bias. *Annu Rev Genet* **42**, 287-299.

Hickerson MJ, Carstens BC, Cavender-Bares J, *et al.* (2010) Phylogeography's past, present, and future: 10 years after. *Molecular Phylogenetics and Evolution* **54**, 291-301.

Hijmans RJ, Cameron SE, Parra JL, Jones PG (2005) Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965-1978.

Hoffmann WA, Franco AC, Moreira MZ, Haridasan M (2005) Specific leaf area explains differences in leaf traits between congeneric savanna and forest trees. *Functional Ecology* **19**, 932-940.

Hopkins R, Schmitt J, Stinchcombe JR (2008) A latitudinal cline and response to vernalization in leaf angle and morphology in *Arabidopsis thaliana* (Brassicaceae). *New Phytologist* **179**, 155-164.

Hostetler SW, Clark PU (1997) Climatic controls of Western U.S. Glaciers at the last glacial maximum. *Quaternary Science Reviews* **16**, 505-511.

Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. *Nat Rev Genet* **11**, 855-866.

Huang X, Wei X, Sang T, *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* **42**, 961-967.

Hübner S, Höffken M, Oren E, *et al.* (2009) Strong correlation of wild barley (*Hordeum spontaneum*) population structure with temperature and precipitation variation. *Molecular Ecology* **18**, 1523-1536.

Jombart T, Pontier D, Dufour AB (2009) Genetic markers in the playground of multivariate analysis. *Heredity* **102**, 330-341.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters* **7**, 1225-1241.

Keller SR, Sowell DR, Neiman M, Wolfe LM, Taylor DR (2009) Adaptation and colonization history affect the evolution of clines in two introduced species. *New Phytologist* **183**, 678-690.

Keller SR, Taylor DR (2008) History, chance and adaptation during biological invasion: separating stochastic phenotypic evolution from response to selection. *Ecology Letters* **11**, 852-866.

Korves TM, Schmid KJ, Caicedo AL, *et al.* (2007) Fitness effects associated with the major flowering time gene *FRIGIDA* in *Arabidopsis thaliana* in the field. *Am Nat* **169**, E141-E157.

Kozak KH, Graham CH, Wiens JJ (2008) Integrating GIS-based environmental data into evolutionary biology. *Trends in Ecology & Evolution* **23**, 141-148.

Kozak KH, Wiens JJ (2006) Does niche conservatism promote speciation? A case study in North American salamanders. *Evolution* **60**, 2604-2621.

Kremer A, Zanetto A, Ducousso A (1997) Multilocus and multitrait measures of differentiation for gene markers and phenotypic traits. *Genetics* **145**, 1229-1241.

Kunstner A, Nabholz B, Ellegren H (2011) Significant selective constraint at 4-fold degenerate sites in the avian genome and its consequence for detection of positive selection. *Genome Biol Evol* **3**, 1381-1389.

Lande R (1979) Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution* **33**, 402-416.

Lee C-R, Mitchell-Olds T (2011) Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. *Molecular Ecology* **20**, 4631-4642.

Lee C-R, Mitchell-Olds T (2013) Complex trait divergence contributes to environmental niche differentiation in ecological speciation of *Boechera stricta*. *Molecular Ecology* **22**, 2204-2217.

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

Leinonen PH, Sandring S, Quilot B, *et al.* (2009) Local adaptation in European populations of *Arabidopsis lyrata* (Brassicaceae). *American Journal of Botany* **96**, 1129-1137.

Lercher MJ, Hurst LD (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**, 337-340.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li H, Handsaker B, Wysoker A, *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Lowry DB, Rockwood RC, Willis JH (2008) Ecological reproductive isolation of coast and inland races of *Mimulus guttatus*. *Evolution* **62**, 2196-2214.

MacColl AD (2011) The ecological causes of evolution. *Trends Ecol Evol* **26**, 514-522.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010) Common factors drive adaptive genetic variation at different spatial scales in Arabis alpina. *Molecular Ecology* **19**, 3824-3835.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends In Ecology & Evolution* **18**, 189-197.

Manel S, Segelbacher G (2009) Perspectives and challenges in landscape genetics. *Molecular Ecology* **18**, 1821-1822.

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nat Genet* **36**, 512-517.

Martin G, Chapuis E, Goudet J (2008) Multivariate Q_{st}-F_{st} comparisons: a neutrality test for the evolution of the G matrix in structured populations. *Genetics* **180**, 2135-2149.

McCormack JE, Zellmer AJ, Knowles LL (2010) Does niche divergence accompany allopatric divergence in Aphelocoma Jays as predicted under ecological speciation?: insights from tests with niche models. *Evolution* **64**, 1231-1244.

McKay JK, Bishop JG, Lin JZ, *et al.* (2001) Local adaptation across a climatic gradient despite small effective population size in the rare sapphire rockcress. *Proceedings of the Royal Society of London - Series B: Biological Sciences* **268**, 1715-1721.

Mckay JK, Richards JH, Mitchell-Olds T (2003) Genetics of drought adaptation in *Arabidopsis thaliana*: I. Pleiotropy contributes to genetic correlations among ecological traits. *Molecular Ecology* **12**, 1137-1151.

Messmer R, Fracheboud Y, Banziger M, Stamp P, Ribaut J-M (2011) Drought stress and tropical maize: QTLs for leaf greenness, plant senescence, and root capacitance. *Field Crops Research* **124**, 93-103.

Mitchell-Olds T (2001) *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends Ecol Evol* **16**, 693-700.

Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat Rev Genet* **8**, 845-856.

Montague JL, Barrett SCH, Eckert CG (2008) Re-establishment of clinal variation in flowering time among introduced populations of purple loosestrife (*Lythrum salicaria*, Lythraceae). *Journal of Evolutionary Biology* **21**, 234-245.

Nakazato T, Bogonovich M, Moyle LC (2008) Environmental factors predict adaptive phenotypic differentiation within and between two wild Andean tomatoes. *Evolution* **62**, 774-792.

Nautiyal PC, Rachaputi NR, Joshi YC (2002) Moisture-deficit-induced changes in leaf-water content, leaf carbon exchange rate and biomass production in groundnut cultivars differing in specific leaf area. *Field Crops Research* **74**, 67-79.

Nicotra AB, Leigh A, Boyce CK, *et al.* (2011) The evolution and functional significance of leaf shape in the angiosperms. *Functional Plant Biology* **38**, 535-552.

Nosil P, Egan SP, Funk DJ (2008) Heterogeneous genomic differentiation between walking-stick ecotypes: "isolation by adaptation" and multiple roles for divergent selection. *Evolution* **62**, 316-336.

Nosil P, Vines TH, Funk DJ (2005) Perspective: Reproductive Isolation Caused by Natural Selection against Immigrants from Divergent Habitats. *Evolution* **59**, 705-719.

Novembre J, Johnson T, Bryc K, *et al.* (2008) Genes mirror geography within Europe. *Nature* **456**, 98-101.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* **40**, 646-649.

Orr MR, Smith TB (1998) Ecology and speciation. *Trends in Ecology & Evolution* **13**, 502-506.

Ovaskainen O, Karhunen M, Zheng C, Arias JMC, Merilä J (2011) A new method to uncover signatures of divergent and stabilizing selection in quantitative traits. *Genetics* **189**, 621-632.

Oyama R, Clauss M, Formanová N, *et al.* (2008) The shrunken genome of *Arabidopsis thaliana*. *Plant Systematics and Evolution* **273**, 257-271.

Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927-931.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* **2**, e190.

Peakall ROD, Smouse PE (2006) Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol Ecol Notes* **6**, 288-295.

Pease K, Freedman A, Pollinger J, *et al.* (2009) Landscape genetics of California mule deer (*Odocoileus hemionus*): the roles of ecological and historical factors in generating differentiation. *Molecular Ecology* **18**, 1848-1862.

Pena-Castillo L, Hughes TR (2007) Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176**, 7-14.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**, 231-259.

Platt A, Horton M, Huang YS, *et al.* (2010) The Scale of Population Structure in Arabidopsis thaliana. *PLoS Genet* **6**, e1000843.

Prasad KVSK, Song BH, Olson-Manning C, *et al.* (2012) A gain-of-function polymorphism controlling complex traits and fitness in nature. *Science* **337**, 1081-1084.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* **20**, R208-R215.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**, 1-17.

Rollins RC (1993) *The Cruciferae of Continental North America* Stanford University Press, Stanford, CA.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters* **8**, 336-352.

Rushworth CA, Song B-H, Lee C-R, Mitchell-Olds T (2011) Boechera, a model system for ecological genomics. *Molecular Ecology* **20**, 4843-4857.

Samis KE, Heath KD, Stinchcombe JR (2008) Discordant Longitudinal Clines in Flowering Time and Phytochrome C in Arabidopsis thaliana. *Evolution* **62**, 2971-2983.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology & Evolution* **16**, 372-380.

Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences* **106**, 9955-9962.

Schranz ME, Dobes C, Koch MA, Mitchell-Olds T (2005) Sexual reproduction, hybridization, apomixis, and polyploidization in the genus Boechera (Brassicaceae). *American Journal of Botany* **92**, 1797-1810.

Schranz ME, Windsor AJ, Song B-h, Lawton-Rauh A, Mitchell-Olds T (2007) Comparative genetic mapping in *Boechera stricta*, a close relative of *Arabidopsis*. *Plant Physiol.* **144**, 286-298.

Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and post-glacial colonization of Europe. *Molecular Ecology* **9**, 2109-2118.

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

Sobel JM, Chen GF, Watt LR, Schemske DW (2010) The biology of speciation. *Evolution* **64**, 295-315.

Song B-H, Windsor AJ, Schmid KJ, *et al.* (2009) Multilocus patterns of nucleotide diversity, population structure and linkage disequilibrium in *Boechera stricta*, a wild relative of *Arabidopsis*. *Genetics* **181**, 1021-1033.

Song BH, Clauss MJ, Pepper A, Mitchell-Olds T (2006) Geographic patterns of microsatellite variation in *Boechera stricta*, a close relative of *Arabidopsis*. *Molecular Ecology* **15**, 357-369.

Sork VL, Davis FW, Westfall R, *et al.* (2010) Gene movement and genetic association with regional climate gradients in California valley oak (Quercus lobata Née) in the face of climate change. *Molecular Ecology* **19**, 3806-3823.

Stajich JE, Block D, Boulez K, *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611-1618.

Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. *Mol Biol Evol* **22**, 63-73.

Stinchcombe JR, Weinig C, Ungerer M, *et al.* (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. *Proceedings of the National Academy of Sciences* **101**, 4712-4717.

Storfer A, Murphy MA, Evans JS, *et al.* (2007) Putting the 'landscape' in landscape genetics. *Heredity* **98**, 128-142.

Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP (2010) Landscape genetics: where are we now? *Molecular Ecology* **19**, 3496-3514.

Templeton AR (2008) The reality and importance of founder speciation in evolution. *Bioessays* **30**, 470-479.

Thibert-Plante X, Hendry AP (2010) When can ecological speciation be detected with neutral loci? *Molecular Ecology* **19**, 2301-2314.

Thomassen HA, Cheviron ZA, Freedman AH, *et al.* (2010) Spatial modelling and landscape-level approaches for visualizing intra-specific variation. *Molecular Ecology* **19**, 3532-3548.

Tonsor SJ, Scheiner SM (2007) Plastic trait integration across a CO_2 gradient in *Arabidopsis thaliana*. *American Naturalist* **169**, E119-E140.

Turelli M, Barton NH (2004) Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and G x E interactions. *Genetics* **166**, 1053-1079.

van Heerwaarden J, Doebley J, Briggs WH, *et al.* (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proceedings of the National Academy of Sciences* **108**, 1088-1092.

Via S, Bouck AC, Skillman S (2000) Reproductive isolation between divergent races of pea aphids on two hosts. II. Selection against migrants and hybrids in the parental environments. *Evolution* **54**, 1626-1637.

Wang IJ, Summers K (2010) Genetic structure is correlated with phenotypic divergence rather than geographic isolation in the highly polymorphic strawberry poison-dart frog. *Molecular Ecology* **19**, 447-458.

Weigel D (2012) Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol* **158**, 2-22.

Whitlock MC (2008) Evolutionary inference from Q_{ST}. *Mol Ecol* **17**, 1885-1896.

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

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

Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. *PLoS Genetics* **4**, e1000212.

Yang L, Gaut BS (2011) Factors that contribute to variation in evolutionary rate among Arabidopsis genes. *Molecular Biology and Evolution* **28**, 2359-2369.

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*, msm088.

Yu J, Pressoir G, Briggs WH, *et al*. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203-208.

Zellmer AJ, Knowles LL (2009) Disentangling the effects of historic vs. contemporary landscape structure on population genetic divergence. *Molecular Ecology* **18**, 3593-3602.

Biography

Cheng-Ruei Lee was born on April 29th, 1983 in Tainan, Taiwan. He received his Bachelor of Science from National Taiwan Normal University in June 2005. The papers he has published are: "Molecular evolution and functional diversification of fatty acid desaturases after recurrent gene duplication in Drosophila" in Molecular Biology and Evolution, "Boechera, a model system for ecological genomics" in Molecular Ecology (ME), "Life history QTLs and natural selection on flowering time in Boechera stricta, a perennial relative of Arabidopsis" in Evolution, "Quantifying effects of environmental and geographical factors on patterns of genetic differentiation" in ME, "Origin, fate, and architecture of ecologically relevant genetic variation" in Current Opinion in Plant Biology, "A novel gain of function polymorphism controlling complex traits and fitness in nature" in Science, "Environmental adaptation contributes to gene polymorphism across the Arabidopsis thaliana genome" in Molecular Biology and Evolution (MBE), "Genetic tradeoffs and conditional neutrality contribute to local adaptation" in ME, "Evolution of flux control in the glucosinolate pathway in Arabidopsis thaliana" in MBE, "Complex trait divergence contributes to environmental niche differentiation in ecological speciation of Boechera stricta" in ME, "3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture" in PNAS, and "Strong selection genome-wide enhances fitness tradeoffs across environments and episodes of selection" in Evolution. Cheng-Ruei Lee received the following honors and fellowships: Hung Taiwan-Duke University Fellowship, Duke Graduate Travel Fellowship, Duke Biology One-Semester Fellowship, Sigma Xi Grant-In-Aid of Research, Duke Biology Grant-In-Aid of Research, and the National Science Foundation Doctoral Dissertation Improvement Grant.