Combining Quantitative and Population Genetics to Map Phenotype to Genotype in *Ipomoea*

by

Sonal Gupta

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in The University of Michigan
2022

Doctoral Committee:

Associate Professor Regina S. Baucom, Chair
Associate Professor Jennifer Blesh
Assistant Professor Dan Chitwood, Michigan State University
Associate Professor Stephen A. Smith
Professor Patricia J. Wittkopp

Sonal Gupta

gsonal@umich.edu

ORCID iD: 0000-0002-4419-2345

To Mumma and Papa, for the values they've instilled in me

To my siblings (Ruchika, Stuti, Muskaan, and Vaagish), for the endless love and support

To Tanmoy for the love, encouragement and advice over the years

Acknowledgements

I would like to thank Regina Baucom, Ph.D. (Gina) for being an incredible advisor. Her endless support and encouragement made me believe in myself and allowed my push my potential as a scientist. I have learned so much from Gina, from critically thinking about science to effectively communicating science. Additionally, her constant guidance has made me a better and a more confident writer. As an international student, I am ever so grateful to Gina for always understanding and checking-in during difficult times which made me truly feel emotionally safe and not lonely. I would like to thank Dan Chitwood, Ph.D. for his excellent inputs regarding leaf shape and morphometric analysis. His suggestions were integral in shaping my research and it furthered my understanding on the subject. Additionally, Dan has been a constant source of positivity and encouragement, always offering to help and making me feel great about my work and science in general. I would also like to thank Trisha Wittkopp, Ph.D. for her crucial inputs throughout my graduation. Her inputs were invaluable, especially in regard to transcriptomics, which made me critically think through the study design and also about the inferences of the results. Additionally, I am thankful to Stephen Smith, Ph.D. and Jennifer Blesh, Ph.D. As committee members they have been so supportive and offered their expertise every step of the way. Through their support and understanding, I was able to find alternative solutions when the experiments did not go as planned. In summary, I am very thankful to all my committee members, all of whom have greatly contributed to my critical and more mature thinking of science, in a way that was most constructive and valuable.

I would also like to thank Shu-Mei Chang, Ph.D. for her professional and financial support, and the collaborators Alex Harkess, Ph.D., David Rosenthal, Ph.D., and John Stinchcombe, Ph.D. for believing in me and always offering to help whenever I was stuck. Of special mention, I would like to thank James Leebens-Mack, Ph.D., for his support through my second chapter, and for his immense assistance during field work at the University of Georgia.

Baucom lab members (past and current) have been instrumental in my academic and personal growth during my time in Ann Arbor. I have learned a great deal from them: Megan Van Etten, Ph.D., Diego Alvarado Serrano, Ph.D. helped me a great deal with acclimatizing with the computing systems and with the different aspects of bioinformatics analyses; Sara Colom, Ph.D. has been with me since the beginning and has been an amazing colleague and a great friend, helping with both science and life; Nia Johnson for always helping me calm down during stressful times and for an amazing friend. I would like to thank other lab members, notably Sasha Bishob, Grace Zhang, Malia Santos, Toshiro Newsum, and Anah Sobel. Baucom lab members have been amazing companions and friends and I am so grateful for each and every one of them.

I would specially like to thank Matthaei Botanical Gardens staff member – Michael Palmer, Paul Girard and Jeremy Moghtader – for all their help through the years in maintaining the plants at the greenhouse and helping me with the field work. Additionally, I am extremely grateful to EEB administrative staff, especially Cindy Carl, Kati Ellis and Gail Kuhnlein for keeping me on schedule with respect to the administrative requirements and helping me through the way.

I am very lucky to have been a part of a super supportive and friendly cohort. I am grateful for their friendships and our times together; they were the reason I felt so ease in a foreign country even in my first year as a graduate student. Their companionship, be it studying together for prelims, blowing away some stress, or just hanging out, has been a major source of strength for me. Shout out especially to Danny, Peter, Mariah, Chatura, Tamara, Sasha, and Haixing. I am also very grateful other friends in Ann Arbor who I love deeply, and have made me feel a sense of belonging, especially Suki, Nikesh, Henry, Charles, Renata, and Arnaud. Of special note, I am thankful to friends from back home who were my constant companions in Ann Arbor and never let me feel homesick – thank you Awanti, Siddhesh, Indira, Vora and Darshini. Forever thankful for all the special Diwali parties and homecooked meals!

To all my close friends from back home (too many to name) who have listened to me talk endlessly and always reminded me of the support and love I can find in them. Lastly, I am most grateful to my family, my mom, dad, siblings (Ruchika, Stuti, Muskaan and Vaagish), and my

husband Tanmoy. They have been my pillars of strength, loving and supporting me unconditionally. I love all of you deeply!

Table of Contents

Dedication ii

Acknowledgements iii

List of Figures ix

List of Tables xi

List of Appendices xiii

Abstract xiv

Chapter 1 Introduction 1

Problem Statement 1

Understanding the complexity of mapping phenotype to genotype 1

Mapping phenotype to genotype: Quantitative and Population Genetic Approaches 4

Study System 6

Thesis overview 7

References 9

Chapter 2 Assessing the Remarkable Morphological Diversity and Transcriptomic Basis of Leaf Shape in *Ipomoea batatas* (sweetpotato) 15

Abstract 15

Introduction 16

Materials and Methods 19

Results 22

Discussion 27

Acknowledgements 32

References 33

Chapter 3 Inter-Chromosomal Linkage Disequilibrium and Linked Fitness Cost Loci Associated with Selection for Herbicide Resistance 49

Abstract 49

Introduction 50

Results 52

Discussion 61

Materials and Methods 64

Acknowledgements 71

References 72

Chapter 4 One Hundred Years of Deciphering Genetic Correlations -- Lessons Learned and Future Perspectives 89

Abstract 89

Introduction 90

How do correlations arise? 92

Estimating genetic correlations and inferences on causality using phenotypic data 93

Assessing genetic correlations at the molecular level 95

Inferences on causality using genetic data 98

Multi-faceted approach for the future 100

Conclusions 103

Glossary 103

Acknowledgements 104

References 105

Chapter 5 Discussion and Future Directions 122

What does the leaf shape diversity in sweetpotato teach us? 122

What have we learned about the mechanistic basis of polygenic glyphosate resistance? 124

What have we learned about teasing apart the mechanistic basis of trait correlations? 126

Conclusions 128

References 129

Appendices 134

List of Figures

Figure 2-1 Geographic diversity of the 68 chosen sweet potato, *Ipomoea batatas*, accessions 39

Figure 2-2 Methodology for RNA-seq data processing for differential gene expression 40

Figure 2-3 Leaf shape variation in 57 diverse, glasshouse-grown, accessions of sweet potato, *Ipomoea batatas*, highlighting exceptionally high morphological variation 41

Figure 2-4 Elliptical Fourier Descriptor (EFDs) of symmetrical shape variation in sweet potato, *Ipomoea batatas* 42

Figure 2-5 Plot of log fold change against log CPM (counts per million) with differentially expressed transcripts highlighted (red and blue dots) for leaf shape in *Ipomoea batatas* 43

Figure 2-6 Elliptical Fourier Descriptor (EFDs) of symmetrical leaf shape variation among 68 accessions of sweet potato, *Ipomoea batatas*, in the common gardens in Michigan (a) and Ohio (b), respectively 44

Figure 3-1 Overview of populations examined in this work and the genomic context of selection 82

Figure 3-2 Region of Chromosome 10 showing signs of selection 83

Figure 3-3 Gene expression variation associated with herbicide resistance, and results of a functional assay supporting the idea that resistance in *I. purpurea* is due to detoxification 84

Figure 3-4 Long-distance linkage disequilibrium and ILD among the four highly differentiated ($G_{ST} > 0.39$) regions under selection associated with glyphosate resistance 85

Figure 3-5 Loci associated with the cost of glyphosate resistance identified by the (a) whole-genome selection-scan and differential expression analysis in the (b) absence and (c) presence of herbicide 86

Figure 4-1 Schematic representation of the mechanistic basis of genetic correlations 115

Figure 4-2 Overview of analytical methods that can be integrated to identify the causal genetic mechanism underlying trait correlations 116

Figure S2-1: Green-house grown accessions selected for transcriptomic analysis for leaf shape traits 136

Figure S2-2: Correlation plot between leaf shape traits 137

Figure S2-3 Leaf shape variation captured by EFDs from MI and OH differing significantly in their order of variation explained 138

Figure S3-1 Chromosome scaffolding and renaming 140

Figure S3-2 Synteny of the *I. purpurea* genome against related Convolvulaceae species, including *I. nil*, *I. trifida*, and *I. triloba* 141

Figure S3-3 Signs of selection across conserved haplotype of multiple glycosyltransferases for each individual on Chromosome10 141

Figure S3-4 PCA of resistant and susceptible populations used 142

List of Tables

Table 2-1 Leaf shape trait values across the 57 chosen sweetpotato accessions 45

Table 2-2 Sequence statistics of the reference transcriptome obtained from EvidentialGene pipeline 45

Table 2-3 Candidate genes maintaining variation in leaf traits (circularity, AR and symPCs) identified from the set of differentially expressed transcripts (DETs) in *Ipomoea batatas* 46

Table 2-4 ANOVA table of the leaf shape traits model showing significant explanatory variables 47

Table 2-5 Broad-sense heritability values for leaf shape traits in differing environments 48

Table 3-1 Overview of the genes under selection for glyphosate resistance that are involved in the process of detoxification, environmental sensing, and stress signaling and response 87

Table 4-1 Example of studies reporting the genetic correlation estimates using family-level phenotypic data 118

Table 4-2 Benefits and drawbacks of the most commonly used family design and analyses method used for estimating genetic correlations from phenotypic data 120

Table 4-3 Example of studies reporting the genetic correlation estimates using genetic data 121

Table S2-1 Accession IDs with their source and location of origin used in this study . 139

Table S2-2 Differentially expressed transcripts associated with leaf shape traits found in this study 139

Table S2-3 Raw read counts of orthologs of homeobox domain genes within the assembled transcriptomes, for accessions chosen for circularity RNA-Seq analysis 139

Table S3-1 SNP outliers (in the top 5% BF and 5% Rho) identified via bayenv2 143

Table S3-2 Functional annotation of genes present within +/- 5KB of bayenv2 SNP outliers 143

Table S3-3 Summary of the genome-wide regions under selection via the Md-rank-P 143

Table S3-4 Functional annotations of genes under selection identified via md-rank-P approach 143

Table S3-5 Functional annotations of genes under selection identified via bayenv2 and md-rank-P approach 143

Table S3-6 Mutations and their effects present within the genes of interest 143

Table S3-7 List of differentially expressed genes between treated (herbicide sprayed) resistant vs susceptible individuals 143

Table S3-8 List of differentially expressed genes between control (non-herbicide sprayed) resistant vs susceptible individuals 143

Table S3-9 Pairwise contrast statistics for normalized above-ground biomass between the four treatment conditions 144

Table S3-10 ILD summary statistic (99th percentile value and max r2) for the five regions under selection that exhibited GST > 0.39 144

Table S3-11 Individual ILD interactions, above the 99 percentile cutoff r2 value, for SNPs within the region under selection 144

Table S3-12 Population information for each population used in the study 145

Table S3-13 RNA-Seq sample information used in the study 146

Table S3-14 Sample information used for the malathion assay 147

List of Appendices

Appendix A Supplementary Methods, Figures and Tables for Chapter 2 134

Appendix B Supplementary Figures and Tables for Chapter 3 140

Abstract

Understanding the genetic basis of adaptive traits is one of the central goals of modern plant biology. Deciphering how phenotype maps to the genotype is a complex problem that requires answering multiple questions: Is the phenotype environmentally induced, or is it genetically controlled? What gene(s) underlie the phenotype? Are the gene(s) associated with other phenotypes which might impact its evolutionary potential? These questions can be answered using a combination of quantitative and population genetic approaches. In this thesis, I apply these approaches to two complex adaptive phenotypes -- leaf shape in *Ipomoea batatas* (sweetpotato) and herbicide resistance in *Ipomoea purpurea* (common morning glory) -- to identify putative genes and mechanisms contributing to phenotype. Further, I synthesize a framework for analyzing genetic correlations that underlie most complex traits. I show that leaf shape variation is extensive and largely under genetic control in sweetpotato, and likewise, identify genes putatively associated with leaf shape variation. I also show that although simple leaf shape descriptors are not environmentally controlled when considered together these can be influenced to a great extent by the environment. Next, using genome-wide resequencing, RNA-Seq and functional tests I show that the detoxification mechanism underlies the polygenic herbicide resistance in the common morning glory. Using linkage analysis, I identify a role of intrerchromosomal linkage disequilibrium in maintaining the resistance alleles over generations. Moreover, I identify loci associated with the cost of resistance and show the potential role of genetic-hitchhiking in maintaining the cost. Lastly, I review the current methods available for studying the genetic basis of trait correlations, and highlight the pitfalls associated with such methods. I thus synthesize an analytical framework which can more precisely identify the genes and mechanisms underlying trait correlations. Together, my thesis identifies the genetic basis of adaptive phenotypes in the genus *Ipomoea* and thus narrows the gap between genotype and adaptive phenotypes.

Chapter 1

Introduction

Problem Statement

One of the major goals of ecology and evolutionary biology is to understand how the genome shapes phenotype in natural populations. Phenotypes can be seen to be a result of a combination of evolutionary processes in some sense and can provide insights into the evolutionary dynamics that shape natural population. Thus, there is an increasing need to understand the mechanistic basis of phenotypes, especially for potentially adaptive phenotypes. Although we are starting to map the genes for crucial phenotypes in model species, much less is known about the mechanistic basis of complex phenotypes, especially in non-model species. In regard to this, questions that still need to be answered are -- Is the phenotype environmentally induced, or is it genetically controlled? What gene(s) underlie the phenotype? Are the gene(s) associated with other phenotypes which might impact its evolutionary potential? In my thesis, I answer these questions for two adaptive traits, leaf shape and herbicide resistance, in members of the genus *Ipomoea* using a combination of quantitative and population genetic approaches.

Understanding the complexity of mapping phenotype to genotype

Understanding how genotype maps to the phenotype is one of the most important challenges in modern plant biology. Expression of a phenotype in natural populations occurs due to a combination of neutral evolutionary processes (genetic drift, gene flow), natural selection, and environmentally induced plasticity (Endler, 1986). Plants display exceptional phenotypic

diversity, not just across species but also within species. This variability is very closely linked to the sessile nature of the plants, which requires the plants to quickly adapt to any changes in their biotic and abiotic environment. Thus, phenotypic diversity reflects the plants' ability to adapt to environmental changes and is an important driver of ecological and evolutionary processes (Via & Lande, 1985; Hahn *et al.*, 2019). As such, there is an increasing need for understanding the genetic basis and the molecular mechanism underlying the naturally occurring plant adaptive phenotypic traits.

Phenotype is a result of both the genotype and the environment. Genotype has been shown to have a considerable influence on the phenotype (Mousseau & Roff, 1987; Lande & Shannon, 1996; Anderson, 2012; van Boheemen & Hodgins, 2020). For instance, multiple plant traits like root and canopy architecture, floral morphology, and even susceptibility to various biotic and abiotic factors have been shown to be genetically controlled, and that these phenotypes have evolved via natural selection (Carvalho & Qualset, 1978; Baucom & Mauricio, 2004; Mertens *et al.*, 2018; Colom & Baucom, 2020). Environment similarly has been shown to alter the phenotype, either directly or by imposing a selection pressure (Smith *et al.*, 2011; Gratani, 2014; Couture *et al.*, 2015; Henn *et al.*, 2018; Arnold *et al.*, 2019). For example, multiple invasive species have been shown to display extensive phenotypic variations in new environments, which allows the species to establish themselves in new heterogeneous environments (Pohlman *et al.*, 2005; Davidson *et al.*, 2011; Colautti & Lau, 2015). Additionally, once an invasive species establishes itself successfully in a new environment, selection on phenotypes favoring higher fitness can occur, given genetic diversity exists in the population (Murren *et al.*, 2005; Lavergne & Molofsky, 2007; Barrett, 2015). Notably, genotype and the environment do not exclusively influence the phenotype but work in concert.

Disentangling the proportion of phenotypic variance that is attributable to the environment vs genetic factors helps in predicting the evolvability of the phenotype (Via & Lande, 1985), which is of special interest to breeders to determine whether a phenotype of interest can be selected upon, and even to conservation biologists to understand how plants will react to changing environments. To understand the importance of a phenotype and its response to selection pressures, we must understand how the genotype and the environment independently, and

interactively impact the phenotype. Moreover, identifying genes and gene networks associated with phenotypes is also equally important since it provides insights into the genetic and molecular basis of the phenotypic variations. Furthermore, understanding the genetic basis of trait variation can give us insights into the functional importance of the trait and the physiological processes responsible for the species' growth in a range of environmental conditions. This can further help us understand the limits to which a population can respond to a novel condition via evolutionary shifts (Ackerly *et al.*, 2000) and help us predict its distribution patterns under future conditions (Aitken *et al.*, 2008; Hancock *et al.*, 2011).

Importantly, identifying the genetic underpinnings of a phenotype can shed light on potential evolutionary constraints associated with the phenotype. Most adaptive plant traits are complex and are controlled by multiple genes, one or more of which might be pleiotropic and/or in linkage with neighboring genes. This has important evolutionary consequences -- the presence of either pleiotropy or linkage can influence an unrelated phenotype and thus can limit the evolution of the phenotype of interest. For example, negative genetic interactions (like the traits involved in allocation trade-offs) can impose a constraint on trait evolution (Via & Lande, 1985; Barton & Turelli, 1989). Understanding the genetic architecture of a phenotype is particularly important for breeding practices since negative interactions between traits can impose a constraint in achieving the maximum potential of each trait, and thus limit the genetic improvement of the trait (Falconer, 1996), (Bandillo *et al.*, 2015). In contrast, positive genetic interactions among traits related to adaptations might increase evolvability by reducing the dimensionality of genetic variation (Wagner, 1988; Orr, 2000). Differentiating between pleiotropy and linkage is crucial since these have different evolutionary fates -- a constraint due to pleiotropy is expected to persist over multiple generations whereas one caused by linkage is expected to be more transient due to recombination breaking down linkage.

We thus see that mapping a phenotype to its genotype is a complex problem that requires answering multiple questions -- Is the phenotype genetically controlled? What are the gene(s) influencing the phenotype? Are one or more of these genes associated with other phenotypic traits either through linkage or pleiotropy?

Mapping phenotype to genotype: Quantitative and Population Genetics Approaches

The first step towards understanding the mechanistic basis of a phenotype is to estimate the relative contribution of genotype and environment in controlling the phenotype. In other words, we need to estimate the extent to which a phenotype is genetically determined, commonly referred to as the broad-sense heritability of the trait.

Classically, this involves estimating the degree of resemblance between relatives in a particular environment (Nyquist & Baker, 1991; Falconer & Mackay, 2009). Thus, the choice of family design is very important. For selfing plants, clonal and inbred crops, a large number of genetically identical individuals allows for true replication of a genotype in and across multiple environments, allowing for a more accurate estimation of heritability. For other species wherein cloning or inbreeding are not possible, multiple choice of family designs exists -- parent-offspring, half-sib, full-sib, mixed family design, etc. (Rice & Borecki, 2001)for comprehensive reviews, see (Rice & Borecki, 2001; Visscher & Goddard, 2015). For example, multiple studies have employed parent-offspring regression to estimate the contribution of genetics in determining the phenotype (Vogel *et al.*, 1980; Åkesson *et al.*, 2008; Sanogo *et al.*, 2019). Alternatively, family means or variance component methods (which partition the phenotypic variance into genetic and environmental variance) can be used for other family designs (Ågren & Schemske, 1992; Ritland & Ritland, 1996; Campbell, 1996). Thus, multiple choice of family designs and methods exists for estimating the extent to which a phenotype is genetically determined. An important consideration for designing experiments though must be the choice of the environment -- these should almost always be conducted in the field since these represent the true environment the plants or crops are exposed to. Multiple studies have shown that heritability is overestimated when performed in a controlled environment (eg. greenhouse) as compared to the field (Conner *et al.*, 2003; Winn, 2004). Additionally, heritability should be estimated in multiple environments for adaptive traits to capture its robustness and potential for selection.

Understanding the genetic basis of an adaptive phenotype is a complex process. Studies have heavily relied on gene expression levels and genetic markers to find an association with the phenotype of interest. For easy to breed/cross species and/or species with genomic resources, the

most commonly used method for identifying genes underlying an adaptive phenotype is QTL (Quantitative Trait Loci) mapping. QTL mapping is based on the intuition that the loci controlling the phenotype will be in linkage to markers nearby, which can be identified by a statistical association test. The earliest studies of QTL in plants date back to the 1980s (Stuber *et al.*, 1987; Paterson *et al.*, 1988). More and more studies have employed QTL mapping to study loci underlying important plant phenotypic traits (reviewed (Nguyen *et al.*, 2019). Some examples include fruit characteristics in tomato (Paterson *et al.*, 1988), flowering time rice (Hori *et al.*, 2016), grain yield and root characteristics in maize (Edwards *et al.*, 1992; Aslam *et al.*, 2015), and disease resistance traits in multiple crop species (Young, 1996). The success of a QTL study is determined by the number of segregating individual and polymorphic markers -- a higher number of markers lead to higher accuracy. The cost and time required for a sufficient number of individuals and markers might thus be a limitation to this method (Jamann *et al.*, 2015). Additionally, QTL has limitations with respect to species wherein a segregating population cannot be created via crossing, and even in the case of polyploid species with genomic resources unavailable. In such cases, a way of gaining insights into the genes related to the phenotype is to use a transcriptomics survey. Essentially, this involves identifying transcripts showing variable expression patterns in individuals representing the two ends of the spectrum of phenotypes (reviewed in (Wang *et al.*, 2009). Multiple studies, using this method, have identified genes related to important agronomic traits like stress response genes (Chen & Zhu, 2004), plant height (Hu *et al.*, 2018; Howlader *et al.*, 2020), and various leaf phenotypic traits (Kim *et al.*, 2002, 2020; Kimura *et al.*, 2008; Chitwood & Sinha, 2016). Thus, depending on the species, either of these commonly used quantitative and population genetic methods can be used to identify genes related to the phenotype.

With the advent of sequencing technologies and the lower costs of sequencing, studies have moved to deep whole-genome scans to identify genes associated with complex phenotypes. These broadly fall into two categories -- a selection scan and a genome-wide association scan (GWAS). Over the last decade alone, more than 1000 studies have employed GWAS in crops to decode the phenotype-genotype associations (Liu & Yan, 2019). Like QTL mapping, GWAS also utilizes linkage between genes to identify the genes associated with the phenotype, but unlike QTL, GWAS does not require the individuals to be related. GWAS has been applied to a

plethora of plant traits like various agronomic, morphological traits, response to biotic and abiotic stress, and biochemical/nutritional traits (for comprehensive reviews, see (Liu & Yan, 2019; Gupta *et al.*, 2019). Although GWAS has a high potential for identifying genes associated with complex phenotypes, it has significant limitations, especially issues related to population structure correlation and low-frequency alleles causing false-positives (reviewed in (Liu & Yan, 2019; Tam *et al.*, 2019). In comparison, a selection scan aims to detect selection signatures that are left behind in the nucleotide sequences after a selection sweep has occurred. This method thus is more applicable and useful for detecting loci leading to adaptations. Multiple studies have used a variety of selection statistics that have been developed to identify signatures of selection and adaptation (Chapman *et al.*, 2008; Beissinger *et al.*, 2014; Gould & Stinchcombe, 2017; He *et al.*, 2017; Van Etten *et al.*, 2020; Derbyshire, 2020). In some cases though, some selection statistics might be compounded by the presence of background noises and/or neutral processes, and thus to sieve out false-positives, studies have suggested the use of multiple selection statistics together. To this effect, composite measures have been developed (Lotterhos *et al.*, 2017). Thus, a choice of methods exists for decoding phenotype to genotype.

Study System

Ipomoea is one of the largest genera in the flowering plant family Convolvulaceae, as it contains more than 600-700 species that are found throughout the tropical and subtropical regions of the world. In my dissertation, I examined questions using *Ipomoea batatas* (cultivated sweetpotato), and *Ipomoea purpurea* (common morning glory).

Sweetpotato is one of the most widely cultivated staple crops worldwide (Khoury *et al.*, 2015). It is thought to have been domesticated at least 5000 years ago in Central America or South America. Like many crops, sweetpotato is a polyploid- specifically a hexaploid with 90 chromosomes (2n=6X=90) with an estimated genome size of 4.4 Gb (Ozias-akins & Jarret, 1994). Sweetpotato displays striking morphological variation in leaf shape across its ~6000 documented varieties (Huaman, 1988), but very few studies have examined the extensive leaf shape diversity in this species (Huaman, 1988; Hue *et al.*, 2012; Rosero *et al.*, 2019). Few studies have examined leaf shape phenotypes in sweetpotato, but these are limited to a few cultivars

and/or present traditional measures of leaf shape traits. Additionally, the genetic or transcriptomic basis of leaf shape variation in this species has yet to be considered.

Ipomoea purpurea is a common agricultural weed in the Southeast and Midwest United States. Populations of this species, which have consistently been exposed to glyphosate-based herbicides since the late 1990s (Kuester *et al.*, 2015, 2016), exhibit varying levels of herbicide resistance, both within- and among-population -- while some populations of this species across its range in the southeastern and Midwest United States exhibit high survival following herbicide application (high resistance), other populations exhibit low survival (high susceptibility) (Kuester *et al.*, 2015). Additionally, there is a fitness cost associated with this resistance: resistant populations show lower germination and deteriorated seed quality compared to susceptible populations (Van Etten *et al.*, 2016). Although previous work in this species has indicated the potential role of detoxification underlying this resistance (Van Etten *et al.*, 2020), this work relied on low-coverage sequencing, and so we don't have a full picture of the loci involved in the resistance, and especially in the cost associated with the resistance.

Thus, the *Ipomoea* genus offers a unique opportunity to study phenotype to genotype in diverse species -- a crop that is hexaploid and an invasive diploid species.

Thesis overview

In chapter 2, I perform a multi-level analysis of leaf shape using diverse accessions of sweetpotato (*Ipomoea batatas*) to uncover the role of genetics, environment, and GxE on this important trait. For this, I use a suite of different methods -- morphometric analyses to identify the extent of variation, transcriptomic survey to identify gene expression changes associated with leaf shape, and a field study in two geographically separate common gardens to examine the role of genetics and environment on leaf shape. I show that extensive leaf shape variation exists within *I. batatas* and identify promising candidate genes underlying this variation. Interestingly, when considering traditional measures, I find that genetic factors are largely responsible for most of the leaf shape variation, but that the environment is highly influential when using more quantitative measures *via* leaf outlines.

In chapter 3, I focus on another important adaptive trait -- herbicide resistance in the weed common morning glory (*Ipomoea purpurea*). I perform a multi-level analysis to uncover putative loci involved in nontarget herbicide resistance (NTSR) and cost associated with NTSR, and to examine evolutionary forces underlying the maintenance of resistance in natural populations. I find loci involved in herbicide detoxification, and stress sensing to be under selection, and confirm that detoxification is responsible for glyphosate resistance using a functional assay. Furthermore, I find the role of interchromosomal linkage disequilibrium (ILD) in potentially mediating resistance through generations. Additionally, by combining the selection screen, differential expression, and LD analysis, I identify putative fitness cost loci that are strongly linked to resistance alleles, indicating the role of genetic hitchhiking in maintaining the cost in this species.

In chapter 4, inspired by the finding of potential fitness cost loci in linkage with resistance loci, I synthesize a conceptual review, focusing on understanding the mechanistic basis of trait correlations required to predict the long-term evolutionary trajectory of the correlated traits. I review the phenotypic and marker-assisted methods available in the literature to assess genetic correlation and outline the analytical strategies that have been implemented for teasing apart the underlying mechanistic basis (pleiotropy vs linkage) of genetic trait correlations. I then discuss the pitfalls associated with the currently available methods and suggest strategies that can avoid and address some of these issues. Next, I outline a path for integrating knowledge from phenotypic family-level data, genetic marker data (like GWAS, omics-QTL), and molecular validation tools, to shed light on the genetic architecture of trait correlations.

Finally, in Chapter 5, I synthesize the outcomes of my three chapters and discuss how future work should be directed to address the remaining gaps and expand our current knowledge in the field of ecology and evolutionary genetics. Additionally, in the appendices, I include a series of supplemental figures and tables accompanying each chapter.

References

Ackerly DD, Dudley SA, Sultan SE, Schmitt J, Coleman JS, Linder CR, Sandquist DR, Geber MA, Evans AS, Dawson TE, *et al.* 2000. The Evolution of Plant Ecophysiological Traits : Recent Advances and Future Directions. *Bioscience* 50: 979–995.

Ågren J, Schemske DW. 1992. Artificial selection on trichome number in Brassica rapa. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 83: 673–678.

Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S. 2008. Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary applications* 1: 95–111.

Åkesson M, Bensch S, Hasselquist D, Tarka M, Hansson B. 2008. Estimating heritabilities and genetic correlations: Comparing the 'animal model' with parent-offspring regression using data from a natural population. *PloS one* 3: e1739.

Anderson CJR. 2012. The Role of Standing Genetic Variation in Adaptation of Digital Organisms to a New Environment. *Artificial Life 13*.

Arnold PA, Kruuk LEB, Nicotra AB. 2019. How to analyse plant phenotypic plasticity in response to a changing climate. *The New phytologist* 222: 1235–1241.

Aslam M, Maqbool MA, Cengiz R. 2015. *Drought Stress in Maize (Zea mays L.): Effects, Resistance Mechanisms, Global Achievements and Biological Strategies for Improvement*. Springer.

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, Lorenz A. 2015. A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. *The plant genome* 8: eplantgenome2015.04.0024.

Barrett SCH. 2015. Foundations of invasion genetics: the Baker and Stebbins legacy. *Molecular ecology* 24: 1927–1941.

Barton NH, Turelli M. 1989. Evolutionary quantitative genetics: how little do we know? *Annual review of genetics* 23: 337–370.

Baucom RS, Mauricio R. 2004. Fitness costs and benefits of novel herbicide tolerance in a noxious weed. *Proceedings of the National Academy of Sciences of the United States of America* 101: 13386–13390.

Beissinger TM, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, Buell CR, Kaeppler SM, Gianola D, de Leon N. 2014. A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. *Genetics* 196: 829–840.

van Boheemen LA, Hodgins KA. 2020. Rapid repeatable phenotypic and genomic adaptation following multiple introductions. *Molecular ecology* 29: 4102–4117.

Campbell DR. 1996. Evolution of floral traits in a Hermaphroditic plant: Field measurements of heritabilities and genetic correlations. *Evolution; international journal of organic evolution* 50: 1442–1453.

Carvalho FIF, Qualset CO. 1978. Genetic variation for canopy architecture and its use in wheat breeding 1. *Crop science* 18: 561–567.

Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, Burke JM. 2008. A Genomic Scan for Selection Reveals Candidates for Genes Involved in the Evolution of Cultivated Sunflower (Helianthus annuus). *The Plant Cell* 20: 2931–2945.

Chen WJ, Zhu T. 2004. Networks of transcription factors with roles in environmental stress response. *Trends in plant science* 9: 591–596.

Chitwood DH, Sinha NR. 2016. Evolutionary and Environmental Forces Sculpting Leaf Development. *Current biology: CB* 26: R297–306.

Colautti RI, Lau JA. 2015. Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. *Molecular Ecology* 24: 1999–2017.

Colom SM, Baucom RS. 2020. Belowground Competition Can Influence the Evolution of Root Traits. *The American naturalist* 195: 577–590.

Conner JK, Franks R, Stewart C. 2003. Expression of additive genetic variances and covariances for wild radish floral traits: comparison between field and greenhouse environments. *Evolution; international journal of organic evolution* 57: 487–495.

Couture JJ, Serbin SP, Townsend PA. 2015. Elevated temperature and periodic water stress alter growth and quality of common milkweed (Asclepias syriaca) and monarch (Danaus plexippus) larval performance. *Arthropod-Plant Interactions* 9: 149–161.

Davidson AM, Jennions M, Nicotra AB. 2011. Do invasive species show higher phenotypic plasticity than native species and, if so, is it adaptive? A meta-analysis. *Ecology letters* 14: 419–431.

Derbyshire MC. 2020. Bioinformatic Detection of Positive Selection Pressure in Plant Pathogens: The Neutral Theory of Molecular Sequence Evolution in Action. *Frontiers in microbiology* 11: 644.

Edwards MD, Helentjaris T, Wright S, Stuber CW. 1992. Molecular-marker-facilitated investigations of quantitative trait loci in maize : 4. Analysis based on genome saturation with isozyme and restriction fragment length polymorphism markers. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 83: 765–774.

Endler JA. 1986. *Natural Selection in the Wild. (MPB-21)*. Princeton University Press.

Falconer DS. 1996. *Introduction to Quantitative Genetics*. Pearson Education.

Falconer DS, Mackay TFC. 2009. *Introduction to Quantitative Genetics*. Pearson.

Gould BA, Stinchcombe JR. 2017. Population genomic scans suggest novel genes underlie convergent flowering time evolution in the introduced range ofArabidopsis thaliana. *Molecular Ecology* 26: 92–106.

Gratani L. 2014. Plant phenotypic plasticity in response to environmental factors. *Advances in botany* 2014.

Gupta PK, Kulwal PL, Jaiswal V. 2019. Association mapping in plants in the post-GWAS genomics era. *Advances in genetics* 104: 75–154.

Hahn PG, Agrawal AA, Sussman KI, Maron JL. 2019. Population Variation, Environmental Gradients, and the Evolutionary Ecology of Plant Defense against Herbivory. *The American naturalist* 193: 20–34.

Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, Toomajian C, Roux F, Bergelson J. 2011. Adaptation to climate across the Arabidopsis thaliana genome. *Science* 334: 83–86.

He Q, Kim K-W, Park Y-J. 2017. Population genomics identifies the origin and signatures of selection of Korean weedy rice. *Plant biotechnology journal* 15: 357–366.

Henn JJ, Buzzard V, Enquist BJ, Halbritter AH, Klanderud K, Maitner BS, Michaletz ST, Pötsch C, Seltzer L, Telford RJ, *et al.* 2018. Intraspecific Trait Variation and Phenotypic Plasticity Mediate Alpine Plant Species Response to Climate Change. *Frontiers in plant science* 9: 1548.

Hori K, Matsubara K, Yano M. 2016. Genetic control of flowering time in rice: integration of Mendelian genetics and genomics. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 129: 2241–2252.

Howlader J, Robin AHK, Natarajan S, Biswas MK, Sumi KR, Song CY, Park J, Nou I. 2020. Transcriptome Analysis by RNA–Seq Reveals Genes Related to Plant Height in Two Sets of Parent-hybrid Combinations in Easter lily (Lilium longiflorum). *Scientific reports* 10: 1–15.

Huaman Z. 1988. Current status on the maintenance of sweet potato genetic resources at CIP. In: Lima (Peru): CIP, 101–120.

Hu F, Chen Z, Zhao J, Wang X, Su W, Qin Y, Hu G. 2018. Differential gene expression between the vigorous and dwarf litchi cultivars based on RNA-Seq transcriptome analysis. *PloS one* 13: e0208771.

Hue SM, Chandran S, Boyce AN. 2012. Variations of leaf and storage roots morphology in Ipomoea batatas L. (Sweet Potato) cultivars. In: Acta Horticulturae. 73–80.

Jamann TM, Balint-Kurti PJ, Holland JB. 2015. QTL mapping using high-throughput sequencing. *Methods in molecular biology* 1284: 257–285.

Khoury CK, Heider B, Castañeda-Álvarez NP, Achicanoy HA, Sosa CC, Miller RE, Scotland RW, Wood JRI, Rossel G, Eserman LA, *et al.* 2015. Distributions, ex situ conservation priorities, and genetic resource potential of crop wild relatives of sweetpotato [ipomoea batatas (L.) lam., I. series batatas]. *Frontiers in plant science* 6: 1–14.

Kim SH, Kim SW, Lim G-H, Lyu JI, Choi H-I, Jo YD, Kang S-Y, Kang B-C, Kim J-B. 2020. Transcriptome analysis to identify candidate genes associated with the yellow-leaf phenotype of a Cymbidium mutant generated by γ-irradiation. *PloS one* 15: e0228078.

Kim GT, Shoda K, Tsuge T, Cho KH, Uchimiya H, Yokoyama R, Nishitani K, Tsukaya H. 2002. The ANGUSTIFOLIA gene of Arabidopsis, a plant CtBP gene, regulates leaf-cell expansion, the arrangement of cortical microtubules in leaf cells and expression of a gene involved in cell-wall formation. *The EMBO journal*.

Kimura S, Koenig D, Kang J, Yoong FY, Sinha N. 2008. Natural Variation in Leaf Morphology Results from Mutation of a Novel KNOX Gene. *Current biology: CB* 18: 672–677.

Kuester A, Chang S-M, Baucom RS. 2015. The geographic mosaic of herbicide resistance evolution in the common morning glory, Ipomoea purpurea: Evidence for resistance hotspots and low genetic differentiation across the landscape. *Evolutionary applications* 8: 821–833.

Kuester A, Wilson A, Chang S-M, Baucom RS. 2016. A resurrection experiment finds evidence of both reduced genetic diversity and potential adaptive evolution in the agricultural weed Ipomoea purpurea. *Molecular ecology* 25: 4508–4520.

Kuester A, Wilson A, Chang S-M, Baucom RS. A resurrection experiment finds evidence of both reduced genetic diversity and potential adaptive evolution in the agricultural weed Ipomoea purpurea.

Lande R, Shannon S. 1996. The role of genetic variation in adaptation and population persistence in a changing environment. *Evolution; international journal of organic evolution* 50: 434–437.

Lavergne S, Molofsky J. 2007. Increased genetic variation and evolutionary potential drive the success of an invasive grass. *Proceedings of the National Academy of Sciences of the United States of America* 104: 3883–3888.

Liu H-J, Yan J. 2019. Crop genome-wide association study: a harvest of biological relevance. *The Plant journal: for cell and molecular biology* 97: 8–18.

Lotterhos KE, Card DC, Schaal SM, Wang L, Collins C, Verity B. 2017. Composite measures of selection can improve the signal-to-noise ratio in genome scans (J Kelley, Ed.). *Methods in ecology and evolution / British Ecological Society* 8: 717–727.

Mertens A, Brys R, Schouppe D, Jacquemyn H. 2018. The impact of floral morphology on genetic differentiation in two closely related biennial plant species. *AoB plants* 10: ly051.

Mousseau TA, Roff DA. 1987. Natural selection and the heritability of fitness components. *Heredity* 59 (Pt 2): 181–197.

Murren CJ, Denning W, Pigliucci M. 2005. Relationships between vegetative and life history traits and fitness in a novel field environment: Impacts of herbivores. *Evolutionary ecology* 19: 583–601.

Nguyen KL, Grondin A, Courtois B, Gantet P. 2019. Next-Generation Sequencing Accelerates Crop Gene Discovery. *Trends in plant science* 24: 263–274.

Nyquist WE, Baker RJ. 1991. Estimation of heritability and prediction of selection response in plant populations. *Critical reviews in plant sciences* 10: 235–322.

Orr HA. 2000. Adaptation and the cost of complexity. *Evolution; international journal of organic evolution* 54: 13–20.

Ozias-akins P, Jarret RL. 1994. Nuclear DNA Content and Ploidy Levels in the Genus Ipomoea. *Journal of the American Society for Horticultural Science. American Society for Horticultural Science* 119: 110–115.

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD. 1988. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335: 721–726.

Pohlman CL, Nicotra AB, Murray BR. 2005. Geographic range size, seedling ecophysiology and phenotypic plasticity in Australian Acacia species. *Journal of biogeography* 32: 341–351.

Rice TK, Borecki IB. 2001. Familial resemblance and heritability. *Advances in genetics* 42: 35–44.

Ritland K, Ritland C. 1996. Inferences about quantitative inheritance based on natural population structure in the yellow monkeyflower, Mimulus guttatus. *Evolution; international journal of organic evolution* 50: 1074–1082.

Rosero A, Granda L, Pérez J-L, Rosero D, Burgos-Paz W, Martínez R, Morelo J, Pastrana I, Burbano E, Morales A. 2019. Morphometric and colourimetric tools to dissect morphological diversity: an application in sweet potato [Ipomoea batatas (L.) Lam.]. *Genetic resources and crop evolution*.

Sanogo O, Tongoona PB, Ofori K, Offei SK, Desmae H. 2019. Parent-offspring regression, correlation and genetic advance of drought and yield traits at early generation in groundnut (Arachis hypogaea L.). *African journal of agricultural research* 14: 1349–1358.

Smith EA, Collette SB, Boynton TA, Lillrose T, Stevens MR, Bekker MF, Eggett D, St Clair SB. 2011. Developmental contributions to phenotypic variation in functional leaf traits within quaking aspen clones. *Tree physiology* 31: 68–77.

Stuber CW, Edwards MD, Wendel JF. 1987. Molecular Marker-Facilitated Investigations of Quantitative Trait Loci in Maize. II. Factors Influencing Yield and Its Component Traits. *Crop science* 27: 639.

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. 2019. Benefits and limitations of genome-wide association studies. *Nature reviews. Genetics* 20: 467–484.

Van Etten ML, Kuester A, Chang S-M, Baucom RS. 2016. Fitness costs of herbicide resistance across natural populations of the common morning glory, Ipomoea purpurea. *Evolution; international journal of organic evolution* 70: 2199–2210.

Van Etten M, Lee KM, Chang S-M, Baucom RS. 2020. Parallel and nonparallel genomic responses contribute to herbicide resistance in Ipomoea purpurea, a common agricultural weed. *PLoS genetics* 16: e1008593.

Via S, Lande R. 1985. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution; international journal of organic evolution* 39: 505–522.

Visscher PM, Goddard ME. 2015. A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. *Genetics* 199: 223–232.

Vogel KP, Haskins FA, Gorz HJ. 1980. Parent-Progeny Regression in Indiangrass: Inflation of Heritability Estimates by Environmental Covariances 1. *Crop Science* 20: 580–582.

Wagner GP. 1988. The influence of variation and of developmental constraints on the rate of multivariate phenotypic evolution. *Journal of evolutionary biology* 1: 45–66.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* 10: 57–63.

Winn AA. 2004. Natural selection, evolvability and bias due to environmental covariance in the field in an annual plant. *Journal of evolutionary biology* 17: 1073–1083.

Young ND. 1996. QTL mapping and quantitative disease resistance in plants. *Annual review of phytopathology* 34: 479–501.

Chapter 2

Assessing the Remarkable Morphological Diversity and Transcriptomic Basis of Leaf Shape in *Ipomoea batatas* (sweetpotato)[1]

Abstract

Leaf shape, a spectacularly diverse plant trait, varies across taxonomic levels, geography, and in response to environmental differences. However, comprehensive intraspecific analyses of leaf shape variation across variable environments is surprisingly absent. Here, we performed a multi-level analysis of leaf shape using diverse accessions of sweetpotato (*Ipomoea batatas*), and uncovered the role of genetics, environment, and GxE on this important trait. We examined leaf shape using a variety of morphometric analyses and complement this with a transcriptomic survey to identify gene expression changes associated with shape variation. Additionally, we examined the role of genetics and environment on leaf shape by performing field studies in two geographically separate common gardens. We showed that extensive leaf shape variation exists within *I. batatas* and identified promising candidate genes underlying this variation. Interestingly, when considering traditional measures, we found that genetic factors are largely responsible for most of leaf shape variation, but that the environment is highly influential when using more quantitative measures *via* leaf outlines. This extensive and multi-level examination of leaf shape shows an important role of genetics underlying a potentially important agronomic

[1] This chapter has been published as Gupta S, Rosenthal DM, Stinchcombe JR, Baucom RS. 2019. The remarkable morphological diversity of leaf shape in sweetpotato (Ipomoea batatas): the influence of genetics, environment, and G×E. *The New phytologist*.

trait, and highlights that the environment can be a strong influence when using more quantitative measures of leaf shape.

Introduction

Leaf shape varies spectacularly among plant species at multiple taxonomic levels (Klein *et al.*, 2017; Shi *et al.*, 2019), across geography (Wyatt & Antonovics, 1981; Gurevitch, 1988), and in response to environmental differences (Andersson, 1991; Jones, 1995; McDonald *et al.*, 2003). Leaves can vary with respect to their degree of dissection, length-to-width ratio, venation patterning, prominence of tips and petiolar sinus, or any combinations of the above, meaning that leaf shape variation across species is multifaceted and complex. Leaf shape diversity is also present within species (Hilu, 1983). For example, accessions of grapevine and cotton vary with respect to leaf complexity whereas lineages within tomato and apple show ample variation in the length-to-width ratio of leaves (Chitwood *et al.*, 2013; Andres *et al.*, 2016; Klein *et al.*, 2017; Migicovsky *et al.*, 2017). Although a large number of species exhibit variation in leaf shape, examinations within species are often limited to only a few accessions, with a few notable exceptions (Conesa *et al.*, 2012; Chitwood *et al.*, 2014a, b). Moreover, these studies often focus on circularity and length-to-width ratio, which are the most common leaf shape descriptors. Thus, for most species, truly quantitative analyses of the diversity of leaf shape variation within species remains largely unexamined.

Leaf shape variation is regulated by genetics, the environment, and the interaction of genes and environment (GxE). Although the genetic and trancriptomic basis underlying leaf shape diversity has been uncovered in only a small number of species (*i.e.*, tomato, Arabidopsis, cotton, and a few others; Kim *et al.*, 2002; Kimura *et al.*, 2008; Vlad *et al.*, 2014; Ichihashi *et al.*, 2014; Andres *et al.*, 2016; Chitwood & Sinha, 2016), there are many examples showing the influence of different environments on leaf shape (McDonald *et al.*, 2003; Zwieniecki *et al.*, 2004; Hopkins *et al.*, 2008; Royer *et al.*, 2009; Nicotra *et al.*, 2011; Royer, 2012; Campitelli & Stinchcombe, 2013; Glennon & Cron, 2015). For example, submerged leaves of aquatic plants are often highly dissected as compared to their aerial counterparts (Arber, 2010) and leaves

growing in colder environments tend to be more complex than similar ones growing in warmer environments (Huff *et al.*, 2003; Royer *et al.*, 2005). Moreover, the environment can interact with genes to further modulate leaf shape. For instance, Nakayama and colleagues (2014) found that changes in temperature leads to abrupt changes in KNOX1 (*KNOTTED1-LIKE HOMOEOBOX1*) activity, a key regulator of circularity in multiple species, thus altering leaf complexity. Although we are beginning to understand how genetics, environment, and GxE separately influence aspects of leaf shape, few studies have partitioned the effect of genetics versus the environment on leaf shape variation, and most examinations are limited to only one environment, such that the role of GxE on leaf shape is often not considered within species.

Leaf shape is most commonly quantified using the 'traditional' leaf shape traits -- circularity (a measure of leaf dissection, or 'lobedness'), aspect ratio (the length-to-width ratio of a leaf) and solidity (the relation of the area and convex hull). These traditional morphometric parameters have previously been used to quantity leaf shape in diverse species, such as grapes (Chitwood *et al.*, 2014b), tomato (Chitwood *et al.*, 2015) and sweetpotato (Rosero *et al.*, 2019), among others. Although these traits are linked to important yield traits in crops (Chitwood *et al.*, 2013; Vuolo *et al.*, 2016; Chitwood & Otoni, 2017; Klein *et al.*, 2017; Rowland *et al.*, 2019), and are important for understanding the broader aspects of plant adaptation to environment, they capture only a few components of leaf shape variation. A more comprehensive quantification of leaf shape can be captured with Elliptical Fourier Descriptor (EFD) analyses, which converts leaf outlines to harmonic coefficients allowing for Fourier analyses (Chitwood & Sinha, 2016). This approach captures extensive leaf shape variation due to both symmetry and asymmetry of the leaf; some examples include shape differences associated with the depth of the petiolar sinus, the prominence of the leaf tip, and the positioning of the lobes. This approach has been applied to a handful of species like tomatoes, passiflora, and grape (Chitwood *et al.*, 2013; Chitwood & Otoni, 2017; Klein *et al.*, 2017), where it was shown that leaf shape based on EFD analysis is highly heritable. Thus, traditional measures along with consideration of leaf outlines holds greater power to comprehensively measure and characterize leaf shape, which may yield important insights about the genetic basis of leaf shape variation. Interestingly, while leaf shape based on EFD analysis is heritable, no studies have yet examined the genetic or transcriptomic basis of leaf shape based on leaf outlines.

Ipomoea batatas, the sweetpotato, is an important staple root crop worldwide (Khoury *et al.*, 2015), as it produces the highest amount of edible energy per hectare (Khoury *et al.*, 2015) and also provides an important source of nutrients in the form of vitamin A, calcium, and iron (Kays & Kays, 1998). Sweetpotato displays striking morphological variation in leaf shape across its ~6000 documented varieties (Huaman, 1987), but very few studies have examined the extensive leaf shape diversity in this species (Huaman, 1987; Hue *et al.*, 2012; Rosero *et al.*, 2019). Studies that have examined leaf shape phenotypes in sweetpotato are limited to a few cultivars and/or present traditional measures of leaf shape traits. Additionally, the genetic or transcriptomic basis of leaf shape variation in this species has yet to be considered. The vast unexamined diversity of leaf shape in this species, along with its role as a staple food crop worldwide makes *I. batatas* an ideal study system to investigate leaf shape diversity at the species level and how this diversity is influenced by the interplay between genetics and environment.

Here, we examine the extensive leaf shape variation within accessions of *I. batatas*, and uncover the role of genetics, environment and GxE in influencing leaf shape traits. We specifically ask: (1) How diverse is leaf shape at a species-wide level? (2) what are the candidate genes associated with leaf shape (extending beyond the traditional shape descriptors)? and (3) to what degree does the environment and GxE influence leaf shape traits? We show that extensive natural variation exists in leaf shape within this species and that most of this variation is largely controlled by genetic factors, with a low proportion of variance in leaf shape attributable to environmental differences. We also identified promising candidate genes that underlie broad differences in multiple leaf shape traits. The results of our work fill critical gaps in current knowledge of leaf shape evolution by expanding analysis beyond that of the traditional measures of leaf shape and by using many distinct lineages of the species. We unite this with the transcriptomic basis of these traits along with a multiple-environment assessment of leaf shape variation in the field. Thus, this work allows us to comprehensively assess leaf shape in this agronomically important species and partition the role of genetics, environment, and GxE on leaf shape within this species.

Materials and Methods

Leaf shape variation within I. batatas

We ordered vegetative slips for 68 publicly available accessions of sweetpotato from USDA and online resources. The location of origin of 68 accessions is represented in Figure 2-1 (*Appendix A,* Table S2-1). The accessions represent the majority of the genetic variation in the species; we identified three of the four population structure clusters among our chosen accessions as per a recent study (Wadl *et al.*, 2018). We grew slips at the UM Matthaei Botanical Garden under standardized growth conditions (16 hrs light/8 hrs night cycle) for approximately six months, at which time we sampled 4-6 mature leaves (third-sixth mature leaves from the beginning of the vine to control for age and exposure to light) of 57 randomly chosen accessions and scanned them for leaf shape analyses.

We used the scanned images to extract leaf shape trait values using custom macros in ImageJ (Abràmoff *et al.*, 2004). Briefly, we converted leaves into binary images and then used outlines from these binary images to measure circularity, aspect ratio and solidity, each capturing a distinct aspect of leaf shape (Li *et al.*, 2018). Circularity, measured as $4\pi \frac{area}{perimeter^2}$, is influenced by serrations and lobing. Aspect ratio, in comparison, is measured as the ratio of the major axis to the minor axis of the best fitted ellipse, and is influenced by leaf length and width. Lastly, solidity measured as $\frac{area}{convex\ hull}$, is sensitive to leaves with deep lobes, or with a distinct petiole, and can be used to distinguish leaves lacking such structures. Solidity, unlike circularity, is not very sensitive to serrations and minor lobings, since the convex hull remains largely unaffected.

For a more global analysis of leaf shape via Elliptical Fourier Descriptor (EFDs), we used the program SHAPE (Iwata & Ukai, 2002) as described in (Chitwood *et al.*, 2014b). EFDs capture variation in shape represented by the outline which is difficult to categorize via traditional shape descriptors. From the EFD coefficients obtained, we used coefficients a and d only, thus analyzing symmetric variation in leaf shape. Principal component analysis (PCA) was performed on the EFD coefficients to identify shape features contributing to leaf morphological variation (referred to as EFD symPCs below). We calculated the correlation matrices using the rcorr()

function of the Hmisc package version 4.0-3 (Harrell *et al*., 2017) with multiple test adjustments using the p.adjust() function in R.

RNA-Seq library construction and sequencing

We sequenced and analyzed transcriptomes of 19 individuals of *I. batatas* to examine gene expression differences associated with leaf shape variation associated with circularity, aspect ratio, and EFD symPCs to obtain an initial set of candidate genes underlying these traits. We selected greenhouse-grown accessions with differing leaf shape trait values (*Appendix A,* Figure S2-1). Since high aspect ratio represents both longitudinally longer or latitudinally broader leaf shape phenotypes, we chose to only examine individuals that had high aspect ratio due to latitudinal elongation. We chose multiple accessions to assess each leaf shape trait; eleven for circularity (six entire, five lobed), eight for aspect ratio (four high and low AR, each), 6 individuals for EFD symPC1 (three high and three low) and four accessions each for EFD symPC2 and EFD symPC3 (two high and two low) (*Appendix A,* Figure S2-1); EFD symPC4 was not considered for differential expression analysis.

We used three to five leaves that were in P4-P6 stage of growth (fourth to sixth youngest primordium), from multiple branches of each individual accession for RNA extractions, and combined replicate leaves per individual to increase the depth of the transcriptome. We sampled all individuals on the same day within 1 hour to reduce variation due to developmental stage and/or time of collection. We froze samples in liquid nitrogen prior to preserving them at -80^{o} for further processing. We performed RNA extraction using Qiagen RNeasy Plant mini kit with the optional DNase digestion step, and constructed libraries using the TruSeq Stranded mRNA Sample Preparation protocol (LS protocol). After barcoding, we bulked all libraries and performed one lane of Illumina HiSeq2500 sequencing.

RNA-Seq data processing and transcriptome analysis

An overview of our RNA-Seq data processing and transcriptome analysis is given in Figure 2-2, with detailed information presented in Method S2-1 (*Appendix A*).

Differential gene expression--We mapped reads from all 19 individuals to the *de novo* assembled transcriptome using BWA-MEM v0.7.15 (Li, 2013) and estimated read counts for uniquely

mapped reads using samtools v1.9 (Li *et al.*, 2009). We then used read counts to filter out lowly expressed transcripts using the Bioconductor package edgeR version 3.18.1 (Robinson *et al.*, 2010) such that transcripts were retained only if they had greater than 0.5 counts-per-million in at least two samples. We then normalized libraries in edgeR (using the trimmed mean of *M*-values method) followed by differential gene expression analysis using classic pairwise comparison of edgeR version 3.18.1. We extracted the significance of differentially expressed transcripts (DETs) with FDR <= 0.05.

Field experiment

We performed a field experiment to determine the extent to which genetics, the environment, and GxE interactions influence leaf shape traits. We generated replicate individuals by planting 5 cm cuttings of the stem of each accession in 4-inch pots, randomly positioned on a mist bench at the Matthaei Botanical Gardens. During the first week of June, we planted three to seven replicates of each of the 68 accessions in two common gardens--one located at the Matthaei Botanical Gardens in Ann Arbor, MI (42.18° N, 83.39° W), and the other at the Ohio University Student Farm, West State Street Research Site in Athens, OH (40.46° N, 81.55° W). Replicates were planted in either three (MI) or seven (OH) blocks in a completely randomized block design with 14-inch spacing between individuals. Blocks were kept relatively weed free but were otherwise allowed to grow undisturbed. We randomly sampled 2-5 mature leaves from each individual in the first week of October, prior to the first frost, and scanned them for leaf shape analyses as explained before.

Data analysis--We first examined the potential for variation in leaf shape due to environmental differences (i.e. variation due to being grown in MI or OH) by performing an ANOVA. To normalize leaf shape traits, we used the function TransformTukey from rcompanion version 2.0.0 (Mangiafico, 2018). TransformTukey is a power transformation based on Tukey's ladder of Powers, which loops through multiple powers and selects the one that normalizes the data most. These normalized leaf shape traits were then used as dependent variables and accession, garden, block effects and an interaction term of accession and garden as independent variables in the following fixed-effects model:

(Trait ~ Accession + garden + block + Accession:garden).

The term accession represents the genetic component, garden represents variation due to environment (plasticity), Accession:garden represents the GxE component and the block effect captures microenvironmental variation (and was nested within each garden). To quantify the relative effects of each of these variables on leaf shape, we calculated eta squared (η2) as a measure of the magnitude of effect size using the Bioconductor package lsr version 0.5 (Navarro, 2013). Eta squared for an effect is measured as SS_{effect}/SS_{total}, where SS_{effect} is the sum of squares of the effect of interest and SS_{total} is the total sum of squares of all the effects, including interactions. In other words, it is a measure of the proportion of variance in the dependent variable associated with independent variable and is one of the most commonly reported estimates of effect size for ANOVA (Levine & Hullett, 2002; Ialongo, 2016). Further, we calculated broad sense heritabilities of leaf shape traits to determine the extent to which traits are genetically controlled within each environment. Broad sense heritability was calculated using linear mixed modeling with the Bioconductor package sommer version 3.4 (Covarrubias-Pazaran, 2016) based on the phenotypic data collected from the two fields. The model used was

Trait~1, random=~Accession + block + Accession:block, rcov= ~units

Variance components from the model were used to calculate the broad-sense heritability (H^2) using the formula:

$$H^2 = \frac{V_g + V_e + V_{gxe} + V_r}{V_g}$$

where V_g is the genotype variance, V_e is the environmental variance due to the blocks, V_{gxe} is the variance associated with V_{gxe} (accession:block), and V_r is the residual variance.

Results

Leaf shape variation among accessions

We found wide variation in leaf traits across 57 *I. batatas* accessions (Table 2-1). Among the three traditional traits examined, circularity is most variable with a phenotypic coefficient of variation (PCV; (standard deviation(x)/mean(x))*100; where x is the trait of interest) of 22.61%

while aspect ratio is least variable with a narrow distribution and PCV of 4.76%. Figure 2-3 shows the phenotypic diversity with respect to two leaf traits, circularity and aspect ratio (AR). Of our 57 accessions, 10 exhibit low circularity (defined as circularity < 0.50). PI 599387, for example, exhibited leaves that are very deeply lobed and thus has a low circularity (0.09) value. In contrast, PI 566647 has no serrations or lobing (entire margins) and thus exhibits high circularity (0.71; Figure 2-3). Additionally, we found 22 of 57 accessions to exhibit high aspect ratio (AR > 1.11). For example, PI 531134 (AR = 1.03) has almost equal values of major and minor axis and thus a low aspect ratio value. In contrast, the leaves of PI 208886 (AR = 1.268) are much wider, i.e., a larger major to minor axis, and thus has high aspect ratio value. Most often this increase in AR in sweetpotato manifests itself with increase leaf width (eg. PI 566646, PI 208886) relative to length (eg. PI 634379). Further, although solidity values range from 0.44-0.95, only 5 accessions had solidity values less than 0.7 (PCV = 11.85%). The lack of low solidity values indicates that only a few accessions have deeply lobed leaves (eg. PI 599387, solidity = 0.44), in contrast to accessions with slightly lobed leaved (eg. PI 566630, solidity = 0.76).

We performed an EFD analysis on leaf outlines to get a more global estimation of leaf shape variation (Figure 2-4). In total, we processed 292 leaves from 57 accessions to identify leaf shape traits that explain symmetrical shape variation in sweetpotato. Low symPC1 values describe leaves with deep lobing, prominent tip and shallow petiolar sinus (PI 573318) whereas high symPC1 values explain non-lobed leaves with flattened leaf tips and enclosed petiolar sinus (PI 566646). symPC2 explains variation in leaf shape due to differences in breadth and lobing of the leaf (low symPC2 values describe broad leaves with two lobes whereas high symPC2 values depicts narrow leaves with no lobes). symPC3 primarily captures leaf shape variation due to the depth of petiolar sinus (low symPC3 values describe leaves with highly enclosed petiolar sinus as compared to high symPC3 eigenleaves which have flattened sinus). Lastly, symPC4 represents variation in leaf shape attributed to the angle of lobe tips -- low symPC4 eigenleaves have lobes with a high obtuse angle (almost 160º) whereas high symPC4 eigenleaves have lobes with a lower obtuse angle (almost 125º). The four symPC components together explain 87.79% of total variance relating to symmetrical leaf shape variance in sweetpotato.

Further, we calculated correlation matrices for traditional shape descriptors and EFD symPCs to determine if they capture different aspects of leaf shape (*Appendix A,* Figure S2-2). We found that symPC1 is correlated with circularity ($r = 0.20$; $P = 0.03$) and solidity ($r = 0.20$; $P = 0.02$), which is expected as symPC1 partially captures shape differences due to lobing. Additionally, circularity was highly correlated with solidity ($r = 0.96$; $P < 0.001$). This is not surprising as circularity is a measure of serrations and lobing whereas solidity is a measure of deep lobing; leaves having deep lobes (and lacking serrations) will thus have similar values of circularity and solidity.

Sequencing and de novo assembly of I. batatas transcriptome

We performed a transcriptomic survey to identify gene expression changes associated with the leaf shape traits described above. For our analyses of the transcriptome, Illumina HiSeq2500 returned a total of 266 million (125bp) paired-end sequence reads; on average, each individual had 14 million (M) reads (GEO Submission ID-GSE128065) which was used to construct a *de-novo* transcriptome assembly (sequence statistics are presented in Table 2-2). The results from BUSCO (Simão *et al.*, 2015) indicate that the *de novo* transcriptome assembly is of high quality with 91.32% (1315/1440) complete genes found (single copy genes ~87%) of which only 4.51% were duplicates. Additionally, only 6.32% of genes were missing from the assembled transcriptome. Thus, our sequencing and assembly strategy produced a relatively complete transcriptome. Using blastx, 24,565 transcripts were annotated by the functional description of their top 20 hits. The transcriptome is available at Transcriptome Shotgun Assembly Database hosted by NCBI (TSA accession # GHHM01000000).

Identification and functional annotation of differentially expressed transcripts (DETs)

As a first step towards understanding the genetic control of leaf shape, we identified gene expression changes associated with multiple leaf shape traits -- circularity, aspect ratio (latitudinal expansion) and the symPCs obtained from the EFD analysis. We did not consider solidity and symPC4 due to their high correlation to circularity and low level of variation captured, respectively. On average, we found that 11 million unique paired-end reads per individual (range 7.66M - 14.23M) mapped back to the reference transcriptome (net mapping efficiency of 89.65% with the paired-end high-quality reads). This indicates that we had

sufficient read depth (>10M) to continue with our differential expression analysis (as shown by Wang *et al.*, 2011).

We uncovered 530 DETs associated with our leaf shape traits (Figure 2-5; *Appendix A,* Table S2-2). Specifically, we found 47 DETs associated with circularity, and 158 DETs associated with aspect ratio. For the symPCs examined, we found 121 DETs associated with symPC1, 148 DETs with symPC2 and 56 DETs with symPC3. Functional annotation of these DETs uncovered putative leaf shape genes (Table 2-3). As an example, for circularity, FAR1-related sequence 5 (or *FRS5*), a putative transcription factor involved in regulating light control of development, is differentially regulated with log fold-change of 5.77. Among other DETs for circularity, we found genes that are involved in regulating cell proliferation and organ morphogenesis (EXO70A1-like and extra-large guanine nucleotide-binding protein) and could be involved in regulating leaf dissection.

Among the 158 transcripts differentially expressed for AR (broad leaves vs rounder leaves), two genes have been shown in literature to alter the longitudinal vs latitudinal expansion of the leaves. These are *CHS* (chalcone synthase), an enzyme involved in the production of chalcones involved in flavonoid biosynthesis, and feruloyl CoA 6′-hydroxylase which is involved in scopoletin biosynthesis and causes post-harvest physiological deterioration in cassava (Liu *et al.*, 2017). Finally, we also found LIGHT-DEPENDENT SHORT HYPOCOTYL 10 (*LSH10*), to be significantly downregulated (log-fold change of -1.85; P-value < 0.001).

Individuals with extreme values of symPC1, a trait differentiating leaf shape based on lobing and prominence of tips and petiolar sinus, were also analyzed for DETs. Of the 121 transcripts showing differential expression, two genes had interesting functional annotations. We found a homeobox gene (*HAT22*) to be upregulated in individuals with high symPC1 (leaves lacking lobes with flattened leaf tips and enclosed petiolar sinus), with a log-fold change of 1.56. We also found another member of the *FRF1* family -- FAR1-related sequence 7 (or *FRS7*) -- to be upregulated in the high symPC1 individuals, like in the case of circularity.

We found a total of 148 DETs for symPC2, which explains variation in leaf shape due to the differences in the broadness and lobing of the leaf. Again, we found two copies of chalcone synthase (*CHS*) were negatively regulated in high symPC2 individuals. We also found Sporamin B transcript, a tuberous root protein (Yeh *et al.*, 1997), to be significantly downregulated (with log-fold change of -2.76; P-value < 0.001). Finally, we identified 56 transcripts that were differentially expressed with respect to symPC3; however, functional annotation revealed that most genes belonged to chloroplastic or mitochondrial genes.

Field experiment

We performed a field experiment to examine leaf shape in different environments, with the specific goal to determine the extent to which genotype, environment, and GxE altered leaf shape. We found significant variation among accessions (indicating genotypic or genetic variation) for circularity, aspect ratio and solidity ($F_{73} = 18.06$, $F_{73} = 4.22$, $F_{73} = 21.09$; $P < 0.001$), with accession explaining 73.23%, 38.40% and 77.18% of the total variation, respectively (Table 2-4). This high variance explained for circularity and solidity is reflected in high heritability values (Table 2-5; $H^2_{MI\_cir} = 0.79$, $H^2_{OH\_cir} = 0.73$; $H^2_{MI\_solidity} = 0.82$, $H^2_{OH\_solidity} = 0.76$). We also found evidence of significant block effect ($F_8 = 3.01$, $P = 0.002$; $\eta^2 = 1.33\%$) for circularity, whereas aspect ratio and solidity were not significantly influenced by block effects. Garden differences between OH and MI contributed 1.93% ($F_1=15.55$, $P <0.001$) of the variability in AR while the accession by garden interaction contributed 12.95% (a significant GxE effect: $F_{69} = 5.01$, $P = 0.009$). AR also had lower heritability within each garden (Table 2-5; $H^2_{MI\_AR} = 0.39$, $H^2_{OH\_AR} = 0.26$). Circularity and solidity were not significantly altered by environment and had no significant differences due to GxE.

We also examined symmetrical leaf shape variation in both field sites by performing an EFD analysis (Figure 2-6). EFDs from MI captured variation in leaf shape homologous to the symPCs estimated from greenhouse grown individuals. There was general congruence in symPCs between greenhouse and field grown leaves in MI (i.e., MIsymPC1 (field) ≈symPC1 (greenhouse)), but leaf shape variation captured by EFDs from OH differed significantly in their order of variation explained (Figure S2-3). OHsymPC1 explained leaf shape variation due to differences in the broadness and lobing of the leaf (similar to MIsymPC2), whereas OHsymPC2

explained variation due to lobing, tip and petiolar sinus differences (similar to MIsymPC1). This indicates that in OH the majority of leaf shape diversity is primarily due to the broadness of the leaf and secondly due to leaf lobing, while in MI, it is the opposite-- the majority of leaf shape diversity is due to the leaf dissection rather than leaf width. Thus, although traditional shape descriptors are only slightly influenced by the environment, leaf shape as a whole can be altered significantly by the environment.

We also calculated broad sense heritability values for the symPCs in their respective environments and found that H^2 values ranged from 0.47-0.80 across the symPCs (Figure 2-6). Heritability values in the OH garden were consistently lower than in the MI garden due to reduced genetic variance and increased environmental variance. Overall, the high heritability values indicate that leaf morphology is controlled to a great extent by genetic factors.

Discussion

In this study, we examined the extent of leaf shape variation within an agronomically important species, determined the role of genetics, the environment and GxE in altering leaf shape traits, and identified potential candidate genes associated with multiple leaf shape traits. We found evidence of extensive intraspecific morphological variation, with shape differences due to lobing, length-to-width ratio of leaves and the prominence of tip and petiolar sinuses explaining the majority of the variation. We also found that leaf shape has a strong genetic basis with most phenotypic variation attributed to accessional variation, with low or limited influence of GxE. Strikingly, we show that although traditional shape descriptors are only slightly influenced by the environment in this species, when measured comprehensively, leaf shape can be significantly altered by the environment (evident by the change in symPC1 across the MI and OH gardens). Below, we expand on each of our findings, and place them in the context of current knowledge about leaf shape diversity at a species-level as well as what is known about the environmental influence on leaf shape in other species.

High morphological diversity of leaf shape in I. batatas

A recurring question among plant morphologists is the extent to which leaf shape varies among genotypes in a species. This study quantified leaf shape variation among multiple replicated accessions of sweetpotato and identified traits contributing most to leaf shape variation. We focused our morphometric study on three traditional shape descriptors (circularity, aspect ratio and solidity) and then expanded into the more comprehensive Elliptical Fourier Descriptor (EFD) measures.

In our analysis of traditional measures, circularity was found to be the most variable whereas aspect ratio was found to be least variable. Further, the first two principal components of the EFD analysis together accounted for 77.46% of the total variation in leaf shape, and described variation associated with petiolar sinus, tips, and positioning of lobes. Additionally, lack of correlation between symPCs and traditional leaf shape metrics suggests that they capture different features of shape. Only symPC1 was slightly correlated with circularity and solidity. This is not surprising since symPC1 captures variation in leaf shape due to lobing, tip and sinus. No other traits were found to be correlated. Thus, variation captured by the EFD symPCs would have been missed by simply quantifying traditional shape descriptors, suggesting that the use of comprehensive morphometric techniques can help quantify the full extent of shape variation across species. Further, combining the results from traditional morphometric approaches with EFDs revealed that variation in leaf dissection (circularity and symPC1) contributes most to the morphological variation in leaf shape in sweetpotato (Figure 2-3 and Figure 2-4), similar to that seen in grape (Chitwood *et al.*, 2014b). In addition, aspect ratio explains a significant proportion of the remaining variation, unlike in tomato and apple where aspect ratio is the primary trait of variation in leaf shape (Chitwood *et al.*, 2013; Migicovsky *et al.*, 2017). This indicates that leaf shape variation does not follow a trend across species which is likely due to multiple independent evolution of leaf shape across phylogenetic taxas (Nicotra *et al.*, 2011).

Gene transcripts underlying leaf shape variation

To further our understanding of gene expression changes underlying leaf shape diversity, we sequenced transcriptomes of 19 accessions and assembled a high-quality gene expression database for performing a differential expression analysis in *I. batatas*. We found 47 genes that

were differentially expressed for circularity and 121 DETs for symPC1 -- a trait that accounts for leaf shape differences due to leaf dissection, prominence of the tip and petiolar sinus. Functional annotations of these genes identified potential candidates that could contribute to leaf shape dissection in *I. batatas* (Table 2-3). The most promising candidate is *FRS* gene; we found *FRS5* and *FRS7* to be upregulated in non-dissected individuals in the differential analysis for circularity and symPC1, respectively. *FRS* is a putative transcription factor and contains the DNA binding domain needed to bind the RB-box promoter region of *STM* (*SHOOT MERISTEMLESS*) (Aguilar-Martínez *et al.*, 2015), a protein required for leaf serrations (Kawamura *et al.*, 2010). *FRS* might bind to *STM* thus regulating its expression. However, we did not find *STM* to be differentially expressed in our datasets. This might be due to no real expression differences or it might indicate that the expression differences is really small and thus the gene is not detected to be differentially expressed.

Furthermore, genes containing homeobox domains have been shown to be associated with leaf dissection in multiple species --*e.g., PTS* in tomato (Kimura et al., 2008), *STM* in *Arabidopsis (Piazza et al., 2010)*, *RCO* in *C. hirsuta* and other Brassicaceae (Vlad *et al.*, 2014; Sicard *et al.*, 2014) and *LMI1* in cotton (Andres *et al.*, 2016). Most of these genes are differentially regulated in the SAM (shoot apical meristem) and P0 (the youngest primordium) to determine the extent of leaf dissection and complexity for the genotype. However, we did not find any homeobox domain containing genes to be differentially expressed in sweetpotato accessions that varied for circularity (*i.e.* lobed vs entire) (*Appendix A,* Table S2-3) but found a homeobox leucine-zipper protein (*HAT22*) to be upregulated for high symPC1 individuals. This mismatch could represent a caveat to our transcriptomic sampling stage (P4-P6), which is past the leaf dissection morphogenic stage of development. Thus, although preliminary, our data indicate that the degree of lobing in *I. batatas* might be maintained in later stages of leaf development (P4-P6) by the action of a gene containing a homeobox domain and that the difference in expression required might be very small.

Further, we found a total of 158 differentially expressed genes associated with aspect ratio and 148 DETs associated with symPC2 (leaf shape due to the differences in the broadness and lobing). Based on the function of the homologs of these genes, we identified promising putative

candidate genes responsible for broad leaved phenotypes (Table 2-3). In apples, a transgenic *CHS* silenced individual developed longer leaves when supplied with naringenin, thus altering leaf AR. This indicates that higher expression of *CHS* (and thus naringenin) is responsible for the longitudinal expansion of the leaves and thus downregulation of *CHS* could lead to broader leaves due to the lack of longitudinal expansion. Another gene of interest that we found differentially expressed for aspect ratio, feruloyl CoA 6′-hydroxylase, produces broader leaved phenotypes of cassava when silenced (Liu *et al.*, 2017). Interestingly, however, we found *higher* expression of feruloyl CoA 6′-hydroxylase2 in broader-leaved, compared to the rounder-leaved individuals. Finally, the differentially expressed *LSH10* belongs to the family of *LSH* genes, which have been shown to interact with BOP (BLADE-ON-PETIOLE) and regulate *PTS* (PETROSELINUM) expression, a gene that regulates *KNOX* genes, and thus leaf complexity (Ichihashi et al. 2014). This indicates the potential role of *LSH* gene in regulating both leaf broadness and complexity in this species.

Factors influencing leaf shape traits in multiple environments

While studies often examine the potential for plasticity in leaf shape traits (McLellan, 2000; Royer *et al.*, 2009; Viscosi, 2015), the relative influence of genetic background, environment and gene by environment interactions are less commonly examined. We show that leaf shape traits (circularity, aspect ratio and solidity) in sweetpotato are influenced by multiple effects. Variation in circularity and solidity were mostly attributed to accession (or genotype) and showed little to no effect due to environment or gene by environment interaction. Circularity and solidity have exceptionally high broad-sense heritability values in *I. batatas* (0.76 and 0.79 respectively, averaged between gardens). These traits have likewise been shown to be highly heritable in tomato with heritability values being 0.65 and 0.67, respectively (Chitwood *et al.*, 2013). The high PCV for circularity and solidity in *I. batatas* (22.61% and 11.85%) along with high broad-sense heritability indicates that there is a lot of standing variation for these traits that can be actively selected for (or against) by breeders. Furthermore, the lack of plasticity and GxE demonstrate the stability of these simple leaf shape descriptor traits, at least in the environments tested.

Contrary to our results, multiple studies have found that leaf dissection--captured here by our measure of circularity--is a plastic trait that responds to changes in temperature. For example, Royer and colleagues (2009, 2012) found that leaves of *Acer rubrum* were more dissected when grown in cooler environments as compared to warmer environments. A similar trend was observed in grapevine (*Vitis* spp.) (Chitwood *et al.*, 2016). However, we found that leaf dissection in sweetpotato is not influenced by the environment. This could reflect that our gardens were not different enough to lead to plastic responses in these two measures of leaf shape. The Ohio garden was consistently warmer (by 2°C on average) and experienced less precipitation than the Michigan garden--the difference between the two gardens was 662.43 mm/month on average throughout the growing season. Although there were environmental differences between gardens, before we conclude that circularity in *I. batatas* is not strongly environmentally responsive, multiple studies in environments that range more widely for temperature will need to be performed.

Comparatively, we found significant variation in aspect ratio due to environment and GxE, explaining 1.93% and 12.95% of the total observed variation in this measure of leaf shape, respectively. This is reflected in the significant alteration of trait values between environments. There was small yet significant differences observed ($P < 0.001$; 95% CI = 0.009-0.03) between gardens, with clones grown in Michigan consistently showing less round, more elliptical leaves than clones grown in the Ohio garden. However, we still found that 38.40% of the variation in the trait was due to accessional variation which was also indicated in the estimated heritability value of the trait ($h^2 = 0.24$). Aspect ratio has been found to be a major source of leaf shape variation in apples and tomatoes with high heritabilities of 0.75 and 0.63, respectively (Chitwood *et al.*, 2013; Migicovsky *et al.*, 2017). In contrast, we found that this important leaf shape trait is globally not as variable in sweetpotato (4.76% PCV), but it still presents a selection potential. The considerable effect of GxE on aspect ratio indicates that this trait has a genetic component that interacts with the environment leading to varied values between environment.

Further, comparing leaf outlines between two environments, we found that although the traits explaining leaf shape variation are homologous between the two environments, these traits vary in the percent of variation they explain. The heritability of EFD symPCs measured in MI and OH

were found to be very high, yet the changes in the amount of variation they explain in their respective environments indicates a strong environmental (and/or GxE) influence on EFD symPCs measured. Although traditional shape descriptors were only slightly controlled by the environment (aspect ratio), we found that the more comprehensive measure of leaf shape can be altered significantly by the environment. This further signifies the importance of measuring leaf shape using methods apart from traditional shape descriptors in multi-environment conditions.

Overall, this work highlights the extensive natural variation in leaf shape within the globally important domesticate *I. batatas*. More broadly, and considering leaf shape analyses from other, mostly domesticated species, leaf shape variation appears to be species specific -- there is no evidence of a shared trait between species that explains the majority of within-species variation. Additionally, we found that most of the variation in the traditional measures of leaf shape appears to be largely controlled by genetic factors in sweetpotato, with a low proportion of variance in leaf shape attributable to environmental differences between gardens. However, when leaf shape was considered more comprehensively and by the use of leaf outlines, we identified a significant influence of the environment, suggesting that studies relying solely on circularity or aspect ratio to describe leaf shape may not capture the extent to which environmental factors can impact leaf development. This multilevel examination highlights the importance of examining morphological variation at the species-level in multiple environments, and using a range of leaf shape phenotypes to comprehensively understand the mechanistic basis (morphological, molecular and environmental) of leaf shape.

Acknowledgements

We are grateful to Robert Jarrett (USDA, Tifton, GA) for providing the sweetpotato accessions. We thank the staff of Matthaei Botanical Garden and Nichols Arboretum (MBGNA, Ann Arbor) for helping us grow and maintain the accessions. We also thank Dan York, Tyler Marrs, Tilottama Roy, Andrew Fox, Jordan Francisco, Yufei Gao, Abby Singletary and Nicholas Tomeo for growing the plants in the field and collecting data. We are also thankful to Daniel H. Chitwood for comments on the manuscript. Funding for this work was provided by the University of Michigan and Ohio University.

References

Abràmoff MD, Magalhães PJ, Ram SJ. 2004. Image processing with imageJ. *Biophotonics International*.

Aguilar-Martínez JA, Uchida N, Townsley B, West DA, Yanez A, Lynn N, Kimura S, Sinha N. 2015. Transcriptional, posttranscriptional, and posttranslational regulation of SHOOT MERISTEMLESS gene expression in Arabidopsis determines gene function in the shoot apex. *Plant physiology* 167: 424–442.

Andersson S. 1991. Geographical variation and genetic analysis of leaf shape in Crepis tectorum (Asteraceae). *Plant systematics and evolution = Entwicklungsgeschichte und Systematik der Pflanzen*.

Andres RJ, Coneva V, Frank MH, Tuttle JR, Samayoa LF, Han S-W, Kaur B, Zhu L, Fang H, Bowman DT, *et al.* 2016. Modifications to a LATE MERISTEM IDENTITY1 gene are responsible for the major leaf shapes of Upland cotton (Gossypium hirsutum L.). *Proceedings of the National Academy of Sciences* 114: E57–E66.

Arber, A. 2010. Water Plants: A Study of Aquatic Angiosperms (Cambridge Library Collection - Botany and Horticulture). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511700675.

Bailey IW, Sinnott EW. 1916. The Climatic Distribution of Certain Types of Angiosperm Leaves. *Source: American Journal of Botany*.

Campitelli BE, Stinchcombe JR. 2013. Natural selection maintains a single-locus leaf shape cline in Ivyleaf morning glory, Ipomoea hederacea. In: Molecular Ecology. 552–564.

Chitwood DH, Kumar R, Headland LR, Ranjan A, Covington MF, Ichihashi Y, Fulop D, Jiménez-Gómez JM, Peng J, Maloof JN, *et al.* 2013. A Quantitative Genetic Basis for Leaf Morphology is Revealed in a Set of Precisely Defined Tomato Introgression Lines. *The Plant cell* 25: 2465–2481.

Chitwood DH, Kumar R, Ranjan A, Pelletier JM, Townsley BT, Ichihashi Y, Martinez CC, Zumstein K, Harada JJ, Maloof JN, *et al.* 2015. Light-Induced Indeterminacy Alters Shade-Avoiding Tomato Leaf Morphology. *Plant physiology* 169: 2030–2047.

Chitwood DH, Otoni WC. 2017. Morphometric analysis of Passiflora leaves: The relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade. *GigaScience*.

Chitwood DH, Ranjan A, Kumar R, Ichihashi Y, Zumstein K, Headland LR, Ostria-Gallardo E, Aguilar-Martinez JA, Bush S, Carriedo L, *et al.* 2014a. Resolving Distinct Genetic Regulators of Tomato Leaf Shape within a Heteroblastic and Ontogenetic Context. *The Plant cell*.

Chitwood DH, Ranjan A, Martinez CC, Headland LR, Thiem T, Kumar R, Covington MF, Hatcher T, Naylor DT, Zimmerman S, *et al.* 2014b. A Modern Ampelography: A Genetic Basis for Leaf Shape and Venation Patterning in Grape. *Plant physiology*.

Chitwood DH, Rundell SM, Li DY, Woodford QL, Yu TT, Lopez JR, Greenblatt D, Kang J, Londo JP. 2016. Climate and developmental plasticity: interannual variability in grapevine leaf morphology. *Plant physiology*.

Chitwood DH, Sinha NR. 2016. Evolutionary and Environmental Forces Sculpting Leaf Development. *Current biology: CB* 26: R297–306.

Conesa MÀ, Mus M, Rosselló JA. 2012. Leaf shape variation and taxonomic boundaries in two sympatric rupicolous species of Helichrysum (Asteraceae: Gnaphalieae), assessed by linear measurements and geometric morphometry: LEAF SHAPE VARIATION IN HELICHRYSUM. *Biological journal of the Linnean Society. Linnean Society of London* 106: 498–513.

Covarrubias-Pazaran G. 2016. Genome-Assisted prediction of quantitative traits using the r package sommer. *PloS one*.

De Vries CA, Ferwerda JD, Flach M. 1967. Choice of food crops in relation to actual and potential production in the tropics. *Netherlands Journal of Agricultural Science* 15:241–248.

Firon N, LaBonte D, Villordon A, Kfir Y, Solis J, Lapis E, Perlman TS, Doron-Faigenboim A, Hetzroni A, Althan L, *et al.* 2013. Transcriptional profiling of sweetpotato (Ipomoea batatas) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. *BMC genomics* 14: 460.

Frank E Harrell Jr, with contributions from Charles Dupont and many others. 2017. Hmisc: Harrell Miscellaneous. R package version 4.0-3. https://CRAN.R-project.org/package=Hmisc.

Glennon KL, Cron GV. 2015. Climate and leaf shape relationships in four Helichrysum species from the Eastern Mountain Region of South Africa. *Evolutionary ecology* 29: 657–678.

Gurevitch J. 1988. Variation in leaf dissection and leaf energy budgets among populations of Achillea from an altitudinal gradient. *American journal of botany* 75: 1298.

Hilu KW. 1983. The Role of Single-Gene Mutations in the Evolution of Flowering Plants. In: Hecht MK, Wallace B, Prance GT, eds. Evolutionary Biology: Volume 16. Boston, MA: Springer US, 97–128.

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

Huaman Z. 1987. Current status on maintenance of sweet potato genetic resources at CIP, in Exploration, maintenance and utilization of sweetpotato genetic resources. CIP: Lima, Peru. 101-120.

Hue SM, Chandran S, Boyce AN. 2012. Variations of leaf and storage roots morphology in Ipomoea batatas L. (Sweet Potato) cultivars. In: Acta Horticulturae. 73–80.

Huff PM, Wilf P, Azumah EJ. 2003. Digital Future for Paleoclimate Estimation from Fossil Leaves? Preliminary Results. *Palaios* 18: 266–274.

Ialongo C. 2016. Understanding the effect size and its measures. *Biochemia medica: casopis Hrvatskoga drustva medicinskih biokemicara / HDMB*.

Ichihashi Y, Aguilar-Martínez JA, Farhi M, Chitwood DH, Kumar R, Millon LV, Peng J, Maloof JN, Sinha NR. 2014. Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. *Proceedings of the National Academy of Sciences of the United States of America* 111: E2616–21.

Iwata H, Ukai Y. 2002. SHAPE: a computer program package for quantitative evaluation of biological shapes based on elliptic Fourier descriptors. *The Journal of heredity*.

Jones C. 1995. Does shade prolong juvenile development? A morphological analysis of leaf shape changes in Cucurbita argyrosperma subsp. sororia (Cucurbitaceae). *American journal of botany*.

Kawamura E, Horiguchi G, Tsukaya H. 2010. Mechanisms of leaf tooth formation in Arabidopsis. *The Plant journal: for cell and molecular biology* 62: 429–441.

Kays SJ, Kays SE. 1998. Sweetpotato chemistry in relation to health. *Proceedings of the Sweet Potato Production System Towards the 21st Century*: 231–272.

Khoury CK, Heider B, Castañeda-Álvarez NP, Achicanoy HA, Sosa CC, Miller RE, Scotland RW, Wood JRI, Rossel G, Eserman LA, *et al.* 2015. Distributions, ex situ conservation priorities, and genetic resource potential of crop wild relatives of sweetpotato [ipomoea batatas (L.) lam., I. series batatas]. *Frontiers in plant science* 6: 1–14.

Kim GT, Shoda K, Tsuge T, Cho KH, Uchimiya H, Yokoyama R, Nishitani K, Tsukaya H. 2002. The ANGUSTIFOLIA gene of Arabidopsis, a plant CtBP gene, regulates leaf-cell expansion, the arrangement of cortical microtubules in leaf cells and expression of a gene involved in cell-wall formation. *The EMBO journal*.

Kimura S, Koenig D, Kang J, Yoong FY, Sinha N. 2008. Natural Variation in Leaf Morphology Results from Mutation of a Novel KNOX Gene. *Current biology: CB* 18: 672–677.

Klein LL, Caito M, Chapnick C, Kitchen C, O'Hanlon R, Chitwood DH, Miller AJ. 2017. Digital Morphometrics of Two North American Grapevines (Vitis: Vitaceae) Quantifies Leaf Variation between Species, within Species, and among Individuals. *Frontiers in plant science*.

Levine TR, Hullett CR. 2002. Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Human Communication Research*.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.

Li M, An H, Angelovici R, Bagaza C, Batushansky A, Clark L, Coneva V, Donoghue MJ, Edwards E, Fajardo D, *et al.* 2018. Topological Data Analysis as a Morphometric Method: Using Persistent Homology to Demarcate a Leaf Morphospace. *Frontiers in plant science*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Liu S, Zainuddin IM, Vanderschuren H, Doughty J, Beeching JR. 2017. RNAi inhibition of feruloyl CoA 6'-hydroxylase reduces scopoletin biosynthesis and post-harvest physiological deterioration in cassava (Manihot esculenta Crantz) storage roots. *Plant molecular biology* 94: 185–195.

Mangiafico SS. 2018. rcompanion: Functions to Support Extension Education Program Evaluation, R package version 2.0.0. https://CRAN.R-project.org/package=rcompanion.

McDonald PG, Fonseca CR, Overton JM, Westoby M. 2003. Leaf-size divergence along rainfall and soil-nutrient gradients: is the method of size reduction common among clades? *Functional ecology* 17: 50–57.

McLellan T. 2000. Geographic variation and plasticity of leaf shape and size in Begonia dregei and B. homonyma (Begoniaceae). *Botanical journal of the Linnean Society. Linnean Society of London*.

Migicovsky Z, Li M, Chitwood DH, Myles S. 2017. Morphometrics Reveals Complex and Heritable Apple Leaf Shapes. *Frontiers in plant science* 8: 2185.

Nakayama H, Nakayama N, Seiki S, Kojima M, Sakakibara H, Sinha N, Kimura S. 2014. Regulation of the KNOX-GA Gene Module Induces Heterophyllic Alteration in North American Lake Cress. *The Plant Cell Online*.

Navarro D. 2013. *Learning statistics with R:A tutorial for psychology students and other beginners*.

Nicotra AB, Leigh A, Boyce CK, Jones CS, Niklas KJ, Royer DL, Tsukaya H. 2011. The evolution and functional significance of leaf shape in the angiosperms. *Functional Plant Biology*.

Peppe DJ, Royer DL, Cariglino B, Oliver SY, Newman S, Leight E, Enikolopov G, Fernandez-Burgos M, Herrera F, Adams JM, *et al.* 2011. Sensitivity of leaf size and shape to climate: global patterns and paleoclimatic applications. *The New phytologist* 190: 724–739.

Piazza P, Bailey CD, Cartolano M, Krieger J, Cao J, Ossowski S, Schneeberger K, He F, De Meaux J, Hall N, *et al.* 2010. Arabidopsis thaliana leaf form evolved via loss of KNOX expression in leaves in association with a selective sweep. *Current biology: CB*.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140.

Rosero A, Granda L, Pérez J-L, Rosero D, Burgos-Paz W, Martínez R, Morelo J, Pastrana I, Burbano E, Morales A. 2019. Morphometric and colourimetric tools to dissect morphological diversity: an application in sweet potato [Ipomoea batatas (L.) Lam.]. *Genetic resources and crop evolution*.

Rowland SD, Zumstein K, Nakayama H, Cheng Z, Flores AM, Chitwood DH, Maloof JN, Sinha NR. 2019. Leaf shape is a predictor of fruit quality and cultivar performance in tomato. *bioRxiv*: 584466.

Royer DL. 2012. Leaf shape responds to temperature but not CO2 in Acer rubrum. *PloS one* 7: e49559.

Royer DL, Meyerson LA, Robertson KM, Adams JM. 2009. Phenotypic plasticity of leaf shape along a temperature gradient in Acer rubrum. *PloS one*.

Royer DL, Wilf P, Janesko DA, Kowalski EA, Dilcher DL. 2005. Correlations of climate and plant ecology to leaf size and shape: potential proxies for the fossil record. *American journal of botany* 92: 1141–1151.

Shi P, Liu M, Yu X, Gielis J, Ratkowsky DA. 2019. Proportional Relationship between Leaf Area and the Product of Leaf Length and Width of Four Types of Special Leaf Shapes. *Forests, Trees and Livelihoods* 10: 178.

Sicard A, Thamm A, Marona C, Lee YW, Wahl V, Stinchcombe JR, Wright SI, Kappel C, Lenhard M. 2014. Repeated evolutionary changes of leaf morphology caused by mutations to a homeobox gene. *Current biology: CB*.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.

Viscosi V. 2015. Geometric morphometrics and leaf phenotypic plasticity: Assessing fluctuating asymmetry and allometry in European white oaks (Quercus). *Botanical journal of the Linnean Society. Linnean Society of London*.

Vlad D, Kierzkowski D, Rast MI, Vuolo F, Dello Ioio R, Galinha C, Gan X, Hajheidari M, Hay A, Smith RS, *et al.* 2014. Leaf shape evolution through duplication, regulatory diversification and loss of a homeobox gene. *Science* 343: 780–783.

Vogel S. 1968. ‘ Sun Leaves’ and‘ Shade Leaves’: Differences in Convective Heat Dissipation. *Ecology* 49: 1203–1204.

Vuolo F, Mentink RA, Hajheidari M, Bailey CD, Filatov DA, Tsiantis M. 2016. Coupled enhancer and coding sequence evolution of a homeobox gene shaped leaf diversity. *Genes and Development* 30: 2370–2376.

Wadl PA, Olukolu BA, Branham SE, Jarret RL, Yencho GC, Jackson DM. 2018. Genetic Diversity and Population Structure of the USDA Sweetpotato (Ipomoea batatas) Germplasm Collections Using GBSpoly. *Frontiers in plant science* 9: 1166.

Wang Y, Ghaffari N, Johnson CD, Braga-Neto UM, Wang H, Chen R, Zhou H. 2011. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC bioinformatics* 12 Suppl 10: S5.

Wolfe J a. 1978. A paleobotanical interpretation of Tertiary climates in the Northern Hemisphere. *American scientist*.

Wyatt R, Antonovics J. 1981. Butterflyweed Re-Revisited : Spatial and Temporal Patterns of Leaf Shape Variation in Asclepias tuberosa. *Evolution1*.

Yeh KW, Chen JC, Lin MI, Chen YM, Lin CY. 1997. Functional activity of sporamin from sweet potato (Ipomoea batatas Lam.): a tuber storage protein with trypsin inhibitory activity. *Plant molecular biology* 33: 565–570.

Zhang K, Wu Z, Tang D, Luo K, Lu H, Liu Y, Dong J, Wang X, Lv C, Wang J, *et al.* 2017. Comparative Transcriptome Analysis Reveals Critical Function of Sucrose Metabolism Related-Enzymes in Starch Accumulation in the Storage Root of Sweet Potato. *Frontiers in plant science* 8: 914.

Zwieniecki MA, Boyce CK, Holbrook NM. 2004. Hydraulic limitations imposed by crown placement determine final size and shape of Quercus rubra L. leaves. *Plant, cell & environment* 27: 357–365

Figures

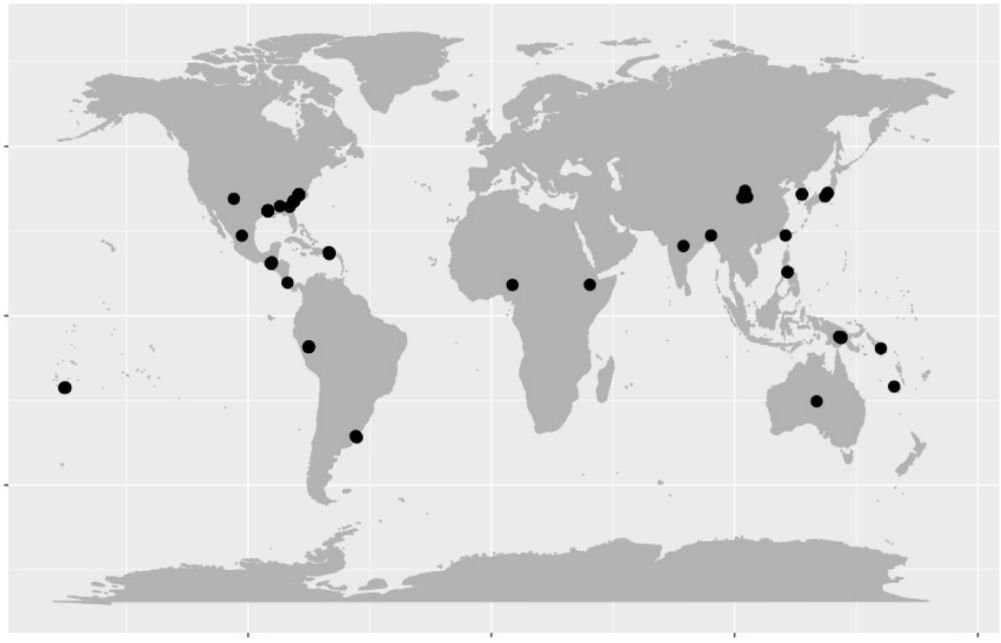

Figure 2-1 Geographic diversity of the 68 chosen sweet potato, *Ipomoea batatas*, accessions. Black dots represent the origin of the chosen samples.

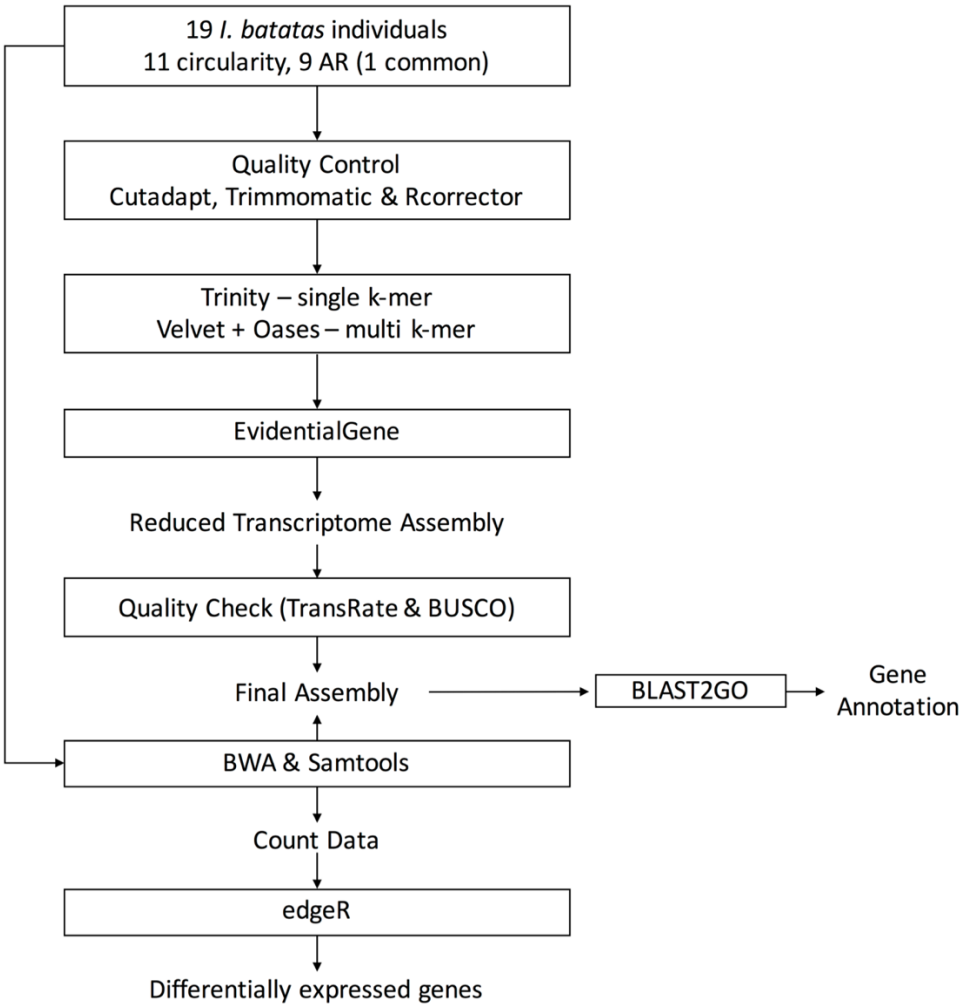

Figure 2-2 Methodology for RNA-seq data processing for differential gene expression.

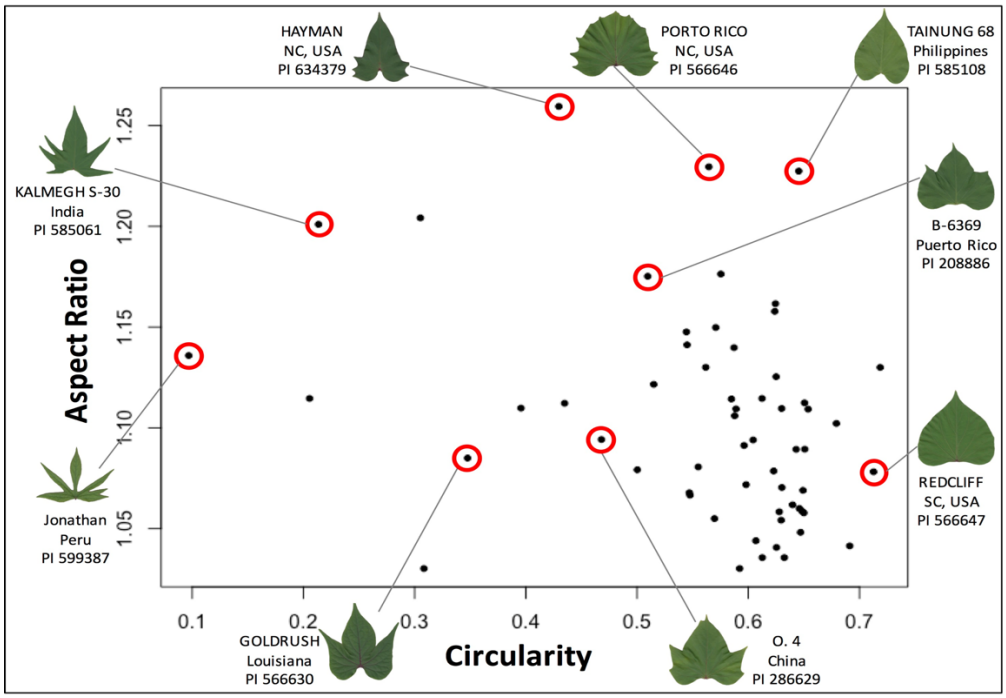

Figure 2-3 Leaf shape variation in 57 diverse, glasshouse-grown, accessions of sweet potato, *Ipomoea batatas*, highlighting exceptionally high morphological variation.

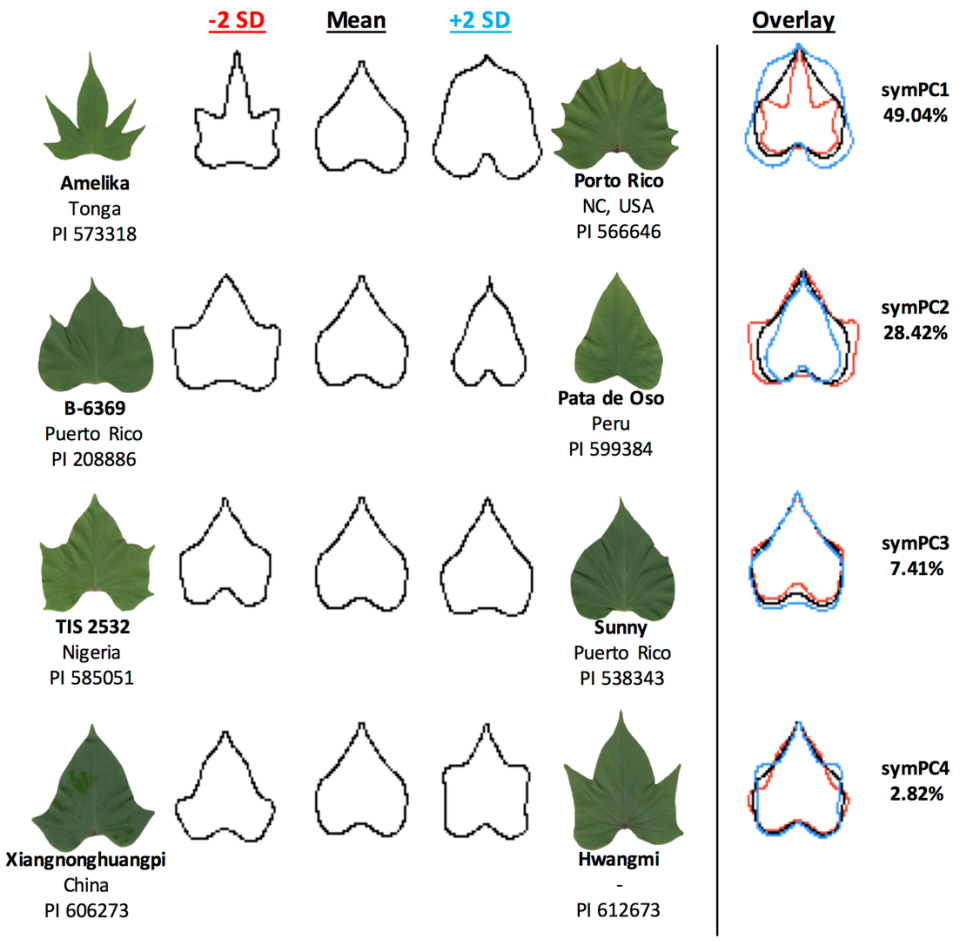

Figure 2-4 Elliptical Fourier Descriptor (EFDs) of symmetrical shape variation in sweet potato, *Ipomoea batatas*.
Contours represent eigenleaves resulting from principal component analysis (PCA) on symmetrical shape (symPC) on EFDs. Shown are the first four symPCs with the per cent variation explained by each; 87.79% of the total variation is explained. _2 SD (red) and +2 SD (blue) represent two units of SD from the mean along the symPC. Representative leaves of accessions with extreme symPC values are shown.

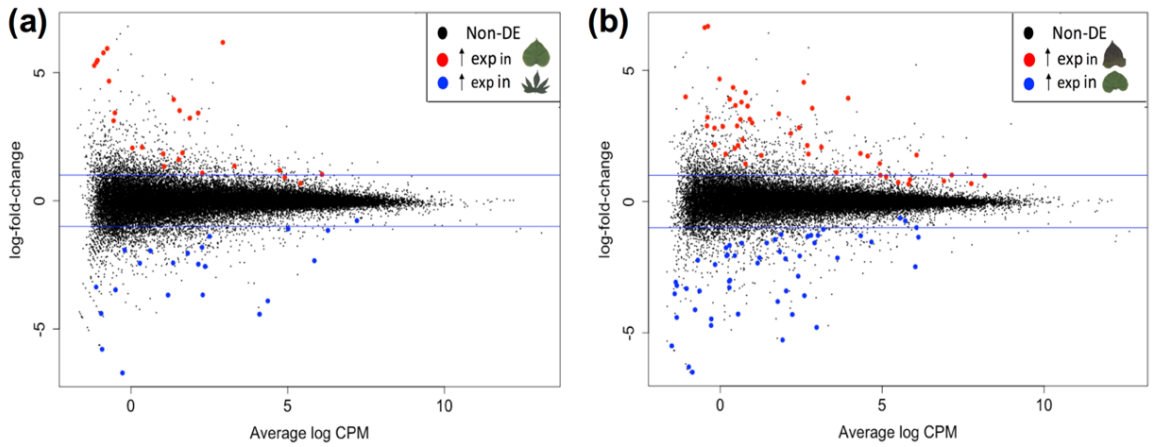

Figure 2-5 Plot of log fold change against log CPM (counts per million) with differentially expressed transcripts highlighted (red and blue dots) for leaf shape in *Ipomoea batatas*. (a) Red and blue dots represent transcripts with higher expression in entire and lobed, respectively, (b) Red and blue dots represent higher expression in high aspect ratio and low aspect ratio individuals, respectively.

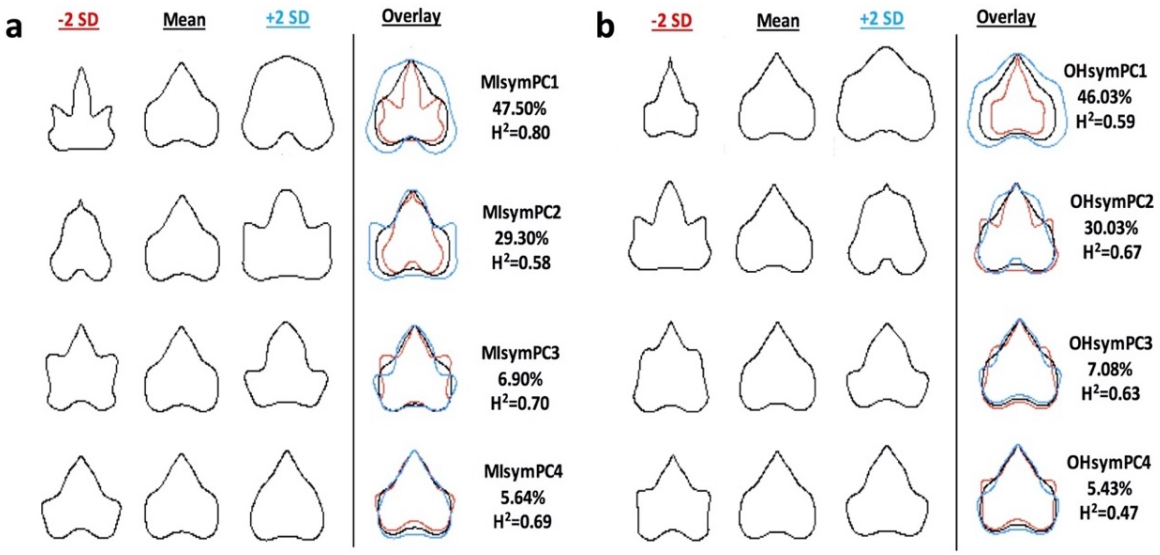

Figure 2-6 Elliptical Fourier Descriptor (EFDs) of symmetrical leaf shape variation among 68 accessions of sweet potato, *Ipomoea batatas*, in the common gardens in Michigan (a) and Ohio (b), respectively.
MIsymPC and OHsymPC represents the Michigan and Ohio symmetrical leaf shape variation, respectively; H^2 represents broad-sense heritability.

Tables

Trait	Range	Mean	SD	PCV (%)
Circularity	0.09-0.71	0.50	0.12	22.61
Aspect Ratio	1.03-1.26	1.10	0.05	4.76
Solidity	0.44-0.95	0.84	0.10	11.85

Table 2-1 Leaf shape trait values across the 57 chosen sweetpotato accessions.
SD represents standard deviation while PCV represents phenotypic coefficient of variation.

Number of transcripts	Min Len (nt)	Max Len (nt)	Number of bases	Mean Len (nt)	ORF percent	n50 (nt)	% reads mapped
33,684	200	16,428	35,769,411	1,062	79.95%	1,608	77%

Table 2-2 Sequence statistics of the reference transcriptome obtained from EvidentialGene pipeline.

Transcript ID	LogFC	FDR	Gene Description
Circularity			
trn22514	5.77	0.003	FAR1-RELATED SEQUENCE
trn27202	2.08	0.021	Exocyst complex component EXO70A1
trn24081	1.33	0.033	Extra-large guanine nucleotide-binding protein
Aspect Ratio			
trn9778	-2.95	0.035	Chalcone Synthase (CHS)
trn24267	2.55	0.00	Feruloyl CoA 6'-hydroxylase 2
trn25053	-1.85	0.021	Protein LIGHT-DEPENDENT SHORT HYPOCOTYLS 10
symPC1			
trn27227	1.56	0.018	Homeobox-leucine zipper HAT22
trn23566	3.54	0.00	FAR1-RELATED SEQUENCE 7
symPC2			
trn27049	-3.09	0.009	Chalcone Synthase
trn28352	-3.52	0.00	Chalcone Synthase
trn9093	-2.21	0.00	Sporamin B

Table 2-3 Candidate genes maintaining variation in leaf traits (circularity, AR and symPCs) identified from the set of differentially expressed transcripts (DETs) in *Ipomoea batatas*.

Variable	df	Circularity			Aspect Ratio			Solidity		
		F	**P**	**η^2 (%)**	***F***	**P**	**η^2 (%)**	***F***	**P**	**η^2 (%)**
Accession	73	18.06	<0.001 ***	73.23	4.22	<0.001 ***	38.40	21.09	<0.001 ***	77.18
Garden	1	3.64	0.056	0.20	15.5	<0.001 ***	1.93	3.37	0.067	0.16
Block	8	3.01	0.002 **	1.33	1.38	0.020	1.38	1.94	0.052	0.70
GxE	69	1.30	0.06	5.01	1.50	0.009 **	12.95	1.30	0.065	0.40
Residuals	364	NA	NA	20.2	NA	NA	45.31	NA	NA	17.56

Table 2-4 ANOVA table of the leaf shape traits model showing significant explanatory variables.
df: degrees of freedom; *F*: value of F-statistic; P: p-value; η^2: eta-squared value.

Env	H^2						
	Circularity	**Aspect Ratio**	**Solidity**	**symPC1**	**symPC2**	**symPC3**	**symPC4**
MI	0.79	0.39	0.82	0.80	0.58	0.70	0.69
OH	0.73	0.26	0.76	0.59	0.67	0.63	0.47

Table 2-5 Broad-sense heritability values for leaf shape traits in differing environments.
* Note: We can not compare heritability values for EFD symPCs between MI and OH because the expression of traits vary between environments, and hence what the symPCs capture differs between the two environments.

Chapter 3

Inter-Chromosomal Linkage Disequilibrium and Linked Fitness Cost Loci Associated With Selection for Herbicide Resistance[2]

Abstract

The adaptation of weedy plants to herbicide is both a significant problem in agriculture and a model for the study of rapid adaptation under regimes of strong selection. Despite recent advances in our understanding of simple genetic changes that lead to resistance, a significant gap remains in our knowledge of resistance controlled by many loci and the evolutionary factors that influence the maintenance of resistance over time. Here, using herbicide resistant populations of the common morning glory (*Ipomoea purpurea*), we perform a multi-level analysis involving whole genome sequencing and assembly, resequencing, and gene expression analysis to both uncover putative loci involved in nontarget-site herbicide resistance (NTSR) and to examine evolutionary forces underlying the maintenance of resistance in natural populations. We found loci involved in herbicide detoxification, and stress sensing to be under selection and confirmed that detoxification is responsible for glyphosate resistance using a functional assay. Furthermore, we found interchromosomal linkage disequilibrium (ILD) to influence NTSR loci found on separate chromosomes thus potentially mediating resistance through generations. Additionally, by combining the selection screen, differential expression, and LD analysis, we identified fitness cost loci that are strongly linked to resistance alleles, indicating the role of genetic hitchhiking in

[2] This chapter is *in revision* at PNAS as Gupta S, Harkess A, Soble A, Van Etten M, Leebens-Mack J, Baucom RS. Inter-chromosomal linkage disequilibrium and linked fitness cost loci associated with selection for herbicide resistance.

maintaining the cost. Overall, our work suggests that NTSR glyphosate resistance in *I. purpurea* is conferred by multiple genes which are potentially maintained through generations *via* ILD and that the fitness cost associated with resistance in this species is a by-product of genetic-hitchhiking.

Introduction

Pesticide and herbicide use has reshaped ecological networks and induced strong selective pressures in the anthropogenic era. How species may adapt to strong selection is a fundamental question in evolution with great importance to the control of pesticide resistant organisms. A striking feature of pesticide resistance evolution is that there are a number of different genetic solutions that can lead to resistance (Liu, 2015; Hawkins *et al.*, 2018). In herbicide resistant plants, for example, resistance can be due to single gene mutations, often found in the herbicide's target protein (target-site resistance, TSR), or due to changes in multiple genes, often underlying nontarget herbicide resistance (NTSR) mechanisms (Powles & Yu, 2010; Mithila & Godar, 2013). A growing body of work has produced a better understanding of resistance controlled by single genes across a variety of species (Jasieniuk *et al.*, 1996; Powles & Yu, 2010; Murphy & Tranel, 2019). However, we currently lack a deep understanding of both the genetic basis and evolutionary potential of nontarget site resistance mechanisms genome wide (Délye, 2013; Baucom, 2019; Beckie, 2020; Leon *et al.*, 2021).

This is due in part to the broad nature of nontarget site herbicide resistance mechanisms more generally. NTSR can be caused by reduced herbicide uptake or penetration, altered translocation or sequestration, metabolism (Vemanna *et al.*, 2017; Pan *et al.*, 2019) and/or herbicide detoxification (Délye *et al.*, 2013; Gaines *et al.*, 2019, 2020) -- mechanisms that likely rely on a complex genetic basis (Busi *et al.*, 2013; Scarabel *et al.*, 2015; Kreiner *et al.*, 2018; Baucom, 2019). For example, detoxification is hypothesized to involve three steps -- uptake of the herbicide by phosphate transporters, chemical modification (*i.e.* the addition of an OH and sugar group by cytochrome P450s and glycosyltransferases, respectively), and transport to vacuoles by ABC transporters and other sugar transporters where the molecule is stored and/or inactivated

(Yuan *et al.*, 2007; Gaines *et al.*, 2020). Combinations of these gene families have been shown to be under selection for resistance to a suite of herbicides (Liu *et al.*, 2018; Yanniccari *et al.*, 2020; Dimaano & Iwakami, 2021; Huang *et al.*, 2021; Pan *et al.*, 2021; Amrhein & Martinoia, 2021). While some investigations have pinpointed a single gene conferring NTSR (Cummins *et al.*, 2013; Pandian *et al.*, 2021), gene expression surveys or whole-genome re-sequencing assays in a small handful of resistant weeds are beginning to shed light on the complexity of nontarget resistance mechanisms (Yuan *et al.*, 2007; Van Etten *et al.*, 2020; Kreiner *et al.*, 2020). In both *Amaranthus tuberculatus* and *Ipomoea purpurea*, a number of different loci found across the genome -- whether structural, regulatory, or both -- exhibit signs of selection and are thus putatively involved in resistance (Yuan *et al.*, 2007; Van Etten *et al.*, 2020; Kreiner *et al.*, 2020). Because we lack a deep understanding of the genetic basis of NTSR in most weeds, however, we lack a firm grasp on the underlying forces that influence the maintenance of resistance in natural populations, such as the prevalence of alleles that may contribute to fitness costs of resistance or the presence of interchromosomal linkage disequilibrium (ILD). The presence of ILD between unlinked regions of the genome would implicate the potential for correlational selection and/or coadaptation between alleles underlying either resistance or its cost.

Ipomoea purpurea is a common agricultural weed in the Southeast and Midwest United States. Populations of this species, which have consistently been exposed to glyphosate based herbicides since the late 1990s (Kuester *et al.*, 2015, 2016), exhibit varying levels of herbicide resistance, with some populations exhibiting low and others high survival post-herbicide application (Kuester *et al.*, 2015, 2016)). There is a fitness cost associated with this resistance: resistant populations show lower germination and deteriorated seed quality compared to susceptible populations (Van Etten *et al.*, 2016). Further, populations from the south and midwest show evidence of genetic admixture, with both microsatellite and SNP data showing low genetic differentiation for a mixed-mating plant species (F_{ST} = 0.11-0.14; (Alvarado-Serrano *et al.*, 2019)) and recent genetic connectivity (Alvarado-Serrano *et al.*, 2019). RADseq and exome sequencing has identified regions of the genome under selection and thus associated with herbicide resistance. These regions are enriched for cytochrome P450s, glycosyltransferases, and ABC transporter genes, indicating a likely role of herbicide detoxification in conferring resistance (Van Etten *et al.*, 2020). Despite evidence that detoxification underlies resistance in

this species and suggestions that loci found on different chromosomes contribute to resistance, previous work relied on low-coverage RADseq sequencing without the benefit of a contiguously assembled genome. Thus, loci that may contribute to NTSR or its cost were likely missed (Lowry *et al.*, 2017), meaning that we lack a thorough understanding of NTSR, the genomic context of NTSR alleles, and the potential for relationships among NTSR alleles in this species -- all crucial to understanding the evolution of resistance more broadly.

Here, we implemented a genome-wide selection screen using whole-genome resequencing of natural populations along with a gene expression survey to characterize the genetic architecture of glyphosate resistance and its cost in *Ipomoea purpurea*. We complemented our survey with a functional assay to test the potential that resistant *I. purpurea* individuals detoxify the herbicide more rapidly. Given previous evidence that multiple loci likely contribute to herbicide resistance in this species, and evidence of fitness cost of resistance, we made two main predictions regarding genome-wide patterns of selection associated with resistance in *I. purpurea*. First, we expected that regions of the genome showing high differentiation and marks of selection when comparing herbicide resistant and susceptible individuals would contain loci with strong functional links to either herbicide resistance or its cost. Second, we anticipated that linkage disequilibrium, the non-random association of alleles at different loci, should be evident among regions of the genome housing resistance loci. Although inter-chromosomal linkage disequilibrium has been identified in other systems assessing ecologically relevant traits such as mate choice and coloration (Petkov *et al.*, 2005; Long *et al.*, 2013; Hench *et al.*, 2019) and recently has even been identified in target-site resistance (Kreiner *et al.*, 2021) it is unknown if loci underlying herbicide resistance that are found across chromosomes exhibit long-distance or inter-chromosomal linkage disequilibrium, as would be expected if adaptation to herbicide is facilitated by multilocus genotypes favored by selection (*i.e.*, coadapted gene complexes) (Wallace, 1953; Dobzhansky, 1971; Schluter, 2000).

Results

A chromosome-scale genome assembly for common morning glory

We assembled a reference *I. purpurea* genome to test these hypotheses, generating the first genome sequence for this common and noxious weed. We generated a total of 48 gigabases of

PacBio Sequel whole genome shotgun data (*Appendix B*, Figure S3-1a). Based on a flow cytometry genome size (Benaroya Institute, Seattle, WA), this amounts to roughly 59X genome coverage for an estimated haploid genome size of 814 Mb. We used 34.79 gigabases of trimmed and self-corrected reads for assembly, scaffolding, and polishing, which produced a 602 Mb assembly in 402 scaffolds (434 contigs), with a scaffold N50 of 5.77 Mb.

We performed pseudomolecule scaffolding with Phase Genomics Hi-C map, which collapsed the assembly into the expected 15 haploid chromosomes (*Appendix B*, Figure S3-1b). We renamed and oriented chromosomes according to a high degree of synteny with the related *Ipomoea nil* genome (Hoshino *et al.*, 2016) (*Appendix B*, Figure S3-1c). No misjoins were identified and broken based on the Hi-C linkage data. BUSCO scores on the unannotated assembly show 97.5% completeness against the Viridiplantae odb10 gene set (*Appendix B*, Figure S3-1d).

Approximately 63% of the assembly was masked as repetitive DNA, with a significant proportion of recently-expanded Long Terminal Repeat (LTR) retrotransposons (*Appendix B*, Figure S3-1e). Given the high degree of synteny with *I. nil* genome, the discrepancy between the flow cytometry genome size (814 Mb) and the assembled size (602 Mb) is likely due to young retrotransposon proliferation. We annotated 53,973 genes by combining ab initio gene predictions and RNA sequencing data from leaf tissue. The assembly shows a high degree of synteny with several genomes in the Convolvulaceae family, including *I. nil, I. trifida, and I. triloba* (*Appendix B*, Figure S3-2).

Detecting loci under selection

Whole-genome analysis of 69 individuals (from three resistant and four susceptible populations, Figure 3-1a) identified genes under selection for herbicide resistance. We did this using a dual approach -- first, we used bayenv2 (Günther & Coop, 2013) followed by the Md-rank-*P* composite test of selection (Lotterhos *et al.*, 2017), taking into account explicitly and implicitly the population structure, respectively. bayenv2 identified 2,016 SNP outliers (*Appendix B*, Table S3-1). Within 5kb flanking regions of these outliers, 1,908 genes were present, of which 1,485 genes could be functionally annotated (*Appendix B*, Table S3-2). The Md-rank-*P* approach -- a composite test of selection that incorporates nucleotide diversity, Tajima's D, Fay and Wu's H,

and H12 -- identified fifteen regions (total 4.47Mb) that exhibited signs of selection (*Appendix B*, Table S3-3). The regions under selection housed 358 genes, 202 of which could be functionally annotated (*Appendix B*, Table S3-4). There were 128 genes (*Appendix B*, Table S3-5) found in common between the two selection scans, and these genes were located on six different chromosomes. These genes were broadly involved in the process of detoxification, environmental sensing, and stress signaling and response (Table 3-1).

The strongest signal of selection we uncovered using both bayenv2 and the M-rank-P selection scans was found on chromosome 10 (Figure 3-2). Within this region, we identified 13 genes -- 6 copies of cytochrome P450 genes (CYP) and 7 copies of glycosyltransferases -- both of which are gene families previously implicated in herbicide detoxification. The six cytochrome P450s belong to the 76A family (three CYP76A1 and three CYP76A2) and were present in tandem within 53kb, which interestingly also contained two additional copies of CYP76. Of the eight tandem CYPs, four copies of the cytochrome P450s exhibited multiple non-synonymous mutations that were almost fixed in the resistant individuals (resistant allele frequency = 0.95; susceptible allele frequency = 0.31; *Appendix B*, Table S3-6). Further, two of the eight cytochrome P450s (CYP76A2) in this block exhibited either a premature stop codon and/or a splice site donor variant (G->C) in the first intron (allele1 susceptible frequency = 0.68, resistant frequency = 0.05) in the majority of the susceptible individuals. The seven glycosyltransferases were also found in a tandem block; one glycosyltransferase copy showed the loss of a stop codon (susceptible frequency = 0.64, resistance frequency = 0.05), whereas the other glycosyltransferases exhibited multiple non-synonymous mutations close to fixation in the resistant individuals (resistant frequency = 0.05, susceptible frequency = 0.60; *Appendix B*, Figure S3-3; Table S3-6). Additionally, the block of glycosyltransferases in this region showed evidence of a hard sweep (glycosyltransferases H12 = 0.87).

On chromosome 11, we found a 6.4-7.3Mb region to show signs of selection with 33 genes identified *via* both bayenv2 and Md-rank-*P* approaches. Among these genes, were five copies of a phosphate transporter gene (*PHO1*), all containing almost fixed non-synonymous mutations in resistant individuals (resistant frequency = 0.99; *SI Dataset*, S6). This region also contained a copy of a cytochrome P450 gene (CYP736A12 family, *SI Dataset*, S2). Interestingly, multiple stress response genes were also present in this region: NAC domain-containing protein 92

(NAC92) involved in salt stress signaling (Franzoni *et al.*, 2019), LOB domain-containing protein 41 (LBD41) involved in hypoxia stress (Giuntoli *et al.*, 2017), and five copies of leaf rust 10 disease-resistance locus receptor-like protein kinase (LRK10L) which are involved in abiotic stress signaling (Lim *et al.*, 2015).

Another region of note showing strong signals of selection was found on chromosome 6 (average Md-rank-*P* = 6.55; average G_{ST} = 0.782, Figure 3-1b), with evidence of strong differentiation continuing further upstream and downstream (40.23Mb - 40.81Mb; mean G_{ST} = 0.727). Within the extended downstream region, we found ethylene responsive transcription factor (*ERF4*) and multiple copies of serine/threonine kinases, genes that are involved in the signal transduction in response to various biotic and abiotic stresses (Hardie, 1999; Lee & Kim, 2011; Thirugnanasambantham *et al.*, 2015; Liu *et al.*, 2017; Ma & Li, 2018). Within this region, we also identified loci that are likely related to the cost of resistance in this species, expanded upon further in 'Signs of selection on potential cost loci' below.

Among other genes exhibiting signs of selection, we found multiple environmental stress response genes (*SI Dataset*, S5): copies of serine/threonine-protein kinases (CTR1, At5g23170), involved in environmental sensing and stress signaling, the GT-3B transcription factor which is responsible for inducing response to salt (Park *et al.*, 2004), AP2-like ethylene-responsive transcription factor PLT1, AT-hook motif nuclear-localized protein 24 (AHL24) (Wang *et al.*, 2021), and two homologs of A20/AN1 zinc-finger containing stress associated protein (SAP), genes involved in response to environmental stress (Giri *et al.*, 2011).

Overall, our selection screens using a WGS resequencing approach identified highly differentiated regions under selection, with these regions containing genes involved in herbicide detoxification (cytochromeP450s, glycosyltransferases, and phosphate transporters), environmental sensing (serine/threonine kinases), and stress response genes (SAPs, PLT1, AHL24, GT-3B transcription factor). Thus, our study expands on our previous work which found detoxification genes to be under selection (Van Etten *et al.*, 2020) by providing strong evidence that glyphosate resistance in *I. purpurea* is a polygenic NTSR mechanism likely involving herbicide detoxification, and response to environmental stimuli and stress.

Gene expression differences implicate herbicide detoxification

We compared gene expression between herbicide treated resistant and susceptible plants and found support for the idea that herbicide detoxification, plant signaling, and stress response underlies resistance. Of the 250 differentially expressed genes (111 upregulated and 139 downregulated; *Appendix B*, Table S3-7), we found cytochrome P450s, glycosyltransferases, ABC transporters, and glutathione S-transferase genes (Figure 3-3a) to be differentially regulated between resistant and susceptible plants. Two copies of the cytochrome P450 family CYP82D7 were significantly upregulated in the resistant individuals (logFC: 2.05 and 1.35), along with two copies of UDP-glycosyltransferases (UGT87A2, logFC: 5.02 and UGT88B1, logFC: 1.65) and a glutathione S-transferase (GST, logFC:2.93). We additionally found a cytochrome P450 (CYP82C4, logFC:-2.15), a glycosyltransferase (UGT89B2, logFC:-2.09), and an ABC transporter (ABCC3, logFC:-8.81) to be downregulated in the resistant individuals.

We likewise uncovered differences in the expression of genes associated with environmental stress responses. Among notable genes were ethylene responsive transcription factors (ERF003, ERF107, TINY (Walper *et al.*, 2016; Xie *et al.*, 2019)), serine/threonine kinase BLUS1 (Takemiya *et al.*, 2013), E3 Ubiquitin protein ligase PUB23 (Lee & Kim, 2011), NAC domain-containing protein 72 (Liu *et al.*, 2016), and WRKY transcription factors (WRKY4, WRKY31, WRKY75 (Lai *et al.*, 2008; Jiang *et al.*, 2017; Zhao *et al.*, 2019)) (Figure 3-3a). Homologs of these genes (ERF4, PLT1, CTR1, HT1, B120, NAC92, NAC25; *Appendix B*, Table S3-5) were also under selection when comparing herbicide resistant and susceptible populations. In comparison to the genes under selection, two genes were also differentially regulated in the treatment group -- an anionic peroxidase and NFD6 (discussed in the '*Signs of selection on potential cost loci*' section).

In the control (non-herbicide) environment, we found 623 differentially expressed genes when comparing resistant and susceptible individuals (319 upregulated and 304 downregulated; *Appendix B*, Table S3-8). We identified multiple copies of cytochrome P450s, glycosyltransferases, and ABC transporters that were differentially expressed, indicating that glyphosate resistance through detoxification could be is constitutive, and not induced, in this species. Interestingly, the specific cytochrome P450s and glycosyltransferase genes that

exhibited signs of selection from our whole-genome scan were not the same as those that exhibited differential expression, which could be due to the non-simultaneous nature of the gene transcription response to glyphosate; members of the same family have been shown to be differentially regulated at different time points after glyphosate exposure (Piasecki *et al.*, 2019). Additionally, this lack of overlap could also be due to the detection of false positives or could represent a transcriptional sampling stage caveat.

Functional assay supports herbicide detoxification as a mechanism of resistance

We performed an assay to determine if the functional mechanism of resistance in *I. purpurea* was herbicide detoxification (following (Christopher *et al.*, 1994; Preston *et al.*, 1996; Yu & Powles, 2014; Yanniccari *et al.*, 2020; Pandian *et al.*, 2021)). We applied malathion, a pesticide that inhibits some cytochrome P450s, to multiple resistant and susceptible *I. purpurea* individuals from the same populations used in the WGS re-sequencing and gene expression studies. The expectation that malathion would act to inhibit *I. purpurea* cytochrome P450s was met; we found a significant overall treatment effect (F-value = 59.33, df = 3, $p < 0.0001$; Figure 3-3b) with individuals treated with both glyphosate and malathion showing lower biomass compared to individuals treated with either malathion, glyphosate, or untreated controls (Figure 3-3b; *Appendix B*, Table S3-9).

As expected, resistant individuals showed significantly greater biomass compared to the susceptible individuals in the presence of glyphosate (F-value = 4.81, df = 1, P-value = 0.03; Figure 3-3c). However, the biomass of resistant individuals in the presence of both malathion and glyphosate was significantly lower than that of resistant individuals treated only with glyphosate (resistant plants, malathion+glyphosate *vs* glyphosate: t = 3.65, df = 78, p-value = 0.001), indicating that the presence of malathion reduces the resistance response. In fact, the presence of both malathion and glyphosate led to similar (and low) remaining biomass of both resistant and susceptible individuals (malathion+glyphosate treatment: resistant vs susceptible plants: t = 0.15, df = 32, p-value = 0.88). This shows that the presence of a cytochrome P450 inhibitor lowers the level of glyphosate resistance in *I. purpurea* plants, supporting the idea that modification to the detoxification pathway underlies glyphosate resistance in this species.

Role of long-distance and interchromosomal linkage disequilibrium in maintaining NTSR alleles

Our whole-genome scan identified regions under selection containing genes involved in environmental sensing, stress responses, and herbicide detoxification. This broad scan implicates a polygenic basis of resistance in *I. purpurea* and shows that multiple regions of the genome likely contribute to resistance. We thus sought to determine if there was evidence of linkage disequilibrium between these regions, which would potentially suggest either epistatic interactions among alleles or the inheritance of coadapted gene complexes (Wallace, 1953; Nei & Li, 1973). We calculated a measure of linkage disequilibrium (r^2) between long-distance and interchromosomal SNPs that showed the most extreme level of differentiation and selection (98th percentile, $G_{ST} > 0.39$) -- regions on Chromosome 6 (two regions, hereon referred to as 6.1 and 6.2), 10, and 11, and compared it to the whole-genome measure. We found that the four highly differentiated regions also under selection showed islands of elevated interchromosomal linkage disequilibrium (*Appendix B*, Table S3-10) in a backdrop of nearly zero genome-wide ILD (background interchromosomal r^2 mean = 0.00096; Figure 3-4). Additionally, the four regions with high differentiation between resistant and susceptible populations that also exhibited signs of selection showed higher linkage with one another (99th percentile ILD = 0.22 ± 0.0002 SE) compared to five randomly pulled, but highly differentiated regions between resistant and susceptible populations that are not under selection (99th percentile ILD = 0.11 ± 0.0001 SE). The region under selection on Chr10 exhibited the strongest linkage to other chromosomal regions under selection (99% ILD Chr6.1-Chr10 = 0.257, Chr6.2-Chr10 = 0.22, Chr11-Chr10 = 0.17).

Interestingly, the highest r^2 values (within the top 1 percentile) within these regions was observed for putative resistance genes identified above. For instance, multiple glycosyltransferases and cytochrome P450s under selection on Chr10 showed high ILD with SNPs on Chr11 (*Appendix B*, Table S3-11). Multiple cytochrome P450 genes (*CYP76A2*) on Chr10 showed a high value of ILD with an uncharacterized protein and a region upstream of GT-3B on Chr6.1 (range of r^2 = 0.256-0.278, *SI Dataset*, S10) as well as the intergenic region between the transcription factors *SPL1* and *DOF1.4*, both of which are responsible for plant growth and development, on Chr6.2 (range of r^2 = 0.249-0.288), perhaps indicating the co-

adaptation of these regions on Chr6 and on Chr10. Thus, the identified resistance alleles within these four highly differentiated regions show signs of linkage and perhaps evidence of co-adaptation.

Local regions of strong long-distance linkage disequilibrium and ILD within species might be aided by demographic processes like population structure (Nei & Li, 1973; Wilson & Goldstein, 2000), genetic drift (Schaper *et al.*, 2012), or could be due to other processes like selection (Hohenlohe *et al.*, 2012; Hench *et al.*, 2019). Furthermore, coadaptation among loci wherein adaptive alleles at two independent loci will be inherited together can generate linkage disequilibrium (Hohenlohe *et al.*, 2012; Sohail *et al.*, 2017; Behrouzi & Wit, 2017; Hench *et al.*, 2019). Given that our sampling design included multiple resistant and susceptible populations from varied locations, low population differentiation among populations, and evidence of recent migration between them (*Appendix B*, Figure S3-4), it is unlikely that the observed ILD is due entirely to demographic processes. Moreover, we observed the strongest ILD between regions under selection harboring resistance associated genes, indicating the potential role of selection in maintaining the observed ILD. Thus, our finding suggests that the highly differentiated regions under selection containing candidate loci for glyphosate resistance are functionally linked and inherited together.

Signs of selection on potential cost loci

Our scan of regions associated with herbicide resistance, paired with a transcriptome survey, identified potential alleles with strong functional connections to the previously identified fitness cost in resistant *I. purpurea*. Cost of resistance can either be due to the same loci conferring resistance and incurring cost, or could also arise from different loci due to the presence of local linkage disequilibrium and ILD (Bergelson & Purrington, 1996; Zhong *et al.*, 2005). Here we tested the latter. We found alternate alleles close to fixation in each population type within the 585kb highly differentiated region on chromosome 6 (40.23Mb - 40.81Mb; mean G_{ST} = 0.727, Figure 3-5a). This region contained the nuclear fission defective 6 (*NFD6*) and NAC transcription factor 25 (*NAC25*) genes, both of which function in seed development. *NAC25,* a gene that is required for normal seed development and morphology (Kunieda *et al.*, 2008), exhibited two missense variants in the resistant individuals (mutant allele resistant frequency =

0.91, susceptible frequency = 0.23). Additionally, *NFD6*, a protein required for nuclear fusion in the embryo sac during the production of the female gametophyte (Portereiko *et al.*, 2006), contained six missense variants in the resistant individuals (mutant allele resistant frequency = 0.88, susceptible frequency = 0.21). The resistant haplotype of this gene also contained 10 SNPs in the promoter region which could potentially alter its expression. Indeed, we found this protein to be downregulated in the presence of the herbicide, with a log-fold change of -3.52 in resistant individuals as compared to the susceptible individuals (Figure 3-5b). Thus, our data suggest that these genes may be responsible for the lower and abnormal germination leading to the observed fitness cost in this species (Van Etten *et al.*, 2016).

Interestingly, this highly differentiated region containing these seed development genes is strongly linked to other regions under selection that harbor resistance alleles (99% ILD value = 0.22; Figure 3-4), indicating the potential role of linkage disequilibrium in maintaining the cost. More specifically, *NFD6* is within 83 kb from, and thus physically linked to ($r^2 = 0.70$), the potential regulatory region on Chr6.2 that exhibits interchromosomal long distance linkage disequilibrium with the CYP76A2 gene on Chr10 (ILD = 0.27). Further, the NAC25 gene is found in close proximity to serine/threonine kinases on Chr6.2 (i.e., 82 kb away), indicating another potential gene involved in the cost phenotype is in physical linkage ($r^2 = 0.67$) with potential NTSR loci.

In addition to the potential cost loci identified from the WGS screen, we also performed gene expression analyses comparing resistant and susceptible individuals. Cost loci are expected to be differentially expressed in the absence of herbicide (*i.e.* the environment in which fitness costs are assessed), but may or may not be differentially expressed in the presence of herbicide, depending on whether the gene expression is constitutive or herbicide dependent. We identified five differentially expressed genes, in the absence of herbicide, that play a role in fertilization and seed maturation and are thus potentially related to the cost (*Appendix B*, Table S3-8). Of special interest, the bud-site selection protein 31 (*BUD31*) was found to be highly upregulated (logFC = 7.22) in resistant plants in the absence of herbicide, whereas its homologue *BUD13* was highly significantly downregulated in resistant individuals in the presence of herbicide

(logFC = -11.39). *BUD13* is involved in pre-mRNA splicing in embryos and is critical for early embryo development (Xiong *et al.*, 2019).

In the control environment, two genes downregulated in the resistant individuals -- *NFD4* (logFC = -3.61) and Agamous-like MADS-box protein AGL61 (*AGL61*; logFC = -4.98) -- are involved in megagametogenesis. The *NFD4* gene, like *NFD6*, is responsible for ovule polar nuclei fusion during female karyogamy (Portereiko *et al.*, 2006), whereas *AGL61* is required for central cell development and differentiation (Steffen *et al.*, 2008). A loss of function mutation in *AGL61* has been shown to cause abnormal morphology and over 50% seed abortion upon fertilization in Arabidopsis (Steffen *et al.*, 2008). We also identified a callose synthase 2 (*CALS2*) to be strongly downregulated in resistant individuals (logFC = -8.72); another member of the callose synthase family (*CALS5*) has been shown to be responsible for pollen viability (Dong *et al.*, 2005). Finally, we also found that E3 ubiquitin-protein ligase BRE1 (*HUB1*), a protein involved in seed germination, was strongly downregulated among the resistant individuals (logFC = -8.27). *HUB1* has been shown to control chromatin remodeling during seed development and leads to alterations in seed dormancy (Liu *et al.*, 2011).

Interestingly, three of these five candidate genes (*AGL61*, *CALS2*, *HUB1*), and homologues of other two (*BUD13* and *NFD6)* were also significantly downregulated in the resistant populations in the presence of herbicide (*Appendix B*, Table S3-7). These candidate cost genes are all essential for plant reproduction and are highly downregulated (except *BUD31*) in the resistant population in both the absence and presence of the herbicide, and thus could potentially explain the phenotypic costs of glyphosate resistance in *I. purpurea* seen by Van Etten and colleagues (Van Etten *et al.*, 2016).

Discussion

While there is an increasing appreciation for the role of nontarget site mechanisms underlying herbicide resistance in agricultural weeds (Délye *et al.*, 2011; Ghanizadeh & Harrington, 2017; Jugulam & Shyam, 2019; Gaines *et al.*, 2020), there are strikingly few comprehensive whole-genome assays of resistant weeds suggesting that the entirety of the NTSR response is rarely

captured. Our study using a sequenced and assembled genome, whole genome resequencing of natural populations, and a gene expression survey offers a unique opportunity to identify loci associated with NTSR and to further investigate the evolutionary forces that underlie the maintenance of resistance alleles in natural populations.

Our results show detoxification underlies resistance in *I. purpurea*. Detoxification is hypothesized to be enriched in phosphate transporters (uptake), cytochrome P450s and glycosyltransferases (chemical modification), and ABC and sugar transporters (transport to vacuoles) (Yuan *et al.*, 2007; Gaines *et al.*, 2020). We found evidence of selection on genes involved in this pathway. We also found evidence of selection (and in some cases, differential expression) of genes involved in plant signaling and environmental stress (i.e., serine/threonine kinases, ERFs, SAPs, PLT1, AHL24, GT-3B). Our results thus expand what we currently know about the detoxification NTSR mechanism in this species to include plant signaling and stress responses, both of which are either hypothesized (Délye, 2013) or shown to be involved in herbicide resistance (Radwan, 2012; Duhoux & Délye, 2013; Dyer, 2018; Vega *et al.*, 2020). While we do not currently have functional genomics resources for this species, our study using a cytochrome P450 inhibitor supports the hypothesis that resistant *I. purpurea* individuals have the ability to detoxify the herbicide. Accordingly, we think that the genes involved in detoxification might be the major effect genes, while stress signaling and response genes might contribute minor effects on the resistance phenotype in this species. The next step in understanding resistance in *I. purpurea* involves determining the contribution of each of the candidate loci under selection (and/or showing differential regulation) to both resistance and its associated cost. With future development of genome editing protocols for *I. purpurea*, we will be able to experimentally test the function of loci hypothesized to be contributing to herbicide resistance.

Due to the involvement of multiple genes involved in the herbicide detoxification pathway, and evidence for selection on regions of the genome found on separate chromosomes, we hypothesized that multiple loci would show evidence of ILD, perhaps indicating co-inheritance. Our results support this hypothesis. Foremost, in contrast to low background ILD, long-distance linkage disequilibrium and ILD were high among intervals under selection, and consistently differentiated between the resistant and susceptible types across multiple populations. The

strongest linkage was observed between putative resistance genes that exhibited signs of selection. This linkage could quickly become very steep in the presence of co-adaptation among loci (Lewontin & Kojima, 1960), as would be the case if genes underlying NTSR worked in concert to produce the resistance phenotype. Indeed, we found high ILD values between regulatory regions and resistance alleles, and between intervals harboring genes involved in the same molecular pathways (e.g. detoxification, and stress signaling and response).

Although linkage should become decoupled over time due to recombination and gene flow, the ongoing selection for herbicide resistance could slow down this decoupling between these functionally interacting genes (Nei, 1967). Even given gene flow between these populations (Alvarado-Serrano *et al.*, 2019), co-adaptation could lead to the fixation of the resistance alleles given strong selection (Takahasi, 2007) or weaker recombination rate (Takahasi & Tajima, 2005). Thus, long-distance linkage disequilibrium and ILD aided with co-adaptation could act to maintain resistance through generations in natural populations.

One evolutionary force that should counteract the continued evolution of resistance is the potential for fitness costs of resistance, either due to the pleiotropic effects of resistance alleles themselves or due to negative fitness effects of loci that are linked to resistance loci. While costs are central to theories of resistance evolution (Simms & Rausher, 1987; Stahl *et al.*, 1999; Baucom & Mauricio, 2004; Vila-Aiub *et al.*, 2009), there are currently no examples, to our knowledge, in which the loci underlying fitness costs of nontarget site resistance have been identified. Our results suggest putative candidate loci associated with the previously identified cost of glyphosate resistance. Specifically, we found a highly differentiated region on Chr6 that exhibited alternate alleles in resistant and susceptible populations, and found this region to contain loci required for normal seed development and maturation (*NAC25, NFD6*). One of these genes, *NFD6*, was differentially regulated in the resistant individuals, further supporting its role in the low seed quality, and thus fitness cost, that we have previously described (Van Etten *et al.*, 2016). Further functional studies are needed to validate the role of these genes in incurring the fitness cost of resistance.

Additionally, our results strongly suggest genetic hitchhiking may act to maintain the cost in this species. Both *NAC25* and *NFD6* are physically linked on chromosome 6 to the regions under selection containing serine-threonine kinase genes and a regulatory region that is itself exhibiting ILD to the *CYP76A2* gene on chromosome 10. Although recombination should decouple cost alleles that are physically linked to resistance alleles, these loci would not completely decouple if the recombination rate (c) is much lower than the selection coefficient(s) of the resistance locus (i.e., c<<s, (Stephan *et al.*, 1992)). The requirement that c<<s is not improbable given the close proximity of cost and resistance loci (< 85kb) and the strong ongoing selection for herbicide resistance. Furthermore, if the ratio $c/s < 10^{-4}$, the hitchhiking would almost be complete and the cost alleles could become fixed in the populations (Fay & Wu, 2000). Alternatively, it is possible that new compensatory mutations arising in the population could increase in frequency over time due to selection, decoupling the cost and resistance alleles and thus reducing fitness cost associated with glyphosate resistance (Vogwill *et al.*, 2016; Lenormand *et al.*, 2018).

Overall, our work identified the potential genes associated with NTSR glyphosate resistance in *I. purpurea* -- our whole genome and transcriptome assays strongly support the role of detoxification conferring herbicide resistance in this species, and we additionally identified a role for plant sensing and stress. Interestingly, we show that NTSR glyphosate resistance in *I. purpurea* involves multiple loci which are maintained through generations via ILD. We also provide strong evidence to support the idea that fitness costs may be due to loci in strong linkage with resistance loci. Our work highlights the importance of multi-level, multi-population study in identifying the genetic mechanisms underlying polygenic defense traits, and for understanding the complex genetic-interplay between defense and cost.

Materials and Methods

Genome sequencing, assembly, and annotation

We used an *I. purpurea* line originally sampled from an agricultural field in Orange County, NC, in 1985 by M. Rausher (*i.e.*, prior to the widespread use of glyphosate) and selfed for >18 generations in the lab for genome sequencing (seeds of this line 'Fred/C' are available upon request). High molecular weight DNA was isolated from flash-frozen leaf tissue using a

modified large-volume CTAB protocol (Doyle & Doyle, 1990) and sequenced on a PacBio Sequel at the University of Georgia. Raw PacBio subreads from 9 cells of Sequel chemistry were error-corrected with Canu (v1.7.1) (Koren *et al.*, 2017) with default parameters for raw PacBio reads (--pacbio-raw). The corrected and trimmed reads from Canu were assembled with Flye (v2.4-release) (Kolmogorov *et al.*, 2019) and anchored onto pseudomolecules by nearly 81 million read pairs of Phase Genomics Hi-C (Seattle, WA) of leaf tissue using Sau3AI cutsites. Within-genome and across-genome synteny was visualized using the CoGE SynMap platform (Lyons, 2008), with DAGChainer options "-D20 -A 5", as well as with jcvi with default parameters (https://github.com/tanghaibao/jcvi). *Ipomoea purpurea* pseudomolecules were numbered and oriented according to chromosome synteny against *Ipomoea nil* pseudomolecules (*Appendix B*, Figure S3-2).

Raw 50nt single-end RNA-seq reads were aligned using STAR (v.2.7.0) (Dobin *et al.*, 2013) with default single-pass parameters. Repetitive elements were first annotated with RepeatModeler (v1.0.11). Long Terminal Repeat (LTR) retrotransposons were annotated with LTRharvest (v1.6.1) with options -similar 85 -mindistltr 1000 -maxdistltr 15000 -mintsd 5 -maxtsd 20". RepeatModeler annotations were combined with all Viridiplantae repeats from Repbase and used as a species-specific repeat database built using RepeatModeler with default options.

Genome annotation was performed using a diverse set of evidence. First, a set of 12 RNA-seq libraries from leaf tissue was aligned with STAR (v2.7.0), and transcripts assembled with Stringtie (v2.1.3). MAKER2 (Holt & Yandell, 2011) was initially run with evidence from the RNA-seq alignments, as well as peptides from *I. trifida, I. triloba,* and *I. nil*. The resulting gene set was used to train SNAP (v2013-11-29) (Korf, 2004) . AUGUSTUS (v3.3.2) (Stanke *et al.*, 2006) was trained with evidence from BUSCO (v4.1.0) (Simão *et al.*, 2015) against the eudicot odb10 set. with default options. MAKER2 was re-run with the *ab initio* SNAP and AUGUSTUS training sets, in addition to the homologous protein and RNA-seq evidence, to build a final gene annotation set.

Sampling and sequencing

We selected eight populations to investigate the genetic basis of glyphosate resistance and its cost following (Van Etten *et al.*, 2020) -- 4 low resistance, from here on referred to as the susceptible population (S: <20% population survival at 1X the field dose of RoundUp) and 4 high resistance populations (R: >70% population survival at 1X the field dose of RoundUp), from here on referred to as the resistant populations (*Appendix B*, Table S3-12). Seeds from 10 maternal lines per population were germinated, except for one susceptible population (RB), wherein 9 maternal lines were used. We extracted DNA from leaf tissue using the Qiagen Plant DNeasy kit. 150 paired-end sequencing was performed using Illumina HiSeq4000 and NovaSeq6000 using three and two lanes, respectively. We sequenced two populations at high coverage (at least 25X) and the remaining six populations at low coverage (10X). Two populations (WG, resistant and RB, susceptible) were run on one lane of HiSeq6000 and NovaSeq6000 each whereas the other lane had the remaining six populations. This yielded a total of 3,300,397,148,700 bases with average coverage of 28.84X for WG and RB. Coverage of the other six populations has an average of 14.66X.

Variant calling

We aligned the reads to our draft genome using BWA mem v0.7.15 (Li & Durbin, 2009) with parameter -M. Since the same sample was sequenced using multiple platforms (HiSeq and NovaSeq), the alignment files were merged and duplicate reads were marked using the MarkDuplicate tool of Picard v2.8.1 (http://broadinstitute.github.io/picard). Next, we prepared a database of true known variants, required for base recalibration. This database was created using data from the top eighteen individuals with the highest read counts, upon which variant call was performed using the HaplotypeCaller tool of GATK v4.1 (McKenna *et al.*, 2010). Low confidence variants were filtered out using the VariantFiltration tool of GATK v4.1 (McKenna *et al.*, 2010) (15 < DP < 60; ReadPosRankSum < -8.0; QD < 2.0; FS > 60.0; SOR > 3.0; MQ < 40.0; MQRankSum < -12.5) and only the high confidence variants were used in the dataset. This was used to recalibrate base qualities using GATK v4.1 tools BaseRecalibrator and ApplyBQSR (McKenna *et al.*, 2010). Variants were called individually on all the individuals using the HaplotypeCaller tool of GATK v4.1 (McKenna *et al.*, 2010) using parameters -ERC GVCF --min-pruning 1 --min-dangling-branch-length 1. The variants from each individual were

combined to one variant file (a raw cohort variant file) using the tools GenomicsDBImport, GenotypeGVCFs, and GatherVcfs (McKenna *et al.*, 2010), with invariants included. Next, multiple rounds of filtration were performed on this variant dataset to filter out potential false positives. First, using the GATK v4.1 tools VariantFiltration and SelectVariants we filtered the variants using the parameters QD<1.5, DP<10 and DP>2000, FS>80, SOR>5, MQ<40, MQRankSum< -6 and MQRankSum>6, and ReadPosRankSum< -4 and ReadPosRankSum> 4 (McKenna *et al.*, 2010). For the next round of filtration, we removed variants that had genotype depth more than twice the average and heterozygosity more than 0.8 using the het packages from VCFtools v0.1.15 (Danecek *et al.*, 2011). In the third round of filtration, we filtered variants that had quality above 20, had no missing information, a minor allele frequency of 0.05, and a minimum mean depth of 10 (vcftools --minQ 20 --max-missing 1.0 --maf 0.05 --min-meanDP 10) (Danecek *et al.*, 2011). Finally, we filtered using BCFtools (v1.7) (Li, 2011) to keep only bi-allelic SNPs (bcftools view -m2 -M2 -v snps). This gave us a total of 3,942,549 high confidence SNPs. These SNPs were used for downstream analyses.

We performed a PCA analysis using the allele frequencies of all the SNPs to investigate the population structure using the package bigsnpr v.1.4.4 (Privé *et al.*, 2018) in R and found that the populations did not segregate into two separate genetic clusters (*Appendix B*, Figure S3-4a-b). Further, we repeated this analysis for SNPs from regions under selection (see below) to test whether we observe the same population structure patterns. We observed that these separated into distinct resistant and susceptible groups, except for a resistant population, BI, which clustered between the susceptible and other resistant populations (*Appendix B*, Figure S3-4c). Thus, for the purposes of this study, we dropped the BI population from further analysis.

Selection analysis

To identify the regions associated with resistance, we used a two step approach. First, we used bayenv2 (Günther & Coop, 2013), which is designed to identify candidate outliers while taking population structure into account by incorporating a covariance matrix of population allele frequencies (Günther & Coop, 2013). bayenv2 was run on a pruned dataset that contained SNPs that were at least 5kb apart (vcftools --thin 5000), resulting in a total of 86,648 SNPs. Ten independent covariance matrices were constructed for sets of 5000 randomly selected SNPs from

the global dataset, by running bayenv2 for 100,000 iterations. The final covariance matrix was constructed by averaging across the 10 independent runs. Using this matrix, we ran bayenv2 with different random seeds for 5 independent runs with 100,000 iterations each on the pruned set of SNPs. We then used the test statistic averaged over the five runs to identify the loci under selection. We considered SNPs as robust candidates if they belonged in the top 5 % of Bayes factors (BFs) and Spearman's ρ. To define genes under selection, we picked up genes in the 5kb up- and downstream spanning regions of the robust candidates.

Secondly, we employed a Md-rank-*P* approach that integrates multiple selection statistics to identify the regions under selection. Recently, composite selection testing has been proposed as a way to confidently identify regions under selection and has been shown to considerably reduce false-positives (Lotterhos *et al.*, 2017; Yurchenko *et al.*, 2018; Brennan *et al.*, 2018). We split the high confidence variant dataset obtained into 'resistant' and 'susceptible' variant datasets using vcf-subset of VCFtools v0.1.15 (Danecek *et al.*, 2011). The 'resistant' and 'susceptible' variant datasets contained 30 and 39 individuals, respectively (*Appendix B*, Table S3-12). We then used these datasets to calculate diversity and selection statistics $G_{ST\ (Weir,\ 1996)}$, pi, Tajima's D (Tajima, 1989), Fu and Way's H (Fay & Wu, 2003) using a custom script from (Baduel *et al.*, 2018) in a 300SNP window, for both the dataset. Furthermore, to detect hard sweep we phased the variants using beagle version 5.1 (Browning & Browning, 2007) which was then used to calculate the haplotype homozygosity statistic (H12, a measure of haplotype homozygosity that detects both hard and soft sweeps) using the scripts provided (Garud *et al.*, 2015). For regions above 95 percentile G_{ST}, we calculated a composite rank-based statistic (Md-rank-*P*) which was computed as the Mahalanobis distance on the negative log10 transformation of raw statistics into rank P-values (Lotterhos *et al.*, 2017). This Md-rank-*P* was calculated using pi, Tajima's D, Fu and Way's H, and H12. To identify potential regions of selection we chose bins with greater than 95 percentile Md-rank-*P*. Genes that were identified via both approaches were considered as the genes associated with herbicide resistance in *I. purpurea*.

Linkage analysis

We calculated linkage disequilibrium (r^2) at three different levels. First, to estimate the background genome-wide long-distance (and interchromosomal) linkage disequilibrium (ILD),

we calculated r^2 values for 5842 SNPs separated by at least 100kb using VCFtools v0.1.15 (Danecek *et al.*, 2011) (--thin 100000 --interchrom-hap-r2). Second, we estimated the r^2 for SNPs separated by at least 1kb in and between broad regions (0.75Mb upstream and downstream) around the five focused regions (with $G_{ST} > 0.39$) under selection using VCFtools v0.1.15 (Danecek *et al.*, 2011). For this, we used the range under selection obtained from the Md-rank-*P* approach since it uses bins of 300 SNPs, as compared to bayenv2 which only provides outlier SNPs. Lastly, since one would expect higher linkage between regions with high differentiation, we also randomly chose four regions with high differentiation (showing no signs of selection) of similar lengths as the five focused regions above and compared its linkage values to those regions.

RNA-Seq

To identify transcripts associated with glyphosate resistance and its potential cost, we sequenced transcriptomes of 17 individuals belonging to four different treatments; resistant control (Rc), susceptible control (Sc), resistant herbicide sprayed (Rh), and susceptible herbicide sprayed (Sh). Each treatment had multiple individuals (Rc-2, Sc-2, Rh-6, Sh-7; *Appendix B*, Table S3-13). These individuals were generated via selfing in the growth chamber and were grown in a controlled environment (growth chamber) to reduce variation due to environmental differences. 20 days after planting, we sprayed glyphosate (concentration of 1.52 kg ai/ha) on the Rh and Sh treatment plants and collected the second and fourth leaf for RNA extractions 8 hours post-spray. These were flash frozen using liquid nitrogen and stored at -80°C. We extracted RNA using Qiagen RNeasy Plant mini kit with the optional DNase digestion step. This was then sequenced using Illumina NovaSeq 6000 at 150bp paired-end sequencing. A total of 132,551,535,000 bp were obtained.

Differential gene expression -- We processed the raw reads obtained to remove adapters using cutadapt v1.18 (Martin, 2011) and then mapped them to the de-novo assembled genome (--sjdbOverhang 149 --outSAMtype BAM SortedByCoordinate Unsorted) using STAR v2.7.5 (Dobin *et al.*, 2013). Next, using HTSeq v0.11.1(Anders *et al.*, 2015), we counted read counts for each gene. These read counts were then used to filter out lowly expressed transcripts using the Bioconductor package edgeR version 3.18.1 (Robinson *et al.*, 2010) such that transcripts were

retained only if they had greater than 0.5 counts-per-million in at least two samples (Rc vs Sc) and four samples (Rh vs Sh). The libraries were then normalized in edgeR (using the trimmed mean of M-values method) followed by differential gene expression analysis using the classic pairwise comparison of edgeR version 3.18.1 (Robinson *et al.*, 2010). We extracted the significance of differentially expressed transcripts (DETs) with FDR <= 0.05. This was done for two contrasts, Rh vs Sh (total sample size = 13; Rh = 6, Sh = 7) and Rc vs Sc (total sample size 4; Rc = 2, Sc = 2). The first contrast informs us of the genes that are regulated in response to the herbicide, and how this gene regulation differs between the resistant and the susceptible populations, whereas the latter informs us of the baseline expression difference due to glyphosate resistance between the two populations.

Malathion Experiment

On May 15th, 2019, we planted a total of 180 replicate seeds from multiple resistant and susceptible populations (*Appendix B*, Table S3-14) in Cone-Tainers (Stewe and Sons). These were allowed to grow for 30 days, after which we subjected them to one of the four treatment environments--malathion (7.81 ml/ 500 mL), glyphosate (3.4 kg ai/ha), glyphosate and malathion, and a control. Malathion was applied using a handheld sprayer, and glyphosate was applied 24 hours later using a hand-held CO2 sprayer (Spraying Systems Co., Wheaton, IL) calibrated to deliver 187 L/ha. Twenty-five days post treatment spray, we harvested the plants. These were dried for 3 days at 70C and weighed for an estimate of dry above ground-biomass.

Using this data, we assessed whether biomass was significantly altered by the different treatments. First, we normalized the above-ground biomass using the transformTukey function from Rcompanion v.2.0.0 (Mangiafico, 2015). We then used a generalized linear model (lm function (Computing, 2013)) with normalized biomass as the dependent variable and population Type (R/S) and treatment as the independent variables. We assessed the significance of the variables using the Anova function of the car package v.3.0.10 (Fox & Weisberg, 2018), and performed a pairwise comparison between groups using the lsmeans function from package lsmeans v2.30.0 (Lenth, 2016), adjusted for multiple tests using tukey correction. Using the same general model, we also compared whether biomass was significantly different between treatments for each population type. To control for the differences in the plant size we

standardized the biomass of the individuals by the average biomass of the respective maternal line in the control treatment, and then normalized it as above.

Acknowledgements

We thank M. Rausher for supplying germplasm used in the *I. purpurea* genome assembly, and T. Newsum, S. Paranjape, and K. Johnson for growth room phenotyping, and M. Palmer and MBGNA for growth room support. We likewise thank J. Opp in the University of Michigan sequencing core for sequencing support, Amanda Cummings for high molecular weight DNA preparation, and the Georgia Genomics and Bioinformatics Core, which provided the PacBio Sequel sequencing service. We also thank E.B. Josephs, R.L. Rogers, and the Ross-Ibarra lab for providing feedback on the manuscript. Funding for this work was provided by the University of Michigan and USDA NIFA awards 24892 & 28497 to RSB, NSF DEB 1442199 and IOS 1444567 to J.L.M., and NSF IOS 1611853 to A.H.

References

Alvarado-Serrano DF, Van Etten ML, Chang S-M, Baucom RS. 2019. The relative contribution of natural landscapes and human-mediated factors on the connectivity of a noxious invasive weed. *Heredity* 122: 29–40.

Amrhein N, Martinoia E. 2021. An ABC transporter of the ABCC subfamily localized at the plasma membrane confers glyphosate resistance. *Proceedings of the National Academy of Sciences of the United States of America* 118.

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166–169.

Baduel P, Hunter B, Yeola S, Bomblies K. 2018. Genetic basis and evolution of rapid cycling in railway populations of tetraploid Arabidopsis arenosa. *PLoS genetics* 14: e1007510.

Baucom RS. 2019. Evolutionary and ecological insights from herbicide-resistant weeds: what have we learned about plant adaptation, and what is left to uncover? *The New phytologist* 223: 68–82.

Baucom RS, Mauricio R. 2004. Fitness costs and benefits of novel herbicide tolerance in a noxious weed. *Proceedings of the National Academy of Sciences of the United States of America* 101: 13386–13390.

Beckie HJ. 2020. Herbicide Resistance in Plants. *Plants* 9.

Behrouzi P, Wit EC. 2017. Detecting Epistatic Selection with Partially Observed Genotype Data Using Copula Graphical Models. *arXiv [stat.AP]*.

Bergelson J, Purrington CB. 1996. Surveying Patterns in the Cost of Resistance in Plants. *The American naturalist* 148: 536–558.

Brennan RS, Healy TM, Bryant HJ, Van La M, Schulte PM, Whitehead A. 2018. Integrative Population and Physiological Genomics Reveals Mechanisms of Adaptation in Killifish. *Molecular biology and evolution* 35: 2639–2653.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* 81: 1084–1097.

Busi R, Neve P, Powles S. 2013. Evolved polygenic herbicide resistance in Lolium rigidum by low-dose herbicide selection within standing genetic variation. *Evolutionary applications* 6: 231–242.

Christopher JT, Preston C, Powles SB. 1994. Malathion Antagonizes Metabolism-Based Chlorsulfuron Resistance in Lolium rigidum. *Pesticide biochemistry and physiology* 49: 172–182.

Computing RFS. 2013. R: A language and environment for statistical computing. *Vienna: R Core Team*.

Cummins I, Wortley DJ, Sabbadin F, He Z, Coxon CR, Straker HE, Sellars JD, Knight K, Edwards L, Hughes D, *et al.* 2013. Key role for a glutathione transferase in multiple-herbicide resistance in grass weeds. *Proceedings of the National Academy of Sciences of the United States of America* 110: 5812–5817.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.

Délye C. 2013. Unravelling the genetic bases of non-target-site-based resistance (NTSR) to herbicides: a major challenge for weed science in the forthcoming decade. *Pest management science* 69: 176–187.

Délye C, Gardin JAC, Boucansaud K, Chauvel B, Petit C. 2011. Non-target-site-based resistance should be the centre of attention for herbicide resistance research: Alopecurus myosuroides as an illustration: Why we need more research on NTSR. *Weed research* 51: 433–437.

Délye C, Jasieniuk M, Le Corre V. 2013. Deciphering the evolution of herbicide resistance in weeds. *Trends in genetics: TIG* 29: 649–658.

Dimaano NG, Iwakami S. 2021. Cytochrome P450-mediated herbicide metabolism in plants: current understanding and prospects. *Pest management science* 77: 22–32.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21.

Dobzhansky T. 1971. *Genetics of the Evolutionary Process*. Columbia University Press.

Dong X, Hong Z, Sivaramakrishnan M, Mahfouz M, Verma DPS. 2005. Callose synthase (CalS5) is required for exine formation during microgametogenesis and for pollen viability in Arabidopsis. *The Plant journal: for cell and molecular biology* 42: 315–328.

Doyle JJ, Doyle JL. 1990. Isolation ofplant DNA from fresh tissue. *Focus* 12: 39–40.

Duhoux A, Délye C. 2013. Reference genes to study herbicide stress response in Lolium sp.: up-regulation of P450 genes in plants resistant to acetolactate-synthase inhibitors. *PloS one* 8: e63576.

Dyer WE. 2018. Stress-induced evolution of herbicide resistance and related pleiotropic effects. *Pest management science* 74: 1759–1768.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413.

Fay JC, Wu C-I. 2003. Sequence Divergence, Functional Constraint, and Selection in Protein Evolution. *Annual review of genomics and human genetics* 4: 213–235.

Fox J, Weisberg S. 2018. *An R Companion to Applied Regression*. SAGE Publications.

Franzoni G, Cocetta G, Trivellini A, Ferrante A. 2019. Transcriptional Regulation in Rocket Leaves as Affected by Salinity. *Plants* 9.

Gaines TA, Duke SO, Morran S, Rigon CAG, Tranel PJ, Küpper A, Dayan FE. 2020. Mechanisms of evolved herbicide resistance. *The Journal of biological chemistry* 295: 10307–10330.

Gaines TA, Patterson EL, Neve P. 2019. Molecular mechanisms of adaptive evolution revealed by global selection for glyphosate resistance. *The New phytologist* 223: 1770–1775.

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS genetics* 11: e1005004.

Ghanizadeh H, Harrington KC. 2017. Non-target Site Mechanisms of Resistance to Herbicides. *Critical reviews in plant sciences* 36: 24–34.

Giri J, Vij S, Dansana PK, Tyagi AK. 2011. Rice A20/AN1 zinc-finger containing stress-associated proteins (SAP1/11) and a receptor-like cytoplasmic kinase (OsRLCK253) interact via A20 zinc-finger and confer abiotic stress tolerance in transgenic Arabidopsis plants. *The New phytologist* 191: 721–732.

Giuntoli B, Licausi F, van Veen H, Perata P. 2017. Functional Balancing of the Hypoxia Regulators RAP2.12 and HRA1 Takes Place in vivo in Arabidopsis thaliana Plants. *Frontiers in plant science* 8: 591.

Günther T, Coop G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics* 195: 205–220.

Hardie DG. 1999. PLANT PROTEIN SERINE/THREONINE KINASES: Classification and Functions. *Annual review of plant physiology and plant molecular biology* 50: 97–131.

Hawkins NJ, Bass C, Dixon A, Neve P. 2018. The evolutionary origins of pesticide resistance. *Biological reviews of the Cambridge Philosophical Society*.

Hench K, Vargas M, Höppner MP, McMillan WO, Puebla O. 2019. Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. *Nature ecology & evolution* 3: 657–667.

Hohenlohe PA, Bassham S, Currey M, Cresko WA. 2012. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367: 395–408.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics* 12: 491.

Hoshino A, Jayakumar V, Nitasaka E, Toyoda A, Noguchi H, Itoh T, Shin-I T, Minakuchi Y, Koda Y, Nagano AJ, *et al.* 2016. Genome sequence and analysis of the Japanese morning glory Ipomoea nil. *Nature communications* 7: 13295.

Huang X-X, Zhao S-M, Zhang Y-Y, Li Y-J, Shen H-N, Li X, Hou B-K. 2021. A novel UDP-glycosyltransferase 91C1 confers specific herbicide resistance through detoxification reaction in Arabidopsis. *Plant physiology and biochemistry: PPB / Societe francaise de physiologie vegetale* 159: 226–233.

Jasieniuk M, Anita L. Brûlé-Babel, Ian N. Morrison. 1996. The Evolution and Genetics of Herbicide Resistance in Weeds. *Weed Science* 44: 176–193.

Jiang J, Ma S, Ye N, Jiang M, Cao J, Zhang J. 2017. WRKY transcription factors in plant responses to stresses. *Journal of integrative plant biology* 59: 86–101.

Jugulam M, Shyam C. 2019. Non-Target-Site Resistance to Herbicides: Recent Developments. *Plants* 8.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* 37: 540–546.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* 27: 722–736.

Korf I. 2004. Gene finding in novel genomes. *BMC bioinformatics* 5: 59.

Kreiner JM, Sandler G, Stern AJ, Tranel PJ, Weigel D, Stinchcombe JR, Wright SI. 2021. Repeated origins, gene flow, and allelic interactions of herbicide resistance mutations in a widespread agricultural weed. *bioRxiv*: 2021.05.10.443516.

Kreiner JM, Stinchcombe JR, Wright SI. 2018. Population Genomics of Herbicide Resistance: Adaptation via Evolutionary Rescue. *Annual review of plant biology* 69: 611–635.

Kreiner JM, Tranel PJ, Weigel D, Stinchcombe JR, Wright SI. 2020. The genetic architecture and genomic context of glyphosate resistance. *Cold Spring Harbor Laboratory*: 2020.08.19.257972.

Kuester A, Chang S-M, Baucom RS. 2015. The geographic mosaic of herbicide resistance evolution in the common morning glory, Ipomoea purpurea: Evidence for resistance hotspots and low genetic differentiation across the landscape. *Evolutionary applications* 8: 821–833.

Kuester A, Wilson A, Chang S-M, Baucom RS. 2016. A resurrection experiment finds evidence of both reduced genetic diversity and potential adaptive evolution in the agricultural weed Ipomoea purpurea. *Molecular ecology* 25: 4508–4520.

Kunieda T, Mitsuda N, Ohme-Takagi M, Takeda S, Aida M, Tasaka M, Kondo M, Nishimura M, Hara-Nishimura I. 2008. NAC Family Proteins NARS1/NAC2 and NARS2/NAM in the Outer Integument Regulate Embryogenesis in Arabidopsis. *The Plant cell* 20: 2631–2642.

Lai Z, Vinod K, Zheng Z, Fan B, Chen Z. 2008. Roles of Arabidopsis WRKY3 and WRKY4 transcription factors in plant responses to pathogens. *BMC plant biology* 8: 68.

Lee J-H, Kim WT. 2011. Regulation of abiotic stress signal transduction by E3 ubiquitin ligases in Arabidopsis. *Molecules and cells* 31: 201–208.

Lenormand T, Harmand N, Gallet R. 2018. Cost of resistance: an unreasonably expensive concept. *Rethinking Ecology* 3: 51–70.

Lenth RV. 2016. Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software, Articles* 69: 1–33.

Leon RG, Dunne JC, Gould F. 2021. The role of population and quantitative genetics and modern sequencing technologies to understand evolved herbicide resistance and weed fitness. *Pest management science* 77: 12–21.

Lewontin RC, Kojima K-I. 1960. The Evolutionary Dynamics of Complex Polymorphisms. *Evolution; international journal of organic evolution* 14: 458–472.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987–2993.

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

Lim CW, Yang SH, Shin KH, Lee SC, Kim SH. 2015. The AtLRK10L1.2, Arabidopsis ortholog of wheat LRK10, is involved in ABA-mediated signaling and drought resistance. *Plant Cell Reports* 34: 447–455.

Liu N. 2015. Insecticide resistance in mosquitoes: impact, mechanisms, and research directions. *Annual review of entomology* 60: 537–559.

Liu W, Bai S, Zhao N, Jia S, Li W, Zhang L, Wang J. 2018. Non-target site-based resistance to tribenuron-methyl and essential involved genes in Myosoton aquaticum (L.). *BMC plant biology* 18: 225.

Liu Y, Geyer R, van Zanten M, Carles A, Li Y, Hörold A, van Nocker S, Soppe WJJ. 2011. Identification of the Arabidopsis REDUCED DORMANCY 2 gene uncovers a role for the polymerase associated factor 1 complex in seed dormancy. *PloS one* 6: e22241.

Liu W, Karemera NJU, Wu T, Yang Y, Zhang X, Xu X, Wang Y, Han Z. 2017. The ethylene response factor AtERF4 negatively regulates the iron deficiency response in Arabidopsis thaliana. *PloS one* 12: e0186580.

Liu S, Li M, Su L, Ge K, Li L, Li X, Liu X, Li L. 2016. Negative feedback regulation of ABA biosynthesis in peanut (Arachis hypogaea): a transcription factor complex inhibits AhNCED1 expression during water stress. *Scientific reports* 6: 37943.

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjálmsson BJ, Korte A, Nizhynska V, *et al.* 2013. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature genetics* 45: 884–890.

Lotterhos KE, Card DC, Schaal SM, Wang L, Collins C, Verity B. 2017. Composite measures of selection can improve the signal-to-noise ratio in genome scans (J Kelley, Ed.). *Methods in ecology and evolution / British Ecological Society* 8: 717–727.

Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A. 2017. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular ecology resources* 17: 142–152.

Lyons EH. 2008. *CoGe, a new kind of comparative genomics platform: Insights into the evolution of plant genomes*. University of California, Berkeley.

Ma L, Li G. 2018. FAR1-RELATED SEQUENCE (FRS) and FRS-RELATED FACTOR (FRF) Family Proteins in Arabidopsis Growth and Development. *Frontiers in plant science* 9: 692.

Mangiafico SS. 2015. An R companion for the handbook of biological statistics. *Available: rcompanion. org/documents/RCompanionBioStatistics. pdf. (January 2016)*.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10–12.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* 2010. The Genome Analysis Toolkit: a MapReduce

framework for analyzing next-generation DNA sequencing data. *Genome research* 20: 1297–1303.

Mithila J, Godar AS. 2013. Understanding Genetics of Herbicide Resistance in Weeds: Implications for Weed Management. *Advances in Crop Science and Technology* 1: 1–3.

Murphy BP, Tranel PJ. 2019. Target-Site Mutations Conferring Herbicide Resistance. *Plants* 8.

Nei M. 1967. Modification of linkage intensity by natural selection. *Genetics* 57: 625–641.

Nei M, Li WH. 1973. Linkage disequilibrium in subdivided populations. *Genetics* 75: 213–219.

Pandian BA, Sathishraj R, Prasad PVV, Jugulam M. 2021. A single gene inherited trait confers metabolic resistance to chlorsulfuron in grain sorghum (Sorghum bicolor). *Planta* 253: 48.

Pan L, Yu Q, Han H, Mao L, Nyporko A, Fan L, Bai L, Powles S. 2019. Aldo-keto Reductase Metabolizes Glyphosate and Confers Glyphosate Resistance in Echinochloa colona. *Plant physiology* 181: 1519–1534.

Pan L, Yu Q, Wang J, Han H, Mao L, Nyporko A, Maguza A, Fan L, Bai L, Powles S. 2021. An ABCC-type transporter endowing glyphosate resistance in plants. *Proceedings of the National Academy of Sciences of the United States of America* 118.

Park HC, Kim ML, Kang YH, Jeon JM, Yoo JH, Kim MC, Park CY, Jeong JC, Moon BC, Lee JH, *et al.* 2004. Pathogen- and NaCl-induced expression of the SCaM-4 promoter is mediated in part by a GT-1 box that interacts with a GT-1-like transcription factor. *Plant physiology* 135: 2150–2161.

Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K. 2005. Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS genetics* 1: e33.

Piasecki C, Yang Y, Benemann DP, Kremer FS, Galli V, Millwood RJ, Cechin J, Agostinetto D, Maia LC, Vargas L, *et al.* 2019. Transcriptomic Analysis Identifies New Non-Target Site Glyphosate-Resistance Genes in Conyza bonariensis. *Plants* 8.

Portereiko MF, Sandaklie-Nikolova L, Lloyd A, Dever CA, Otsuga D, Drews GN. 2006. NUCLEAR FUSION DEFECTIVE1 encodes the Arabidopsis RPL21M protein and is required for karyogamy during female gametophyte development and fertilization. *Plant physiology* 141: 957–965.

Powles SB, Yu Q. 2010. Evolution in Action: Plants Resistant to Herbicides. *Annual review of plant biology* 61: 317–347.

Preston C, Tardif FJ, Christopher JT. 1996. Multiple Resistance to Dissimilar Herbicide Chemistries in a Biotype ofLolium rigidumDue to Enhanced Activity of Several Herbicide Degrading Enzymes. *Pesticide biochemistry and physiology*.

Privé F, Aschard H, Ziyatdinov A, Blum MGB. 2018. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* 34: 2781–2787.

Radwan DEM. 2012. Salicylic acid induced alleviation of oxidative stress caused by clethodim in maize (Zea mays L.) leaves. *Pesticide biochemistry and physiology* 102: 182–188.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140.

Scarabel L, Pernin F, Délye C. 2015. Occurrence, genetic control and evolution of non-target-site based resistance to herbicides inhibiting acetolactate synthase (ALS) in the dicot weed Papaver rhoeas. *Plant science: an international journal of experimental plant biology* 238: 158–169.

Schaper E, Eriksson A, Rafajlovic M, Sagitov S, Mehlig B. 2012. Linkage disequilibrium under recurrent bottlenecks. *Genetics* 190: 217–229.

Schluter D. 2000. *The Ecology of Adaptive Radiation*. OUP Oxford.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.

Simms EL, Rausher MD. 1987. Costs and Benefits of Plant Resistance to Herbivory. *The American naturalist* 130: 570–581.

Sohail M, Vakhrusheva OA, Sul JH, Pulit SL, Francioli LC, Genome of the Netherlands Consortium, Alzheimer's Disease Neuroimaging Initiative, van den Berg LH, Veldink JH, de Bakker PIW, *et al.* 2017. Negative selection in humans and fruit flies involves synergistic epistasis. *Science* 356: 539–542.

Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J. 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature* 400: 667–671.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* 34: W435–9.

Steffen JG, Kang I-H, Portereiko MF, Lloyd A, Drews GN. 2008. AGL61 interacts with AGL80 and is required for central cell development in Arabidopsis. *Plant physiology* 148: 259–268.

Stephan W, Wiehe THE, Lenz MW. 1992. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theoretical population biology* 41: 237–254.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.

Takahasi KR. 2007. Evolution of coadaptation in a subdivided population. *Genetics* 176: 501–511.

Takahasi KR, Tajima F. 2005. Evolution of coadaptation in a two-locus epistatic system. *Evolution; international journal of organic evolution* 59: 2324–2332.

Takemiya A, Sugiyama N, Fujimoto H, Tsutsumi T, Yamauchi S, Hiyama A, Tada Y, Christie JM, Shimazaki K-I. 2013. Phosphorylation of BLUS1 kinase by phototropins is a primary step in stomatal opening. *Nature communications* 4: 2094.

Thirugnanasambantham K, Durairaj S, Saravanan S, Karikalan K, Muralidaran S, Islam VIH. 2015. Role of Ethylene Response Transcription Factor (ERF) and Its Regulation in Response to Stress Encountered by Plants. *Plant molecular biology reporter / ISPMB* 33: 347–357.

Van Etten ML, Kuester A, Chang S-M, Baucom RS. 2016. Fitness costs of herbicide resistance across natural populations of the common morning glory, Ipomoea purpurea. *Evolution; international journal of organic evolution* 70: 2199–2210.

Van Etten M, Lee KM, Chang S-M, Baucom RS. 2020. Parallel and nonparallel genomic responses contribute to herbicide resistance in Ipomoea purpurea, a common agricultural weed. *PLoS genetics* 16: e1008593.

Vega T, Gil M, Martin G, Moschen S, Picardi L, Nestares G. 2020. Stress response and detoxification mechanisms involved in non-target-site herbicide resistance in sunflower. *Crop science* 60: 1809–1822.

Vemanna RS, Vennapusa AR, Easwaran M, Chandrashekar BK, Rao H, Ghanti K, Sudhakar C, Mysore KS, Makarla U. 2017. Aldo-keto reductase enzymes detoxify glyphosate and improve herbicide resistance in plants. *Plant biotechnology journal* 15: 794–804.

Vila-Aiub MM, Neve P, Powles SB. 2009. Fitness costs associated with evolved herbicide resistance alleles in plants. *The New phytologist* 184: 751–767.

Vogwill T, Kojadinovic M, MacLean RC. 2016. Epistasis between antibiotic resistance mutations and genetic background shape the fitness effect of resistance across species of Pseudomonas. *Proceedings. Biological sciences / The Royal Society* 283.

Wallace B. 1953. On Coadaptation in Drosophila. *The American naturalist* 87: 343–358.

Walper E, Weiste C, Mueller MJ, Hamberg M, Dröge-Laser W. 2016. Screen Identifying Arabidopsis Transcription Factors Involved in the Response to 9-Lipoxygenase-Derived Oxylipins. *PloS one* 11: e0153216.

Wang M, Chen B, Zhou W, Xie L, Wang L, Zhang Y, Zhang Q. 2021. Genome-wide identification and expression analysis of the AT-hook Motif Nuclear Localized gene family in soybean. *BMC genomics* 22: 361.

Weir BS. 1996. Genetic Data Analysis II Sinauer Associates. *Inc. , Sunderland, MA*.

Wilson JF, Goldstein DB. 2000. Consistent Long-Range Linkage Disequilibrium Generated by Admixture in a Bantu-Semitic Hybrid Population. *American journal of human genetics* 67: 926–935.

Xie Z, Nolan T, Jiang H, Tang B, Zhang M, Li Z, Yin Y. 2019. The AP2/ERF Transcription Factor TINY Modulates Brassinosteroid-Regulated Plant Growth and Drought Responses in Arabidopsis. *The Plant cell* 31: 1788–1806.

Xiong F, Ren J-J, Yu Q, Wang Y-Y, Kong L-J, Otegui MS, Wang X-L. 2019. AtBUD13 affects pre-mRNA splicing and is essential for embryo development in Arabidopsis. *The Plant journal: for cell and molecular biology* 98: 714–726.

Yanniccari M, Gigón R, Larsen A. 2020. Cytochrome P450 Herbicide Metabolism as the Main Mechanism of Cross-Resistance to ACCase- and ALS-Inhibitors in Lolium spp. Populations From Argentina: A Molecular Approach in Characterization and Detection. *Frontiers in plant science* 11: 600301.

Yuan JS, Tranel PJ, Stewart CN Jr. 2007. Non-target-site herbicide resistance: a family business. *Trends in plant science* 12: 6–13.

Yu Q, Powles S. 2014. Metabolism-based herbicide resistance and cross-resistance in crop weeds: a threat to herbicide sustainability and global crop production. *Plant physiology* 166: 1106–1118.

Yurchenko AA, Daetwyler HD, Yudin N, Schnabel RD, Vander Jagt CJ, Soloshenko V, Lhasaranov B, Popov R, Taylor JF, Larkin DM. 2018. Scans for signatures of selection in Russian cattle breed genomes reveal new candidate genes for environmental adaptation and acclimation. *Scientific reports* 8: 1–16.

Zhao X-Y, Qi C-H, Jiang H, You C-X, Guan Q-M, Ma F-W, Li Y-Y, Hao Y-J. 2019. The MdWRKY31 transcription factor binds to the MdRAV1 promoter to mediate ABA sensitivity. *Horticulture Research* 6: 66.

Zhong D, Pai A, Yan G. 2005. Costly Resistance to Parasitism. *Genetics* 169: 2127–2135.

Figures

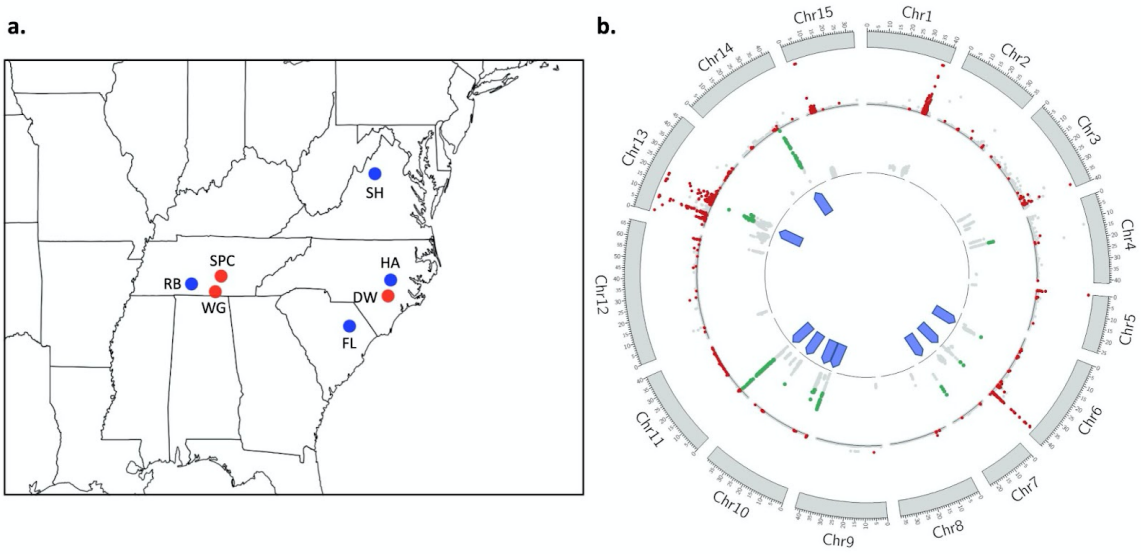

Figure 3-1 Overview of populations examined in this work and the genomic context of selection.
(a) Ten individuals were sampled from each population (except RB, wherein 9 individuals were sampled, with resistant populations (70-100% survival post-glyphosate) indicated in red and susceptible populations (< 20% survival post-glyphosate) indicated by blue markers (*Appendix B*, Table S3-12). (b) Circos plot depicting the regions of the genome that show signs of selection associated with herbicide resistance. The genome assembly resulted in 15 scaffolds which are represented here by grey bars. Significant values of Bayes Factor (BF) from bayenv2 (top 5% BF and top 5% Spearman's Rho) indicating outlier SNPs for herbicide resistance are depicted by red dots, and the significant Md-rank-*P* values (top 5% *Md-rank-P* value) identifying signatures of selection are presented in green dots. Regions of the genome that were significant using both bayenv2 and *Md-rank-P* are identified by the blue arrows.

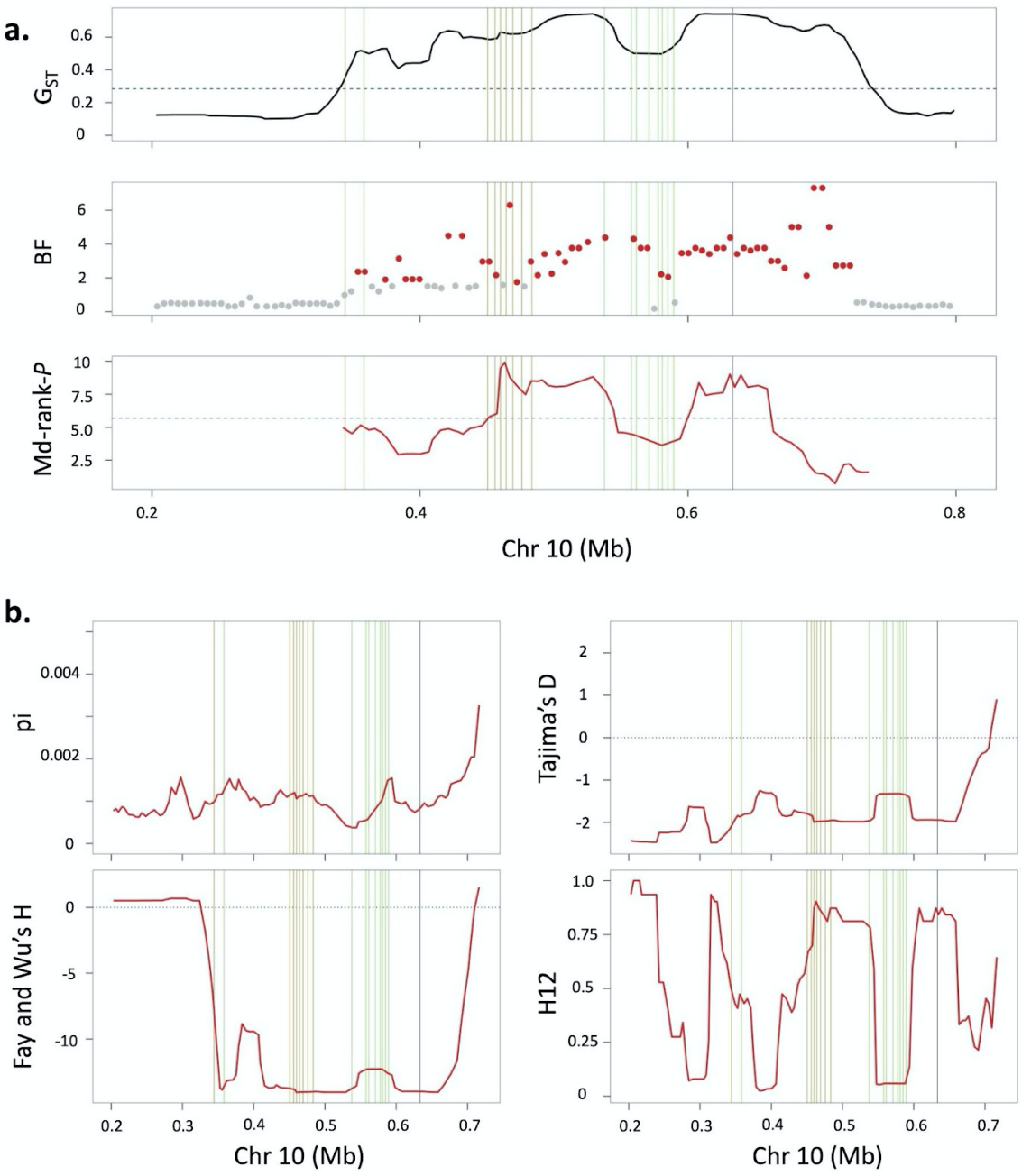

Figure 3-2 Region of Chromosome 10 showing signs of selection.
Shown is the (a) G_{ST} (upper), BayesFactor (BF) (middle) and Md-rank-*P* (lower) for the resistant individuals, and statistics used to estimate Md-rank-P (b) clockwise starting from upper left, pi, Tajima's D, H12 and Fay and Wu's H. Red lines indicate respective values for the resistant populations. Khaki vertical lines represent copies of glycosyltransferases, green vertical lines are the cytochrome P450, and the grey vertical line represents CTR1 (see below). The black dashed line in (a) represents 95 percentile values.

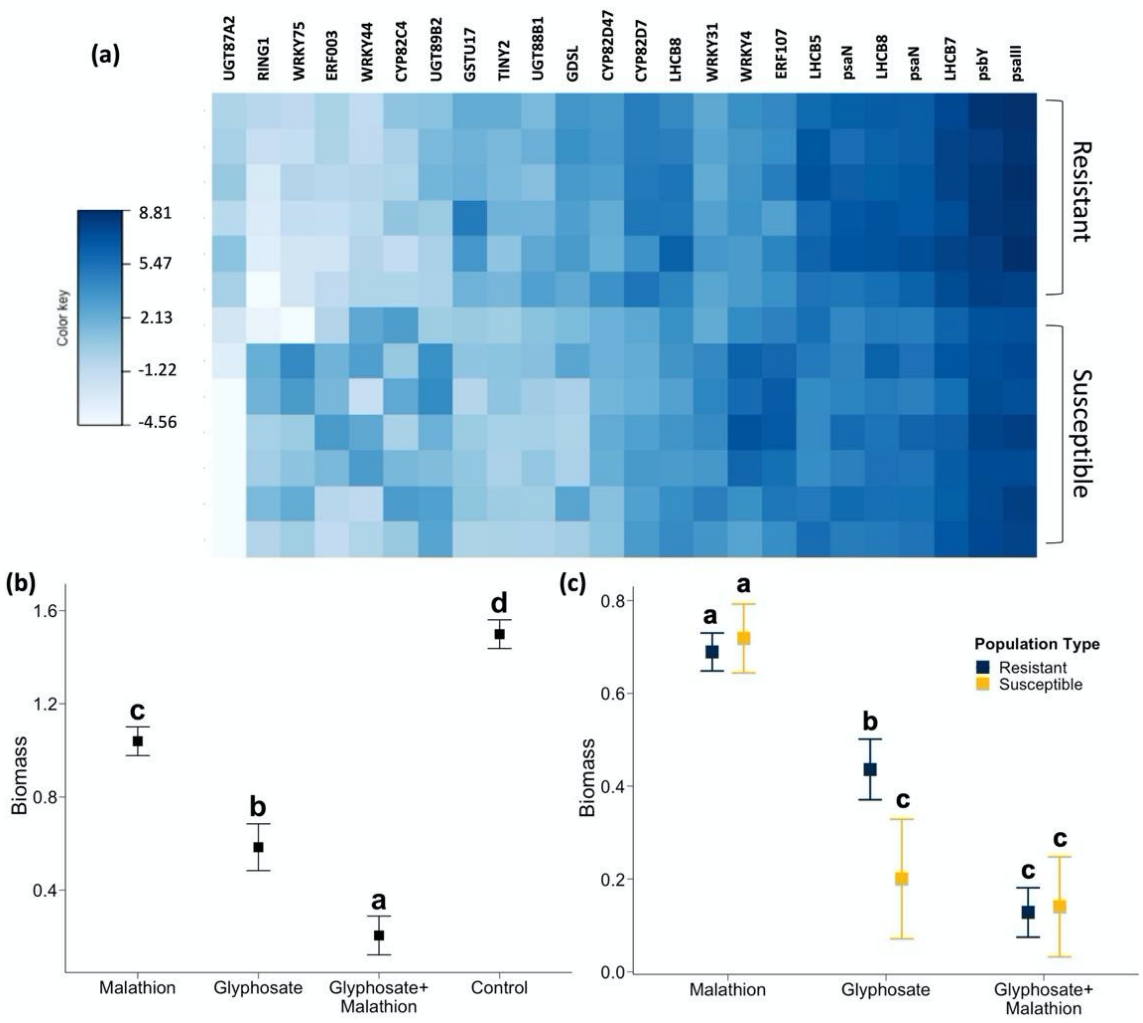

Figure 3-3 Gene expression variation associated with herbicide resistance, and results of a functional assay supporting the idea that resistance in *I. purpurea* is due to detoxification. (a) Loci associated with glyphosate resistance identified by differential expression analysis with P-value < 0.0005. Color key represents log2 fold-change values. (b) Least square means of above-ground biomass according to treatment (malathion, glyphosate, glyphosate plus malathion, and a control (no treatment)) and (c) summarized according to resistance type (R/S), normalized over the control. Letters in (b) and (c) indicate significant differences between treatment environments. The addition of the cytochrome P450 inhibitor malathion reverses glyphosate resistance (glyphosate vs glyphosate+malathion, contrast estimate = 0.379, t-ratio = 2.946, p-value = 0.019), with the resistant individuals showing the same phenotype as the susceptible individuals in the presence of glyphosate and malathion but not in the presence of glyphosate only. Error bars represent the one standard deviation from the least square means.

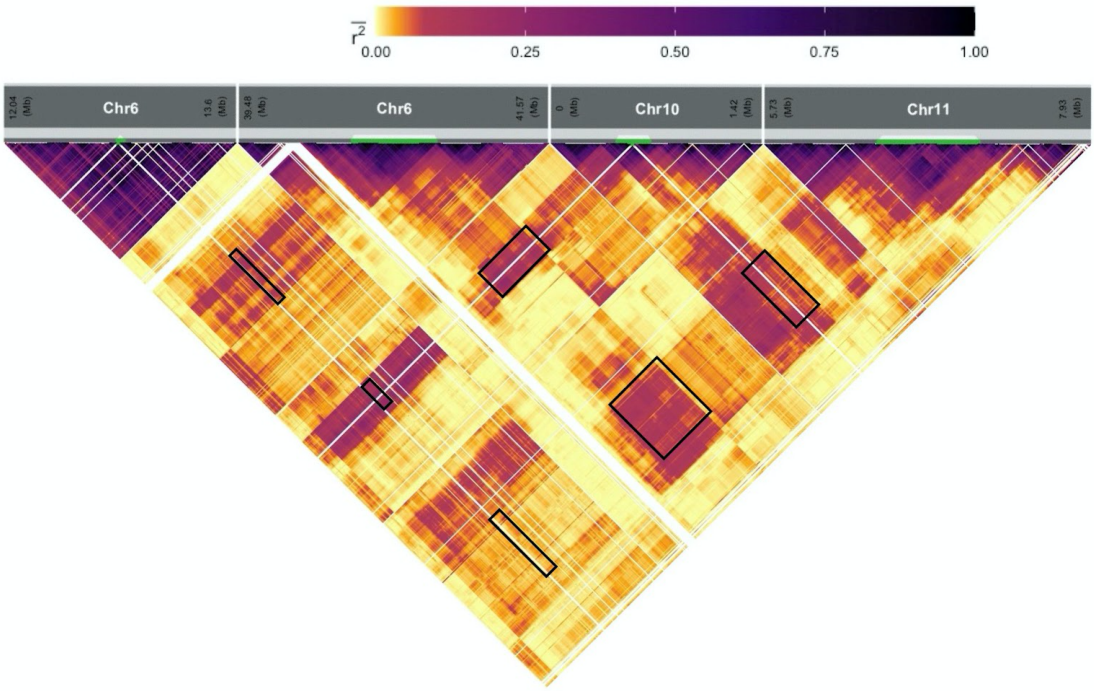

Figure 3-4 Long-distance linkage disequilibrium and ILD among the four highly differentiated (G_{ST} > 0.39) regions under selection associated with glyphosate resistance.
The four intervals displayed islands of increased linkage disequilibrium as estimated by r^2 for SNPs separated by at least 1 kb in and between broad regions under selection. The white lines represent the absence of SNPs (missing data) whereas the black boxes represent linkage between the five selection intervals. r^2 values are averaged over two-dimensional bins of 10 x 10 kb.

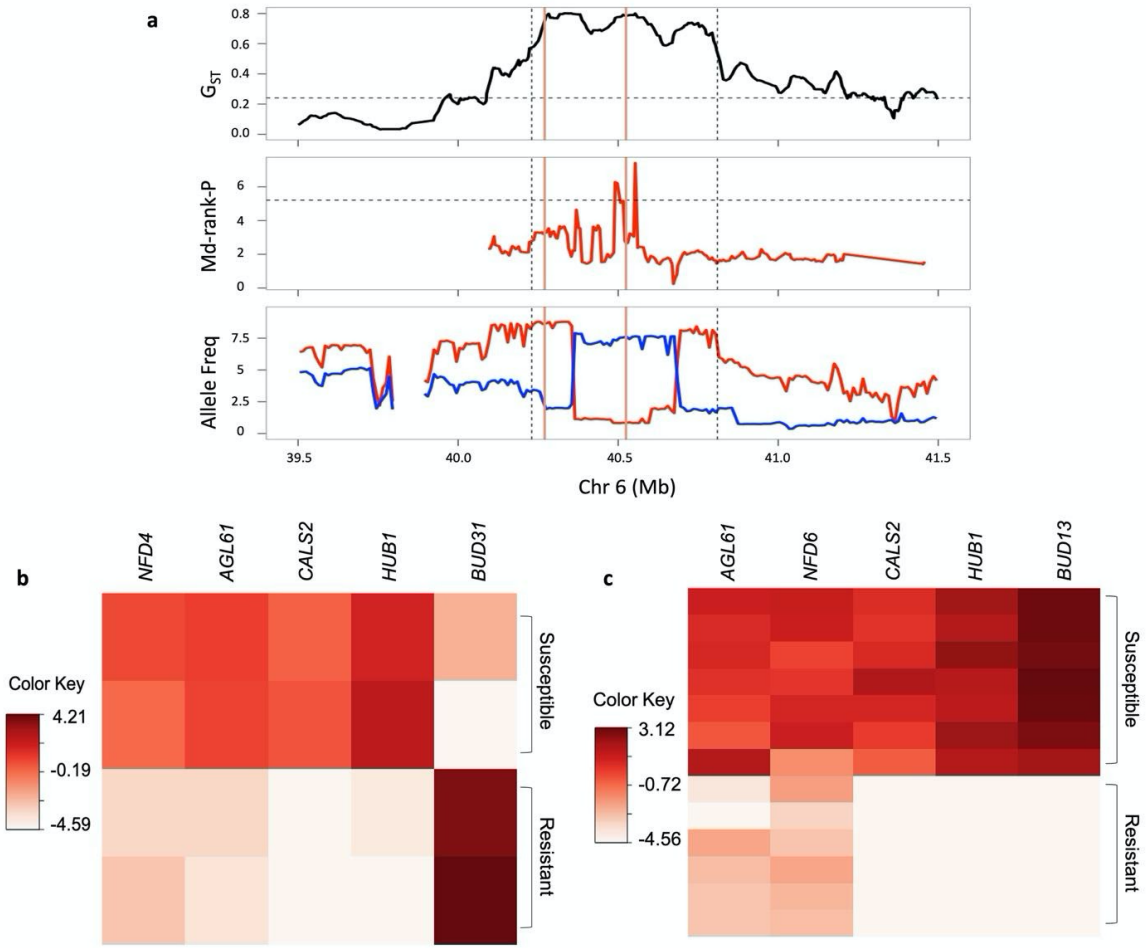

Figure 3-5 Loci associated with the cost of glyphosate resistance identified by the (a) whole-genome selection-scan and differential expression analysis in the (b) absence and (c) presence of herbicide.
Top panel of (a) represents G_{ST} between the resistant and the susceptible populations, mid-panel is the Md-rank-*P* value, and the lower panel represents the allele frequency. Salmon vertical lines represent NFD6 and NAC25, in that order. Red and blue represent resistant and susceptible populations, respectively. Black horizontal dotted lines represent 95 percentile values while vertical lines represent regions with G_{ST} above 0.6. The differentially expressed cost genes shown here were chosen based on their functional annotation and had FDR < 0.005 and P-value < 0.00005. Color key represents log2 fold-change values.

Tables

Chr	Location (Mb)	Genes of interest	Functional annotations for genes of interest	Process involved In
Chr6	12.82-12.84	XP_019190118.1	trihelix transcription factor GT-3b-like	Response to stress
Chr6	40.40-40.57	XP_019170142.1	MOR1-like	Microtubules organization during mitosis and cytokinesis
Chr7	16.78-16.83	XP_019163643.1	AP2-like ethylene-responsive transcription factor PLT1, zinc finger A20 and AN1 domain-containing stress-associated protein 8	Responses to environmental stimuli
Chr10	0.43-0.66	XP_019187799.1, XP_019187801.1, XP_019187802.1, XP_019187803.1, XP_019187804.1, XP_019188107.1, XP_019186932.1, XP_019186934.1, XP_019186939.1, XP_019186942.1	Glycosyltransferase (x7), CYP76A1 (x3), CYP76A2 (x3)	Detoxification
		XP_019187692.1	serine/threonine-protein kinase CTR1	Stress signaling
Chr11	6.45-7.31	XP_019154582.1	CYP736A12	Detoxification
		XP_019195164.1, XP_019195154.1, XP_019195155.1, XP_019195149.1	phosphate transporter PHO1 homolog 3-like (x5),	Transport of glyphosate into the cell
		XP_019195564.1, XP_019154554.1, XP_019154738.1, XP_019154605.1, XP_019154740.1, XP_019154739.1, XP_019154737.1	NAC domain-containing protein 92, LOB domain-containing protein 41, Leaf Rust 10 Disease-resistance locus Receptor-like Protein Kinase (x6)	Stress response
Chr13	16.45-17.73	XP_019171785.1, XP_019172641.1, XP_019172676.1	serine/threonine-protein kinase-like protein At5g23170, AT-hook motif nuclear-localized protein 24-like, zinc finger A20 and AN1	Stress sensing and response

			domain-containing stress-associated protein 1	
Chr13	20.91-21.22	XP_019179507.1	CYP87A3	Detoxification

Table 3-1 Overview of the genes under selection for glyphosate resistance that are involved in the process of detoxification, environmental sensing, and stress signaling and response.

Chapter 4

One Hundred Years of Deciphering Genetic Correlations -- Lessons Learned and Future Perspectives

Abstract

Deciphering the genetic basis of trait correlations is crucial for understanding trait evolution and adaptation. Most of the effort towards understanding genetic correlations have made use of phenotypic data from a pedigree of individuals. With the advent of sequencing technologies, the focus is moving towards analytical methods to identify the specific genetic variants and thus the mechanisms underlying trait correlations. Here, we first review the phenotypic and molecular marker methods that are commonly used to estimate the strength of genetic correlations (magnitude and direction), and to deconstruct the underlying genetic mechanism (pleiotropy vs linkage). Next, we discuss the pitfalls associated with these methods and provide an outline of strategies that can address some of these pitfalls. We believe an integrative approach, combining phenotypic studies with genetic marker data and molecular validation tools, is required to shed light on the genetic architecture of trait correlations. A deeper understanding of genetic correlations will aid in understanding their evolutionary potential and further our understanding of how pervasive pleiotropy or linkage is in shaping the evolution of traits.

Introduction

Trait correlations -- covariation between two or more phenotypic traits -- are pervasive in nature and may act to constrain the adaptive evolution of natural populations (Roff, 1996; Conner *et al.*, 2011). Genetic correlations between traits are also crucial to breeding programs attempting to improve crops (Falconer, 1996), play a central role in life-history theory (Via & Lande, 1985; Barton & Turelli, 1989), and may facilitate the adaptation of invasive species introduced into new or changing environments (Hämälä *et al.*, 2020; Dutta *et al.*, 2021). Trait correlations are one of the most commonly studied types of evolutionary constraints (Conner *et al.*, 2011), with their magnitude, direction, and genetic basis all of primary interest given these factors are predicted to influence how correlated traits will evolve over time (Lande, 1979; Cheverud, 1982, 1984; Falconer, 1996). Despite their theoretical importance, however, a rich literature shows that the influence of trait correlations on evolutionary dynamics is complex, with evidence that correlations can constrain, facilitate, or even have little effect on trait evolution (Agrawal & Stinchcombe, 2009).

This apparent complexity is likely due, at least in part, to the stability of genetic correlations. The persistence of correlations between traits and their evolution is influenced by selection (Roff, 2000; Jones *et al.*, 2003; Revell, 2007; Arnold *et al.*, 2008; Chantepie & Chevin, 2020; Svensson *et al.*, 2021), and can also be altered due to mutations, genetic drift and founder effects (Jones *et al.*, 2003; Steven *et al.*, 2020) in addition to the genetic background and environment (Saltz *et al.*, 2017b; Geiler-Samerotte *et al.*, 2020; Mitchell & Houslay, 2021). Importantly, the genetic basis of trait correlations plays a major role in the persistence of the correlation. In the simplest scenario, correlations due to linkage disequilibrium (LD) can be expected to be transient since, assuming no correlated selection, recombination should remove the linked variants rather rapidly whereas genetic correlations due to pleiotropy are expected to persist over longer time scales (Lande & Arnold, 1983; Futuyma, 1986; Endler, 1986; Lynch *et al.*, 1998; Falconer & Mackay, 2009). A number of studies have successfully used crossing designs, artificial selection regimes, and QTL mapping to determine the causal mechanism underlying trait correlations (Fenster & Carr, 1997; Beldade *et al.*, 2002; Conner, 2002; Conner *et al.*, 2011; Delph *et al.*, 2011).

However, as we show below, pleiotropy can arise from both direct and indirect effects of variants on traits, and can likewise be influenced by an intermediate genetic factor (Stearns, 2010; Wagner & Zhang, 2011; Paaby & Rockman, 2013). The resolution achieved from crossing designs, artificial selection, and QTL studies cannot differentiate these possibilities. Further, in some scenarios (*i.e.* inversions), pleiotropy and linkage may effectively be one and the same, suggesting that our expectations for trait evolution in those situations should be modified (Saltz *et al.*, 2017a). A more precise understanding of the mechanism of correlation--*i.e.*, down to the particular genetic variants involved--would likely shed light on long term stability of association and perhaps help us understand how certain combinations of evolutionary forces and trait associations lead to constraints whereas others lead to facilitation. There are a number of analytical methods based on GWAS that we can employ to determine if pleiotropy or linkage may underlie a trait correlation, and, if the correlation is due to pleiotropy, further clarify the type of pleiotropy (*i.e.* biological vs mediated pleiotropy). These methods may be of interest to evolutionary biologists, but their use remains limited.

Here, we first describe the genetic mechanisms hypothesized to underlie trait correlations -- pleiotropy and linkage -- and then review the phenotypic and genetic methods commonly used to estimate the magnitude and direction of genetic correlations. Next, we outline analytical strategies that have been implemented for teasing apart the underlying mechanistic basis of trait correlations. We then discuss the pitfalls associated with these methods and suggest strategies that can address some of these issues. We end this review with a path forward for integrating information from phenotypic family-level data, genetic marker data (GWAS, omics-QTL), and molecular validation tools to shed light on the genetic architecture of trait correlations. Increased precision in our understanding of the genetics of trait correlations will aid predictions on how traits may be expected to evolve, and more importantly will broaden our understanding of the pervasiveness of pleiotropy versus linkage in shaping current patterns of the phenotypic diversity that we see today.

I. How do correlations arise?

Genetic correlations among quantitative traits can arise due to pleiotropy (biological or mediated) or linkage disequilibrium (LD) (Figure 4-1). Biological pleiotropy, also referred to as horizontal pleiotropy, is defined as a causal variant (or gene) that independently influences two traits, either directly or indirectly via an intermediate phenotype (Figure 4-1a-c). In contrast, mediated pleiotropy, also referred to as vertical pleiotropy, is when a genetic variant directly alters one trait which then causally influences the other trait, *i.e.* the genetic variant associated with the first trait is indirectly associated with the second (Figure 4-1d-e). Additionally, genetic correlations can also arise when the causal variant (or gene) for the two traits are physically linked and will thus be inherited together (linkage disequilibrium, Figure 4-1f). Although biological pleiotropy and LD have a rich literature, our understanding of mediated pleiotropy remains limited (Auge *et al.*, 2019). Mediated pleiotropy can be invoked whenever a trait lies in the causal path of the other trait and as such traits within the same functional pathway diverging to two phenotypic traits, or a trait that regulates multiple pathways are more likely to be correlated via mediated pleiotropy.

Delineating pleiotropy from linkage is important due to their inherently different evolutionary fates. A trait correlation arising from pleiotropy is expected to persist over generations -- it is likely that neutral evolutionary processes like drift will not break down the trait associations (Lande & Arnold, 1983; Conner, 2002) since a common functional mechanism generates the correlation. Moreover, natural selection has the potential to lead to the coevolution of the correlated pleiotropic traits, even when the traits are maladaptive (the case of antagonistic pleiotropy) (Lande & Arnold, 1983). In contrast, trait correlations arising from LD are expected to decay rapidly due to recombination and segregation, and thus are more transient (Futuyma, 1986; Falconer & Mackay, 2009). Alternatively, the presence of strong correlational selection could balance out the recombination, especially if the genes are closely linked, making the trait correlations due to linkage less transient (Hartl *et al.*, 1997; Lynch *et al.*, 1998). Taken together, genetic correlation arising due to pleiotropy has a higher potential for impacting the course of evolution of genetically correlated traits.

II. Estimating genetic correlations and inferences on causality using phenotypic data

In the simplest sense, estimation of the genetic correlation coefficient can be thought of as the analysis of variance and covariance for two traits, and ANOVA and REML are routinely utilized for this purpose (see examples in Table 4-1). In experimental and natural populations, genetic correlations have traditionally been estimated using family-level phenotypic data, focusing mainly on parent-offspring, half- or full-sib families (see examples in Table 4-1). The choice of experimental and family design is crucial in such cases since it directly affects the estimation of genetic correlations. For example, inbreeding (Rose, 1984) and the presence of a common stressful environment (Service, Philip M. & Rose, 1985; Kasule, 1991) might introduce biases. Additionally, it was shown that genetic correlation estimates using full-sib data lead to more positive estimates due to the presence of additive and dominance variances (Hill, 2013), and also the presence of maternal effects in such families (Falconer 1996; Hunt and Simmons 1998). To address this, an increasing number of studies have begun employing (Fox, 1998; Fox *et al.*, 2004; Hallsson & Björklund, 2012) mixed family designs like a nested paternal half-sib design. Each design and analysis method incurs its own set of benefits and drawbacks (see Table 4-2). For example, methods based on sampling variance of phenotypic data have been shown to be heavily biased for small sample sizes (Reeve, 1955; Robertson, 1959; Grossman, 1970). In contrast, methods like bootstrapping and jackknifing (Roff & Preziosi, 1994; Reale & Roff, 2001) have been shown to provide a lower SE, and thus might be more favorable over other methods. This highlights the complexity involved in estimating a reliable genetic correlation estimate which is required to increase the predictability of the correlated response to selection.

As discussed above, an important aspect of investigating the presence of genetic correlations is identifying its underlying genetic mechanism -- biological vs mediated pleiotropy vs linkage disequilibrium (LD). One way of teasing apart the causality of trait correlation using phenotypic data is to test for the consistency of the genetic correlation structure between traits -- if it is consistent among multiple populations of the species, it is indicative of pleiotropy. Fenster and Carr (1997) applied this logic to inter- and intra-specific crosses between multiple populations of *Mimulus* and concluded that linkage is responsible for the positive genetic correlation between pollen and ovule production. Using different populations for evaluating the differences between

genetic correlation structures is inherently problematic, however, as the differences in the environment can directly contribute to the differences in genetic correlation structure.

Other studies have used multi-generation approaches for testing the presence of linkage between correlated traits (Conner, 1997; Beldade *et al.*, 2002; Conner *et al.*, 2003, 2011). In multi-generational studies, one would expect the magnitude of genetic correlation to decline after multiple generations of recombination if the genetic correlation is solely due to linkage disequilibrium (recombination will weaken linkage), and for the genetic correlation to be the same in case of pleiotropy. Ideally, the study design should use a random-mating (to minimize LD being maintained by selection) and should be carried for a minimum of eight to ten generations, to confidently differentiate between pleiotropy and linkage (Conner, 1997; Conner *et al.*, 2003, 2011). Beldade and colleagues (Beldade *et al.*, 2002) similarly performed a ten-generation artificial selection study and found that the correlation among wing spots quickly decoupled under selection in *Bicyclus anynana* butterflies, indicating LD as the underlying correlation mechanism. Interestingly, similar logic was recently applied to correlated behavioral traits in domesticated dogs (modern vs ancient) and found that the traits temporally decouple due to the varying selection pressures during the domestication process (Hansen Wheat *et al.*, 2019). These methods are again incapable of differentiating biological vs mediated pleiotropy. Additionally, multi-generational studies have limited power to detect tight linkage, unless a *very* large segregating population is available, creating which can either be very time-intensive or not possible depending on the species.

Using phenotypic data for deconstructing the mechanistic basis of the genetic correlation -- i.e., pleiotropy vs linkage -- is very challenging (Brooks, 2000). Although feasible, the caveats of differentiating between pleiotropy and linkage using phenotypic data are that these methods (1.) are very time-intensive, (2.) provide no insights into the identity of the causal genes underlying the traits, (3.) can only work for genes that are not tightly linked to each other, and (4.) are only possible when the loci under consideration are not located within a coldspot recombination region and/or within structurally rearranged regions (regions with inherent low recombination).

III. Assessing genetic correlations at the molecular level

With the advent of sequencing technologies, estimation of genetic correlations lately has largely depended upon large-scale genetic data, like those available in GWAS datasets (see examples in Table 4-3). A major advantage of using GWAS for the estimation of genetic correlations is that it does not require the individuals to be related and the phenotypes of interest can be measured on different individuals. Together, these enable the inclusion of larger and more diverse samples. Although there is a rich human literature estimating genetic correlation coefficient using GWAS datasets (van Rheenen *et al.*, 2019), this is limited in other species, especially in the plant literature (see Table 4-3). Broadly, the methods that employ GWAS data to estimate the genetic correlation can be classified into two categories -- one that makes use of the individual-level data and another that uses the GWAS summary statistic (see Box 4-1 for an overview of methods).

Korte and colleagues (Korte *et al.*, 2012) developed a multi-trait mixed model (MTMM), to estimate genetic correlation among traits using individual-level data. Using this method, they showed that the blood metabolites involved in cardiovascular diseases are significantly genetically correlated. This method has also been applied to *Arabidopsis thaliana* (Thoen *et al.*, 2017), where it was shown that stress responses are highly correlated -- stress response that induces the salicylic acid pathway (parasitic plants, aphids) is negatively correlated with the response that induces jasmonic acid pathways (necrotrophic fungi, thrips). Another routinely used method to estimate trait correlations using individual-level GWAS data is GREML (Genome-based Restricted Maximum Likelihood). This method has been widely applied to a suite of traits (Table 4-3). For instance, human height and body mass index between men and women have been shown to be highly positively correlated (Yang *et al.*, 2015), and this has recently been shown to be consistent across different continental populations (Guo *et al.*, 2021). Studies in cattle have found a positive genetic correlation between milk and protein yield (Calus *et al.*, 2018) and more interestingly, have found that the resistance to different bovine pathogens are all strongly correlated (both positively and negatively) (Mahmoud *et al.*, 2018). Although accurate and powerful, the individual-level approach can seldom be used due to the lack of availability of individual-level data (since not all studies publish the individual-level data) and a large computational power requirement.

In contrast to individual-level methods, the GWAS summary statistic method for estimating genetic correlation is routinely applied, especially to human traits (Table 4-3). This is in part due to the wide availability of GWAS summary statistics, and the low computational requirements of the method.

One of the most commonly used methods is Linkage Disequilibrium Scores Regression (LDSR) which has been applied to an array of human traits (Zheng *et al.*, 2017; Watanabe *et al.*, 2019; Zhang *et al.*, 2021). Interestingly, using LDSR on 17 traits in humans has provided support for Chevrud's conjecture (Cheverud, 1988) which states that phenotypic correlations are a good estimate of genetic correlations (Sodini *et al.*, 2018). Recently, LDSR was also applied to cattle where it was found that milk, fat and protein yield are all significantly positively correlated, but these have a negative correlation with mastitis resistance (Cai *et al.*, 2020). To date though, no studies in the plant literature have used LDSR to estimate the genetic correlations between traits, despite over 1000 crop GWAS studies published in the last decade (Liu & Yan, 2019). Although LDSR can be used readily, researchers should be aware of the limitations of this method given the high potential standard error associated with the estimation (see Box 4-1).

Box 4-1: Brief overview of the available methods for estimating genetic correlation coefficient using GWAS data.

Methods utilizing individual-level data: Multi-trait mixed model (MTMM), was developed by Korte and colleagues (Korte *et al.*, 2012) which is an extension of the mixed models used for association mapping in GWAS and can be used to estimate genetic correlation among traits using individual-level data. Their model considers both within and between-trait variance components for multiple traits simultaneously, thus partitioning the genetic and environmental covariances influencing the traits. Briefly, it involves building a GRM (Genomic Relationship Matrix), that describes the variance-covariance structure of the SNPs which are then used to estimate the contribution of SNPs to phenotypic variance. For a comprehensive overview of, advantages and pitfalls of MTMM see Yang *et al.*, (2014). Another method utilizing individual-level data is a bivariate GREML (Genome-based Restricted Maximum Likelihood) (Yang *et al.*, 2011a; Lee *et al.*, 2012). Briefly, this method also constructs a GRM but estimates genetic correlation via

LMM (Linear Mixed Model), in which GRM is used to model phenotypes as a function of the genotype of the individuals to estimate a statistic called SNP-based heritability. Caution must be taken when interpreting genetic correlations using this method since SNP-based heritability aims to capture only the causal variants that are in LD with the measured SNPs (van Rheenen *et al.*, 2019), and as such is always underestimated. Additionally, it is important to note that the genetic correlation estimate is associated with errors of three estimated parameters -- the two heritability values and the genetic correlation itself and thus the SE of estimated genetic correlation can thus be high, especially if the SNP density is low (<500k SNPs) (Ni *et al.*, 2018).

Methods utilizing summary statistics: Linkage disequilibrium scores regression (LDSR) was the first method developed that made use of GWAS summary statistics to estimate genetic correlation, and extensions have been made since (Bulik-Sullivan *et al.*, 2015b,a). In the LDSR approach, the association test statistic of a SNP is regressed on their LD scores, which is the sum LD r^2 measured with all other SNPs. The method is based on the observation that variants in LD with a causal variant show a higher association test statistic, which is proportional to the strength of the linkage (Pritchard & Przeworski, 2001; Yang *et al.*, 2011b). Using this LD score for a SNP and the association statistic of the SNP for the two traits of interest, SNP-based heritabilities, and genetic correlation can be estimated. Since this method uses LD between SNPs, it also requires a reference panel with the same LD structure to adjust for linkage in the population, ideally from the same population (Bulik-Sullivan *et al.*, 2015a). The further the reference panel's LD structure is from those of the sample, the higher would be the standard error associated with the estimated genetic correlation (Ni *et al.*, 2018). Additional sources like the inherent loss of information in the GWAS summary statistic, low SNP density, and genomic partitioning can further lead to increased standard errors. Another caveat of the LDSR methodology is the key assumption that the allele frequency differences between subpopulations used in GWAS are independent of the LD scores (Bulik-Sullivan *et al.*, 2015a), but linked selection (like background selection) can lead to a correlation between them (Berg *et al.*, 2019). Thus, although LDSR can be used readily, its reliability is still questionable.

IV. Inferences on causality using genetic data

There is a rich literature of studies that have shown that pleiotropy is a pervasive underlying cause of genetic correlation among traits. For example, in a review of QTL for 238 trait correlations, (Gardner & Latta, 2007) found that correlated traits share on average two QTLs, and thus are interpreted as pleiotropic QTLs. Another meta-analysis of 558 GWAS (representing 558 unique traits) (Watanabe *et al.*, 2019), found that 90% of the identified loci were pleiotropic among two or more traits. A major caveat of these studies (and other studies examining causality of genetic correlations) is the definition of pleiotropy, which is defined as the loci that are significantly associated with more than one trait. This is problematic as it does not exclude the presence of linkage between the two trait SNPs -- when loci for two traits are closely linked, the loci would be identified as a statistically significant association for both traits. It has been proposed that multitrait analysis in GWAS can identify real pleiotropy, but Fernandes and colleagues (Fernandes *et al.*, 2020) have recently shown using simulations that this is not true and LD often gets misclassified as pleiotropy using these methods. Since most QTL and GWAS studies do not have a high resolution, they cannot entirely differentiate between pleiotropy and linkage (Gardner & Latta, 2007). In some cases though, GWAS studies with high SNP density can distinguish between pleiotropy and loose linkage, but tight linkage vs pleiotropy can still not be teased apart. Thus, we still are lacking a clear picture of how often pleiotropy or linkage gives rise to genetic correlations.

One of the most commonly used methods to detect mediated pleiotropy is Mendelian Randomization (MR). MR was initially developed to improve the inference of causality in epidemiology (Davey Smith & Ebrahim, 2003) but has been extended to identify the presence of mediated pleiotropy. Briefly, MR tests whether trait 1 causally affects trait 2 by testing whether the genetic variant of trait 1 is also the genetic variant for trait 2, in the absence of horizontal pleiotropy (see Box 4-2). This has been applied to multiple disease traits in humans (Hartwig *et al.*, 2016; Sun *et al.*, 2019; Choi *et al.*, 2019; Qian *et al.*, 2020; Wu *et al.*, 2020). Interestingly, it was shown using MR that mediated pleiotropy exists between type 2 diabetes and hypertension -- type 2 diabetes causally increases the risk of hypertension (Sun *et al.*, 2019). Although MR has been applied in the plant literature, these have been limited to identifying whether the identified

variants causally influence the trait (Liu *et al.*, 2020; Su *et al.*, 2021), and not to test the presence of mediated pleiotropy between traits. A major caveat to applying MR is the a priori knowledge of the genetic variant of a trait, which in most cases is not known. Also, MR makes the assumption that the traits being tested do not have any direct causal variants in common, which is violated in the presence of horizontal pleiotropy. Further, it is important to note that MR methods cannot distinguish between tight linkage and pleiotropy. Thus, the use of MR remains limited at best.

Bolormaa and colleagues (Bolormaa *et al.*, 2014) used a novel approach to test between pleiotropy and linkage. First, using a multi-trait meta-analysis method on GWAS summary statistics they identified the SNPs associated with two or more traits (genetically correlated variants). Multi-trait analysis has previously been shown to increase power in marker detection (Korol *et al.*, 2001; Turley *et al.*, 2018). Then to test for pleiotropy vs linkage, they fit the most significant SNPs in a region associated with two traits in a regression model and reperformed GWAS. If other SNPs in the region were no longer significantly associated with the traits, pleiotropy is most likely the underlying genetic mechanism, whereas, if other SNPs in the region are still significantly associated with the traits, then linkage is at play. They applied this new method to identify groups of traits that are pleiotropic in Beef Cattle (Bolormaa *et al.*, 2014). Although quite elegant, the usage of this method depends on the identification of the genetically correlated variants, which can have high error rates due to the loss of information associated with GWAS summary statistics usage. Additionally, this method cannot distinguish between biological vs mediated pleiotropy.

Box 4-2: Mendelian Randomization to differentiate between biological and mediated pleiotropy.

Mendelian randomization is based on the intuition that if trait 1 influences trait 2, then a genetic variant that influences trait 1 is also expected to influence trait 2. A genetic variant of trait1 can then be used to determine if trait 1 causally influences trait 2, thus testing whether the association between trait 1 and trait 2 is due to mediated pleiotropy. To perform MR, three assumptions regarding the genetic variant (causal variant of trait 1) must hold -- the variant causally

influences trait 1, the variant is not associated with confounding variables (like environmental variables, etc), and the variant influences trait 2 *only* through trait 1, and not through other traits or even directly (absence of true biological pleiotropy). The second assumption is easily violated in the presence of population stratification, assortative mating, and other indirect genetic effects (VanderWeele *et al.*, 2014). The first and the third assumption implies a priori knowledge of a causal variant of trait1 and sufficient knowledge of the pathways trait 1 and trait 2 are involved in, respectively, which is not always available (Lawlor *et al.*, 2008). Further, due to the widespread pleiotropy among complex traits, the assumption of the absence of biological pleiotropy is not always met, in which case MR is highly biased and cannot be relied upon. Although methods have been developed to correct for biological pleiotropy (Zhu *et al.*, 2018; Verbanck *et al.*, 2018), their effectiveness again depends upon extensive prior knowledge of the confounding pathways the traits are involved in and are very sensitive to confounding factors (Hemani *et al.*, 2018; Koellinger & de Vlaming, 2019).

V. Multi-faceted approach for the future

As discussed above, the current methodologies available for inferring the causality of genetic correlations are limited in their scope -- just using either phenotypic or genetic data have proven to be insufficient for discerning the genetic basis of genetic correlations, and the role of pleiotropy (biological/mediated) and LD in quantitative traits. Here, we discuss some of the potential flaws that are inherent with these methods and suggest ways forward.

Caution should be taken when using genetic data (GWAS) to deconstruct the genetic architecture of quantitative traits since GWAS studies have inherent pitfalls that can lead to false positives and/or overestimation of the importance of the significant loci (for comprehensive reviews, see Josephs *et al.*, 2017; Young *et al.*, 2019; Tam *et al.*, 2019). GWAS studies are infamous for only identifying common variants with higher effect size -- rare causal variants are particularly difficult to identify. To address this, GWAS studies can be combined with other approaches like whole-exome GWAS or haplotype-based association mapping. Haplotype-based association mapping focuses on haplotype blocks rather than single SNPs like in GWAS, and thus can be

used to reveal the complex mechanism of causal haplotypes. Since common SNPs can combine to form rare haplotypic variants, this method offers a higher power of detecting rare causal variants. Depending on the size of the haplotype blocks though, association mapping of genome-wide haplotypes can be very computationally intensive. A round-about way to do this might be first performing GWAS and then performing haplotype-based association mapping only for the significant regions identified via GWAS. Further GWAS studies are often biased due to the presence of either confounding factors (population structure, environment, etc.) and/or indirect effects (Young *et al.*, 2019). Although methods have been developed to incorporate population structure as a covariate, the results can still be biased (Sul *et al.*, 2018). To address confounding factors, especially environment, one should ideally use an experimental population that is artificially designed to control for allele frequency differences. This would also address any biases that may result from the choice of the control group. Importantly, a traditional GWAS study of unrelated individuals estimates the combined effects of direct (phenotype directly altered by genotype) and indirect genetic effects (phenotype altered due to the gene expression of parents and close relatives). To tease apart these effects, GWAS with family data can be performed, as the presence of family-level data will enable the estimation of any indirect effects that may be present which will allow us to identify unbiased causal marker regions associated with the trait(s) of interest.

Identifying the underlying genetic architecture of the correlated traits has been particularly challenging in the past, but taking an integrative approach to identify the mechanistic basis of genetic correlation would be most insightful (Figure 4-2). There are multiple integrative approaches one can take depending on the species characteristics and the availability of data. For example, if the species under consideration has a short generation time and is easy to breed, it might be preferable to perform a multi-generational study (with random mating or a large segregating population) in a single environment (to reduce noise and confounding factors) to differentiate between the presence of linkage and pleiotropy (Figure 4-2a). Since this would still not be able to differentiate biological vs mediated pleiotropy, and even does not discount the presence of tight linkage, supplementing this with either molecular validation tool or genetic marker data (like GWAS) would provide a deeper understanding of the mechanisms underlying the trait correlation (Figure 4-2b-e).

For many species, creating crosses and/or controlling for the environment is not feasible, and thus one has to rely on the genetic markers data to disentangle the genetic architecture of trait correlations. For example, performing two single trait GWAS would help locate the causal loci for trait correlations and can further help disentangle between linkage and pleiotropy (Figure 4-2b). One can use association studies to gain insight into the region underlying trait correlations, but often cannot identify the underlying causal variants due to the presence of linkage between SNPs (synthetic associations). Thus, an association study almost always has to be followed by fine-mapping which takes into account the LD structure of the SNPs and thus reduces the potential causal SNPs underlying trait correlation at the gene or variant level (Figure 4-2c). Additionally, fine mapping can also prevent misclassification of local LD as pleiotropy -- if two distinct causal variants are detected for the traits it would indicate the presence of linkage between the two traits, whereas if the same causal variant is detected it would be indicative of pleiotropy (Figure 4-2c). To note though, fine mapping rarely leads to the identification of *one* causal variant but rather gives a set of potential causal variants (which are a subset of that identified via association mapping). Thus, to distinguish between linkage vs pleiotropy, one can perform a colocalization test, which is a statistical test to determine whether the overlap between SNPs for the two traits is significant or not. Pleiotropy can be suggested if the SNPs for the two traits colocalize; if no colocalization is observed linkage is more likely (Figure 4-2c).

Further, integrating this with functional validation methods like omic-QTL (eQTL, pQTL, meQTL, etc.) data, can help to differentiate between biological and mediated pleiotropy and to test the effect of the variant on the two traits. For example, eQTL which is used to test whether the variant changes the expression of the gene of interest can be employed. In a scenario of a correlation between two traits which exhibit the same GWAS variants, but different eQTLs, biological pleiotropy would be indicated (Figure 4-2d). In comparison, a correlation between two traits which exhibit the same GWAS and same eQTL variants, biological pleiotropy through the action of intermediate phenotype, or mediated pleiotropy would be indicated (Figure 4-2d). In this case if one has a priori knowledge, MR can be used to test for mediated pleiotropy. If no colocalization occurs though, other omics-QTL can be performed to understand the functional underpinning of the genetic correlations. For example, pQTL can be employed

wherein a colocalization of GWAS would indicate the variant influences the traits via changes in the protein levels. Lastly, in species with genome-editing tools available, various molecular validation methods like gene editing (CRISPR/Cas9), overexpression, knockout, and cloning can also be used to confirm the findings and understand the mechanistic basis of traits correlations (Figure 4-2e).

Conclusions

Knowledge of the genetic basis of trait correlations is important for understanding the stability of trait correlations, and their evolutionary potential -- will a trait correlation facilitate or constrain evolution? Although phenotypic pedigree data have made advancements towards understanding the genetic mechanisms underlying trait correlations, these have limitations. Over the last decade, great strides have been made towards identifying the causal gene of trait correlations by making use of marker-assisted data (mainly GWAS), but these studies lack a deconstruction of the underlying mechanism (biological vs mediated pleiotropy or linkage). By outlining an integrative approach to comprehensively study genetic correlations, we hope to foster a deeper analysis of genetic causes of trait correlations. These approaches can be applied depending on the species, *a priori* knowledge, and the availability of genome editing tools for the species. If followed serially, the suggested analysis approach not just be able to identify causal loci, but also identify the functional mechanism through which the variant alters the traits. This would further help us understand how pervasive pleiotropy and linkage may respectively be in shaping trait correlations and predict more precisely how these correlations will evolve over time.

Glossary

Traits: Phenotypes or measurements that can be either dichotomous (presence/absence) or quantitative (range of values).

Pleiotropy: The phenomenon wherein a single variant or allele influences two or more phenotypic traits, either directly or indirectly.

Linkage: The physical state of loci, typically on the same chromosome, being linked.

Linkage Disequilibrium: The phenomenon wherein loci that are physically close on the genome (typically on the same chromosome) are inherited together, thus creating a statistical non-random association of these loci.

Causal variant: Genetic variants that are responsible for the trait, and in the context of association studies, the genetic variants that are responsible for the association signal at a locus.

Genome-Wide Association Studies (GWAS): A statistical approach utilizing linkage disequilibrium to identify the genetic basis of a trait variation. Briefly, the method looks for an association between the genotype and the phenotypic value.

Quantitative Trait Loci (QTL): A genomic region of DNA that is statistically significantly associated with a phenotype of interest. Many QTLs are typically associated with a single trait and consequently, each QTL varies in length and the degree to which it can explain the phenotypic variation (effect size).

omic-QTL: Extension of QTL to other 'omic variations' like transcriptomics (eQTL), proteomics (pQTL), metabolomics (mQTL), epigenomics (histone modification QTL (hmQTL), methylation QTL (meQTL)), to characterize the underlying molecular and functional mechanism controlling the phenotype.

Acknowledgments

I would like to thank Tanmoy Rouchowdhury for his valuable insights on the marker-assisted techniques.

References

Agrawal AF, Stinchcombe JR. 2009. How much do genetic covariances alter the rate of adaptation? *Proceedings. Biological sciences / The Royal Society* 276: 1183–1191.

Åkesson M, Bensch S, Hasselquist D. 2007. Genetic and phenotypic associations in morphological traits: a long term study of great reed warblersAcrocephalus arundinaceus. *Journal of avian biology* 38: 58–72.

Åkesson M, Bensch S, Hasselquist D, Tarka M, Hansson B. 2008. Estimating heritabilities and genetic correlations: Comparing the 'animal model' with parent-offspring regression using data from a natural population. *PloS one* 3: e1739.

Arnold SJ, Bürger R, Hohenlohe PA, Ajie BC, Jones AG. 2008. UNDERSTANDING THE EVOLUTION AND STABILITY OF THE G-MATRIX. *Evolution* 62: 2451–2461.

Auge GA, Penfield S, Donohue K. 2019. Pleiotropy in developmental regulation by flowering-pathway genes: is it an evolutionary constraint? *The New phytologist* 224: 55–70.

Barton NH, Turelli M. 1989. Evolutionary quantitative genetics: how little do we know? *Annual review of genetics* 23: 337–370.

Bégin M, Roff DA. 2001. An analysis of G matrix variation in two closely related cricket species, Gryllus firmus and G. pennsylvanicus. *Journal of evolutionary biology* 14: 1–13.

Beldade P, Koops K, Brakefield PM. 2002. Modularity, individuality, and evo-devo in butterfly wings. *Proceedings of the National Academy of Sciences of the United States of America* 99: 14262–14267.

Bemmels JB, Anderson JT. 2019. Climate change shifts natural selection and the adaptive potential of the perennial forb Boechera stricta in the Rocky Mountains. *Evolution; international journal of organic evolution* 73: 2247–2262.

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, *et al.* 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* 8: e39725.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24: 127–135.

Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, Tier B, Savin K, Hayes BJ, Goddard ME. 2014. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS genetics* 10: e1004198.

Brooks R. 2000. Negative genetic correlation between male sexual attractiveness and survival. *Nature* 406: 67–70.

Browne WJ, Draper D. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* 1: 473–514.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, Duncan L, Perry JRB, Patterson N, Robinson EB, *et al.* 2015a. An atlas of genetic correlations across human diseases and traits. *Nature genetics* 47: 1236–1241.

Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson N, Daly MJ, Price AL, Neale BM. 2015b. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* 47: 291–295.

Cai Z, Dusza M, Guldbrandtsen B, Lund MS, Sahana G. 2020. Distinguishing pleiotropy from linked QTL between milk production traits and mastitis resistance in Nordic Holstein cattle. *Genetics, selection, evolution: GSE* 52: 19.

Calus MPL, Goddard ME, Wientjes YCJ, Bowman PJ, Hayes BJ. 2018. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. *Journal of dairy science* 101: 4279–4294.

Chantepie S, Chevin L-M. 2020. How does the strength of selection influence genetic correlations? *Evolution letters* 4: 468–478.

Cheverud JM. 1982. Phenotypic, genetic, and environmental morphological integration in the cranium. *Evolution; international journal of organic evolution* 36: 499–516.

Cheverud JM. 1984. Quantitative genetics and developmental constraints on evolution by selection. *Journal of theoretical biology* 110: 155–171.

Cheverud JM. 1988. A COMPARISON OF GENETIC AND PHENOTYPIC CORRELATIONS. *Evolution; international journal of organic evolution* 42: 958–968.

Choi KW, Chen C-Y, Stein MB, Klimentidis YC, Wang M-J, Koenen KC, Smoller JW, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. 2019. Assessment of Bidirectional Relationships Between Physical Activity and Depression Among Adults: A 2-Sample Mendelian Randomization Study. *JAMA psychiatry* 76: 399–408.

Conner JK. 1997. Floral Evolution in Wild Radish: The Roles of Pollinators, Natural Selection, and Genetic Correlations Among Traits. *International journal of plant sciences* 158: S108–S120.

Conner JK. 2002. Genetic mechanisms of floral trait correlations in a natural population. *Nature* 420: 407–410.

Conner JK, Franks R, Stewart C. 2003. Expression of additive genetic variances and covariances for wild radish floral traits: comparison between field and greenhouse environments. *Evolution; international journal of organic evolution* 57: 487–495.

Conner JK, Karoly K, Stewart C, Koelling VA, Sahli HF, Shaw FH. 2011a. Rapid independent trait evolution despite a strong pleiotropic genetic correlation. *The American naturalist* 178: 429–441.

Davey Smith G, Ebrahim S. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International journal of epidemiology* 32: 1–22.

Debes PV, Solberg MF, Matre IH, Dyrhovden L, Glover KA. 2021. Genetic variation for upper thermal tolerance diminishes within and between populations with increasing acclimation temperature in Atlantic salmon. *Heredity* 127: 455–466.

Delph LF, Steven JC, Anderson IA, Herlihy CR, Brodie ED 3rd. 2011. Elimination of a genetic correlation between the sexes via artificial correlational selection. *Evolution; international journal of organic evolution* 65: 2872–2880.

Dutta A, Hartmann FE, Francisco CS, McDonald BA, Croll D. 2021. Mapping the adaptive landscape of a major agricultural pathogen reveals evolutionary constraints across heterogeneous environments. *The ISME journal* 15: 1402–1419.

Endler JA. 1986. *Natural Selection in the Wild. (MPB-21)*. Princeton University Press.

Falconer DS. 1996. *Introduction to Quantitative Genetics*. Pearson Education.

Falconer, D.S. (1981) Introduction to Quantitative Genetics. 2nd Edition, Longman Group Ltd., London, 1-133. - References - Scientific Research Publishing.

Falconer DS, Mackay TFC. 2009. *Introduction to Quantitative Genetics*. Pearson.

Fenster CB, Carr DE. 1997. Genetics of sex allocation in Mimulus (Scrophulariaceae). *Journal of evolutionary biology*.

Fernandes SB, Zhang KS, Jamann TM, Lipka AE. 2020. How Well Can Multivariate and Univariate GWAS Distinguish Between True and Spurious Pleiotropy? *Frontiers in genetics* 11: 602526.

Fox CW. 1998. Genetic and Maternal Influences on Body Size and Development Time in the Seed Beetle Stator limbatus (Coleoptera: Bruchidae). *Annals of the Entomological Society of America* 91: 128–134.

Fox CW, Czesak ME, Wallin WG. 2004. Complex genetic architecture of population differences in adult lifespan of a beetle: nonadditive inheritance, gender differences, body size and a large maternal effect. *Journal of evolutionary biology* 17: 1007–1017.

Futuyma DJ. 1986. *Evolutionary Biology*. Sinauer Associates.

Gao Y-H, Jiang Q-W, Meng W, Chen Q-M, Zhao N-Q, Shen F-M. 2003. Comparison of REML and MCMC for Genetic Variance Component Model of Quantitative Trait in Nuclear Family [J]. *Shanghai yixue jianyan zazhi = Shanghai journal of medical laboratory sciences* 4.

Gardner KM, Latta RG. 2007. Shared quantitative trait loci underlying the genetic correlation between continuous traits. *Molecular ecology* 16: 4195–4209.

Geiler-Samerotte KA, Li S, Lazaris C, Taylor A, Ziv N, Ramjeawan C, Paaby AB, Siegal ML. 2020. Extent and context dependence of pleiotropy revealed by high-throughput single-cell phenotyping. *PLoS biology* 18: e3000836.

Grossman M. 1970. Sampling variance of the correlation coefficients estimated from analyses of variance and covariance. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 40: 357–359.

Guo J, Bakshi A, Wang Y, Jiang L, Yengo L, Goddard ME, Visscher PM, Yang J. 2021. Quantifying genetic heterogeneity between continental populations for human height and body mass index. *Scientific reports* 11: 5240.

Hallsson LR, Björklund M. 2012. Sex-specific genetic variances in life-history and morphological traits of the seed beetle Callosobruchus maculatus. *Ecology and evolution* 2: 128–138.

Hämälä T, Gorton AJ, Moeller DA, Tiffin P. 2020. Pleiotropy facilitates local adaptation to distant optima in common ragweed (Ambrosia artemisiifolia). *PLoS genetics* 16: e1008707.

Hangartner S, Lasne C, Sgrò CM, Connallon T, Monro K. 2020. Genetic covariances promote climatic adaptation in Australian Drosophila. *Evolution; international journal of organic evolution* 74: 326–337.

Hansen Wheat C, Fitzpatrick JL, Rogell B, Temrin H. 2019. Behavioural correlations of the domestication syndrome are decoupled in modern dog breeds. *Nature communications* 10: 2422.

Hartl DL, Clark AG, Clark AG. 1997. *Principles of population genetics*. Sinauer associates Sunderland, MA.

Hartwig FP, Bowden J, Loret de Mola C, Tovo-Rodrigues L, Davey Smith G, Horta BL. 2016. Body mass index and psychiatric disorders: a Mendelian randomization study. *Scientific reports* 6: 32730.

Hemani G, Bowden J, Davey Smith G. 2018. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human molecular genetics* 27: R195–R208.

Hill WG. 2013. On estimation of genetic variance within families using genome-wide identity-by-descent sharing. *Genetics, selection, evolution: GSE* 45: 32.

Hunt J, Simmons LW. 1998. Patterns of parental provisioning covary with male morphology in a horned beetle (Onthophagus taurus) (Coleoptera: Scarabaeidae). *Behavioral ecology and sociobiology* 42: 447–451.

Hunt J, Simmons LW. 2000. Maternal and paternal effects on offspring phenotype in the dung beetle Onthophagus taurus. *Evolution; international journal of organic evolution* 54: 936–941.

Jones AG, Arnold SJ, Bürger R. 2003. Stability of the G-matrix in a population experiencing pleiotropic mutation, stabilizing selection, and genetic drift. *Evolution; international journal of organic evolution* 57: 1747–1760.

Josephs EB, Stinchcombe JR, Wright SI. 2017. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *The New phytologist* 214: 21–33.

Kasper C, Schreier T, Taborsky B. 2019. Heritabilities, social environment effects and genetic correlations of social behaviours in a cooperatively breeding vertebrate. *Journal of evolutionary biology* 32: 955–973.

Kasule FK. 1991. Associations of fecundity with adult size in the cotton stainer bug Dysdercus fasciatus. *Heredity* 66: 281–286.

Klein TW, DeFries JC, Finkbeiner CT. 1973. Heritability and genetic correlation: standard errors of estimates and sample size. *Behavior genetics* 3: 355–364.

Koellinger PD, de Vlaming R. 2019. Mendelian randomization: the challenge of unobserved environmental confounds. *International journal of epidemiology* 48: 665–671.

Korol AB, Ronin YI, Itskovich AM, Peng J, Nevo E. 2001. Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits. *Genetics* 157: 1789–1803.

Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics* 44: 1066–1071.

Kruuk LEB. 2004. Estimating genetic parameters in natural populations using the ‘animal model’. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359: 873–890.

Lande R. 1979. Quantitative Genetic Analysis of Multivariate Evolution, Applied to Brain: Body Size Allometry. *Evolution; international journal of organic evolution* 33: 402–416.

Lande R, Arnold SJ. 1983. THE MEASUREMENT OF SELECTION ON CORRELATED CHARACTERS. *Evolution; international journal of organic evolution* 37: 1210–1226.

Lande R, Price T. 1989. Genetic correlations and maternal effect coefficients obtained from offspring-parent regression. *Genetics* 122: 915–922.

Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine* 27: 1133–1163.

LazareviĆ J, PeriĆ-Mataruga V, IvanoviĆ J, AndjelkoviĆ M. 1998. Host plant effects on the genetic variation and correlations in the individual performance of the Gypsy Moth. *Functional ecology* 12: 141–148.

Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. 2012. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28: 2540–2542.

Liu S, Li C, Wang H, Wang S, Yang S, Liu X, Yan J, Li B, Beatty M, Zastrow-Hayes G, *et al.* 2020. Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. *Genome biology* 21: 163.

Liu H-J, Yan J. 2019. Crop genome-wide association study: a harvest of biological relevance. *The Plant journal: for cell and molecular biology* 97: 8–18.

Lynch M, Walsh B, Others. 1998. Genetics and analysis of quantitative traits.

Mahmoud M, Zeng Y, Shirali M, Yin T, Brügemann K, König S, Haley C. 2018. Genome-wide pleiotropy and shared biological pathways for resistance to bovine pathogens. *PloS one* 13: e0194374.

McGlothlin JW, Kobiela ME, Wright HV, Mahler DL, Kolbe JJ, Losos JB, Brodie ED 3rd. 2018. Adaptive radiation along a deeply conserved genetic line of least resistance in Anolis lizards. *Evolution letters* 2: 310–322.

Mitchell DJ, Houslay TM. 2021. Context-dependent trait covariances: how plasticity shapes behavioral syndromes. *Behavioral ecology: official journal of the International Society for Behavioral Ecology* 32: 25–29.

Ni G, Moser G, Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, *et al.* 2018. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. *American journal of human genetics* 102: 1185–1194.

Paaby AB, Rockman MV. 2013. The many faces of pleiotropy. *Trends in genetics: TIG* 29: 66–73.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *American journal of human genetics* 69: 1–14.

Qian Y, Ye D, Huang H, Wu DJH, Zhuang Y, Jiang X, Mao Y. 2020. Coffee Consumption and Risk of Stroke: A Mendelian Randomization Study. *Annals of neurology* 87: 525–532.

Reale D, Roff DA. 2001. Estimating Genetic Correlations in Natural Populations in the Absence of Pedigree Information: Accuracy and Precision of the Lynch Method. *Evolution; international journal of organic evolution* 55: 1249–1255.

Reeve ECR. 1955. The Variance of the Genetic Correlation Coefficient. *Biometrics* 11: 357–374.

Revell LJ. 2007. The G matrix under fluctuating correlational mutation and selection. *Evolution; international journal of organic evolution* 61: 1857–1872.

van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. 2019. Genetic correlations of polygenic disease traits: from theory to practice. *Nature reviews. Genetics* 20: 567–581.

Robertson A. 1959. The Sampling Variance of the Genetic Correlation Coefficient. *Biometrics* 15: 469–485.

Roff DA. 1996. THE EVOLUTION OF GENETIC CORRELATIONS: AN ANALYSIS OF PATTERNS. *Evolution; international journal of organic evolution* 50: 1392–1403.

Roff D. 2000. The evolution of the G matrix: selection or drift? *Heredity* 84: 135–142.

Roff DA. 2008. Comparing sire and dam estimates of heritability: jackknife and likelihood approaches. *Heredity* 100: 32–38.

Roff DA. 2012. *Evolutionary Quantitative Genetics*. Springer Science & Business Media.

Roff DA, Preziosi R. 1994. The estimation of the genetic correlation: the use of the jackknife. *Heredity* 73: 544–548.

Rose MR. 1984. Genetic Covariation in Drosophila Life History: Untangling the Data. *The American naturalist* 123: 565–569.

Royauté R, Hedrick A, Dochtermann NA. 2020. Behavioural syndromes shape evolutionary trajectories via conserved genetic architecture. *Proceedings. Biological sciences / The Royal Society* 287: 20200183.

Saltz JB, Hessel FC, Kelly MW. 2017a. Trait Correlations in the Genomics Era. *Trends in ecology & evolution* 32: 279–290.

Saltz JB, Lymer S, Gabrielian J, Nuzhdin SV. 2017b. Genetic Correlations among Developmental and Contextual Behavioral Plasticity in Drosophila melanogaster. *The American naturalist* 190: 61–72.

Service, Philip M., Rose MR. 1985. Genetic Covariation Among Life-History Components: The Effect of Novel Environments. *Evolution; international journal of organic evolution* 39: 943–945.

Sodini SM, Kemper KE, Wray NR, Trzaskowski M. 2018. Comparison of Genotypic and Phenotypic Correlations: Cheverud's Conjecture in Humans. *Genetics* 209: 941–948.

Stearns FW. 2010. One hundred years of pleiotropy: a retrospective. *Genetics* 186: 767–773.

Steven JC, Anderson IA, Brodie ED 3rd, Delph LF. 2020. Rapid reversal of a potentially constraining genetic covariance between leaf and flower traits in Silene latifolia. *Ecology and evolution* 10: 569–578.

Steven JC, Delph LF, Brodie ED 3rd. 2007. Sexual dimorphism in the quantitative-genetic architecture of floral, leaf, and allocation traits in Silene latifolia. *Evolution; international journal of organic evolution* 61: 42–57.

Sul JH, Martin LS, Eskin E. 2018. Population structure in genetic studies: Confounding factors and mixed models. *PLoS genetics* 14: e1007309.

Sun D, Zhou T, Heianza Y, Li X, Fan M, Fonseca VA, Qi L. 2019. Type 2 diabetes and hypertension. *Circulation research* 124: 930–937.

Su J, Xu K, Li Z, Hu Y, Hu Z, Zheng X, Song S, Tang Z, Li L. 2021. Genome-wide association study and Mendelian randomization analysis provide insights for improving rice yield potential. *Scientific reports* 11: 6894.

Svensson EI, Arnold SJ, Bürger R, Csilléry K, Draghi J, Henshaw JM, Jones AG, De Lisle S, Marques DA, McGuigan K, *et al.* 2021. Correlational selection in the age of genomics. *Nature ecology & evolution* 5: 562–573.

Swallow WH, Monahan JF. 1984. Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences* 26: 47–57.

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. 2019. Benefits and limitations of genome-wide association studies. *Nature reviews. Genetics* 20: 467–484.

Thoen MPM, Davila Olivas NH, Kloth KJ, Coolen S, Huang P-P, Aarts MGM, Bac-Molenaar JA, Bakker J, Bouwmeester HJ, Broekgaarden C, *et al.* 2017. Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *The New phytologist* 213: 1346–1362.

Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, Nguyen-Viet TA, Wedow R, Zacher M, Furlotte NA, *et al.* 2018. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics* 50: 229–237.

Ungerer MC, Halldorsdottir SS, Modliszewski JL, Mackay TFC, Purugganan MD. 2002. Quantitative trait loci for inflorescence development in Arabidopsis thaliana. *Genetics* 160: 1133–1151.

VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. 2014. Methodological challenges in mendelian randomization. *Epidemiology* 25: 427–435.

Verbanck M, Chen C-Y, Neale B, Do R. 2018. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics* 50: 693–698.

Via S, Lande R. 1985. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution; international journal of organic evolution* 39: 505–522.

Vieira C, Pasyukova EG, Zeng ZB, Hackett JB, Lyman RF, Mackay TF. 2000. Genotype-environment interaction for quantitative trait loci affecting life span in Drosophila melanogaster. *Genetics* 154: 213–227.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature reviews. Genetics* 12: 204–213.

Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, Posthuma D. 2019. A global overview of pleiotropy and genetic architecture in complex traits. *Nature genetics* 51: 1339–1348.

Wilson AJ, Réale D, Clements MN, Morrissey MM, Postma E, Walling CA, Kruuk LEB, Nussey DH. 2010. An ecologist's guide to the animal model. *The Journal of animal ecology* 79: 13–26.

Wu F, Huang Y, Hu J, Shao Z. 2020. Mendelian randomization study of inflammatory bowel disease and bone mineral density. *BMC medicine* 18: 312.

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, Lifelines Cohort Study, Esko T, *et al.* 2015. Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Human molecular genetics* 24: 7445–7449.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011a. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* 88: 76–82.

Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR, Mangino M, *et al.* 2011b. Genomic inflation factors under polygenic inheritance. *European journal of human genetics: EJHG* 19: 807–812.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics* 46: 100–106.

Young AI, Benonisdottir S, Przeworski M, Kong A. 2019. Deconstructing the sources of genotype-phenotype associations in humans. *Science* 365: 1396–1400.

Zas R, Sampedro L. 2015. Heritability of seed weight in Maritime pine, a relevant trait in the transmission of environmental maternal effects. *Heredity* 114: 116–124.

Zhang Y, Cheng Y, Jiang W, Ye Y, Lu Q, Zhao H. 2021. Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. *Briefings in bioinformatics*.

Zhang Z, Ma P, Li Q, Xiao Q, Sun H, Olasege BS, Wang Q, Pan Y. 2018. Exploring the Genetic Correlation Between Growth and Immunity Based on Summary Statistics of Genome-Wide Association Studies. *Frontiers in genetics* 9: 393.

Zhang Y, Zhang J, Gong H, Cui L, Zhang W, Ma J, Chen C, Ai H, Xiao S, Huang L, *et al.* 2019. Genetic correlation of fatty acid composition with growth, carcass, fat deposition and meat quality traits based on GWAS data in six pig populations. *Meat science* 150: 47–55.

Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, Hemani G, Tansey K, Laurin C, Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, *et al.* 2017. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 33: 272–279.

Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, Robinson MR, McGrath JJ, Visscher PM, Wray NR, *et al.* 2018. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature communications* 9: 1–12.

Figures

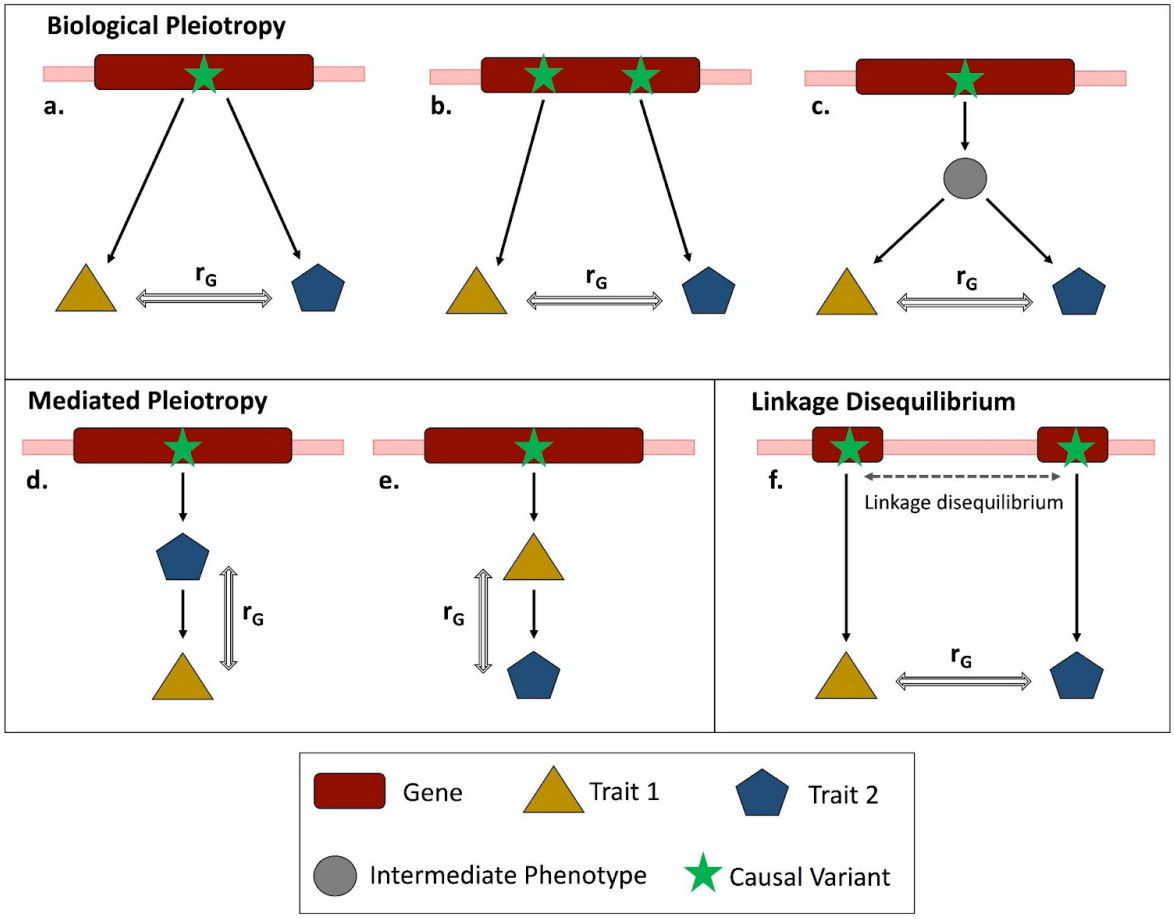

Figure 4-1. Schematic representation of the mechanistic basis of genetic correlations. Genetic correlations can arise due to biological pleiotropy (a-c) where a genetic variant (a) or a gene (b) directly influences both traits (also referred to as horizontal pleiotropy), or indirectly (c) through an intermediate phenotype, Mediated pleiotropy (d-e), wherein a genetic variant directly alters Trait 1 or Trait 2 which then influences Trait 2 or Trait 1 (also referred to as vertical pleiotropy), or through Linkage disequilibrium (f) wherein the causal genes for the two traits are in linkage and thus inherited together. r_G represents estimated genetic correlations between the traits.

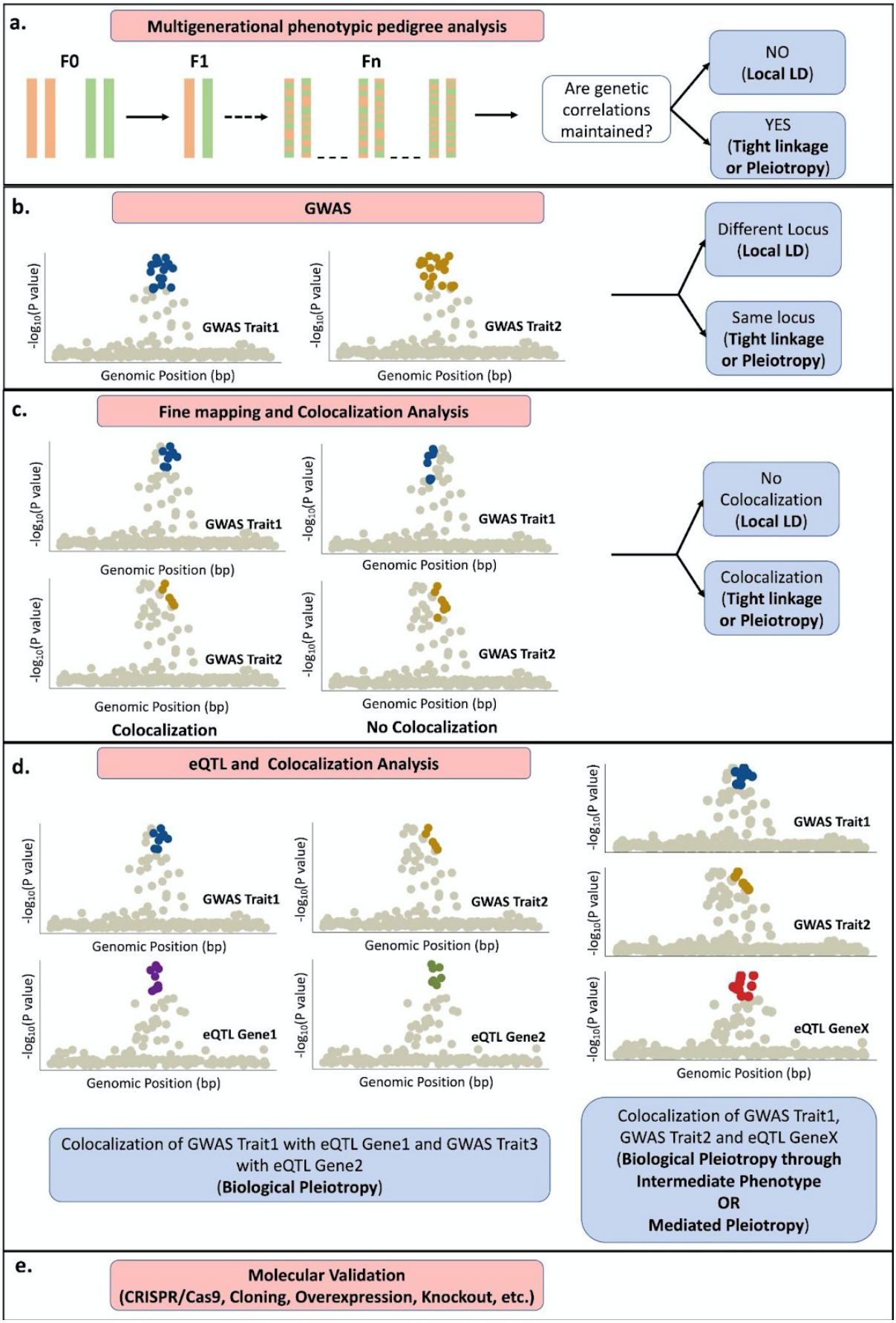
a.
Multigenerational phenotypic pedigree analysis
F0
F1
Fn
Are genetic correlations maintained?
NO (Local LD)
YES (Tight linkage or Pleiotropy)
b.
GWAS
-log10(P value)
GWAS Trait1
Genomic Position (bp)
-log10(P value)
GWAS Trait2
Genomic Position (bp)
Different Locus (Local LD)
Same locus (Tight linkage or Pleiotropy)
c.
Fine mapping and Colocalization Analysis
-log10(P value)
GWAS Trait1
Genomic Position (bp)
-log10(P value)
GWAS Trait2
Genomic Position (bp)
Colocalization
-log10(P value)
GWAS Trait1
Genomic Position (bp)
-log10(P value)
GWAS Trait2
Genomic Position (bp)
No Colocalization
No Colocalization (Local LD)
Colocalization (Tight linkage or Pleiotropy)
d.
eQTL and Colocalization Analysis
-log10(P value)
GWAS Trait1
Genomic Position (bp)
-log10(P value)
GWAS Trait2
Genomic Position (bp)
-log10(P value)
eQTL Gene1
Genomic Position (bp)
-log10(P value)
eQTL Gene2
Genomic Position (bp)
Colocalization of GWAS Trait1 with eQTL Gene1 and GWAS Trait3 with eQTL Gene2 (Biological Pleiotropy)
-log10(P value)
GWAS Trait1
-log10(P value)
GWAS Trait2
-log10(P value)
eQTL GeneX
Genomic Position (bp)
Colocalization of GWAS Trait1, GWAS Trait2 and eQTL GeneX (Biological Pleiotropy through Intermediate Phenotype OR Mediated Pleiotropy)
e.
Molecular Validation (CRISPR/Cas9, Cloning, Overexpression, Knockout, etc.)

Figure 4-2 Overview of analytical methods that can be integrated to identify the causal genetic mechanism underlying trait correlations.
(a) Multigenerational phenotypic pedigree analysis can be conducted to identify whether the trait correlations are due to linkage disequilibrium (LD), or tight linkage or pleiotropy, (b) GWAS can be used to identify variants associated with the traits; different genomic loci for the two traits indicates linkage underlying trait correlation as compared to the same locus which could be due to either tight linkage or pleiotropy, (c) Fine mapping can enable prioritization of causal loci obtained from GWAS, which can be followed by colocalization tests to see if the set of causal loci colocalize, indicating tight linkage or pleiotropy; no colocalization indicates linkage, (d) omic-QTL can be performed to potentially characterize the effect of the causal variant and colocalizing results of omic-QTL with GWAS can help differentiate between biological and mediated pleiotropy; shown is an example of eQTL wherein the colocalized GWAS for trait 1 and trait 2 is colocalized with eQTL -- in the first scenario GWAS for the two traits colocalizes with two distinct genes' eQTL, indicating biological pleiotropy and in the second scenario GWAS for the two traits colocalizes with the same gene eQTL indicating either biological pleiotropy through an intermediate phenotype or mediated pleiotropy, (e) Molecular validation tools can then be employed to further tease these apart and to validate the findings.

Tables

Species	Family Design	Correlation Method	# of Traits	Ref
Mimulus guttatus and *Mimulus micranthus*	Full-sib	Pearson correlation of family means	6	(Fenster & Carr, 1997)
		Parent-offspring regression	3	
Lymantria dispar	Full-sib	Pearson correlation of family means	7	(LazareviĆ *et al.*, 1998)
Drosophila melanogaster	Full-sib (RIL)	Variance Component (ANOVA)	2	(Vieira *et al.*, 2000)
Gryllus firmus and *Gryllus pennsylvanicus*	Full-sib	Variance Component (ANOVA)	5	(Bégin & Roff, 2001)
Raphanus raphanistrum	Parent-offspring	Variance Component (REML)	6	(Conner, 2002)
Arabidopsis thaliana	Full-sib (RIL)	Variance Component (ANOVA)	13	(Ungerer *et al.*, 2002)
Silene latifolia	paternal and maternal half-sibling full-sibling	Variance Component (REML)	7	(Steven *et al.*, 2007)
Acrocephalus arundinaceus	Parent-offspring	Parent-offspring regression	7	(Åkesson *et al.*, 2008)
	Mixed family design	Variance Component (REML)		
Silene latifolia	paternal and maternal half-sibling full-sibling	Variance Component (REML)	1*	(Delph *et al.*, 2011)
Pinus pinaster	Half-sib	Variance Component (REML)	3	(Zas & Sampedro, 2015)

Anolis lizards (7 species)	Half-sib	Variance Component (REML)	8	(McGlothlin *et al.*, 2018)
Neolamprologus pulcher	paternal half-sibling full-sibling	Variance Component (MCMC)	8	(Kasper *et al.*, 2019)
Boechera stricta	full-sibling	Variance Component (MCMC)	7	(Bemmels & Anderson, 2019)
Gryllus integer	paternal and maternal half-sibling full-sibling	Variance Component (REML)	3	(Royauté *et al.*, 2020)
Australian Drosophila	paternal half-sibling full-sibling	Variance Component (REML)	4	(Hangartner *et al.*, 2020)
Salmo salar	Mixed family design	Variance Component (REML)	4	(Debes *et al.*, 2021)

Table 4-1: Example of studies reporting the genetic correlation estimates using family-level phenotypic data.
RIL: Recombinant Inbred Line; ANOVA: Analysis of Variance; REML: Restricted Maximum Likelihood; MCMC: Markov Chain Monte Carlo;
*One trait compared between sexes.

Variable	Benefits	Drawbacks
Choice of Family Design		
Parent-Offspring	Simple; time efficient; data relatively easy to obtain (Roff, 2012)	biased by maternal effects, selection, and shared environment (Lande & Price, 1989; Åkesson *et al.*, 2007); standard errors are underestimated (Robertson, 1959)
Full-sibling	Low standard error (Klein *et al.*, 1973)	Confounded by dominance effects (Lynch *et al.*, 1998); (Hill, 2013), and the presence of maternal effects in such families (Falconer 1996; Hunt and Simmons 1998)
Half-sibling	No Dominance effect; maternal effects can be estimated (Lynch *et al.*, 1998; Wilson *et al.*, 2010)	High standard error unless large sample size used (Roff, 2008)
Mixed family designs	Most accurate as they can account for dominance, maternal effects efficiently, common environment ((Fox, 1998; Hunt & Simmons, 2000; Fox *et al.*, 2004; Hallsson & Björklund, 2012))	Complex design; data not easy to obtain
Choice of Method		
Parent-Offspring Regression (offspring values of a trait are regressed on the parent values of other trait)	Simple and easy to use	High standard error (Åkesson *et al.*, 2007)
Line Means (Pearson product-moment correlation between family means)	Relatively easy to use	Large family size needed (Roff & Preziosi, 1994)
Variance Component (phenotypic variance partitioned into into	ANOVA- Simplest to use, especially when the family design is not complex.	ANOVA- Higher standard error when data is unbalanced (Swallow & Monahan, 1984)

components of genetic variance)	REML- Can handle unbalanced data (Swallow & Monahan, 1984); individuals with unknown paternity can be included (Kruuk, 2004) MCMC- High accuracy even when the sample size is small (Gao *et al.*, 2003)	REML- Is not accurate when applied to non-Gaussian traits (Bolker *et al.*, 2009) and high standard error for a small sample size (Gao *et al.*, 2003) MCMC- Computationally very expensive (Browne & Draper, 2006)

Table 4-2 Benefits and drawbacks of the most commonly used family design and analyses method used for estimating genetic correlations from phenotypic data. ANOVA: Analysis of Variance; REML: Restricted Maximum Likelihood; MCMC: Markov Chain Monte Carlo.

Species	**Correlation Method**	**# of Traits**	**Ref**
Arabidopsis thaliana	MTMM	11	(Thoen *et al.*, 2017)
Pig	GREML	32	(Zhang *et al.*, 2019)
Cattle	GREML	23	(Mahmoud *et al.*, 2018)
Human	GREML and LDSR	2	(Ni *et al.*, 2018)
Human	LDSR	17	(Sodini *et al.*, 2018)
Human	LDSR	25	(Zhang *et al.*, 2018)
Cattle	LDSR	4	(Cai *et al.*, 2020)

Table 4-3 Example of studies reporting the genetic correlation estimates using genetic data. MTMM: Multi-trait Mixed Model; GREML: Genome-based Restricted Maximum Likelihood; LDSR: Linkage Disequilibrium Scores Regression.

Chapter 5

Discussion and Future Directions

The overarching goal of my thesis was to gain an understanding of the genetic basis of adaptive traits in the genus *Ipomoea*. I examined the highly diverse leaf shape trait in sweetpotato, a hexaploid crop with very limited genomic resources, and glyphosate resistance evolution in the common morning glory, a diploid invasive species. The characteristic differences in the traits examined and the species' genomic architecture provided a unique opportunity to examine the underlying mechanisms controlling the traits. Specifically, I found that leaf shape in sweetpotato is largely genetically controlled and identified putative candidate genes associated with leaf shape. Further, I identified potential genes that underlie the polygenic glyphosate resistance, and its associated cost, in *I. purpurea*. I also showed the role of linkage disequilibrium in maintaining the resistance alleles over generations, and potentially also conferring cost. Moreover, I formulated a conceptual chapter wherein I suggest how one can potentially identify the molecular mechanisms that underlie cost-benefit relationships (like the resistance-cost relationship above). Altogether, this thesis provides novel insights into the genetic underpinnings of complex polygenic adaptive traits in the genus *Ipomoea*. Conclusions from each chapter compel multiple future directions.

What does the leaf shape diversity in sweetpotato teach us?

Leaf shape is a highly variable trait that varies across taxonomic levels, geography, and in response to environmental differences (Ashby, 1948; Hilu, 1983; Gurevitch, 1988; Harris *et al.*, 1998). However, comprehensive intraspecific analyses of leaf shape variation across variable environments, disentangling the role of genetics vs environment, is surprisingly absent. In

chapter two, I aimed to answer (1) how diverse is leaf shape at a species-wide level? (2) what are the candidate genes associated with leaf shape? and (3) to what degree does the environment and GxE influence leaf shape traits? I found evidence of extensive intraspecific morphological variation in leaf shape and showed that most of this variation is controlled by the genotype, with low or limited influence of GxE. Next, I also identified putative genes associated with leaf shape (extending beyond the simple shape descriptors). Unexpectedly, I found that although simple leaf shape traits are individually only slightly influenced by the environment, combinations of simple leaf shapes are significantly altered by the environment.

An important question among plant morphologists is the extent to which leaf shape varies among genotypes in a species, and how much of this variation can be attributed to the genotype. Using three traditional leaf shape descriptors and more comprehensive EFD (Elliptical Fourier Descriptors), I showed that the commonly used shape descriptors are inefficient in capturing the entirety of leaf shape variation and that one should use comprehensive morphometric techniques. Importantly, my results showed that leaf shape variation does not follow a trend across species -- I found that leaf dissection contributes most to the morphological variation in leaf shape in sweetpotato, as compared to aspect ratio (ratio of length-to-width) in apple and tomato (Chitwood *et al.*, 2013; Migicovsky *et al.*, 2017). This is most likely due to multiple independent evolutions of leaf shape across phylogenetic taxa (Nicotra *et al.*, 2011).

I also showed that most of the traditional leaf shape descriptors are majorly genetically controlled (with little to no significant effect of the environment), and thus have high heritability values that can be actively selected for (or against) by breeders. Contrary to this result, multiple studies have found that leaf dissection is a plastic trait that responds to changes in temperature (Royer *et al.*, 2009; Royer, 2012; Chitwood *et al.*, 2016)**.** This could reflect that the difference between the gardens used (MI and OH) was not sufficient to capture leaf shape variation in response to change in the environment. Thus, multiple studies in environments that range more widely for temperature will need to be performed in the future to confirm that leaf shape in sweetpotato does not vary significantly with change in the environment.
To further the understanding of genes associated with leaf shape variation, due to the limited genomic resources available for this species, I performed a gene expression study to identify

genes showing expression changes in varying leaf shape types. Using functional annotations of the differentially expressed genes, I identified potential candidate genes that could contribute to leaf shape variation. For example, I identified the FRS genes to be upregulated in nondissected individuals as compared to the highly dissected ones. FRS is a transcription factor that potentially binds to the promoter region of STM gene which has been shown to alter leaf serrations in other species (Kawamura *et al.*, 2010; Aguilar-Martínez *et al.*, 2015). Similarly, we identified the CHS and feruloyl CoA 6′-hydroxylase genes to be differentially regulated for AR; these genes have been shown in the literature to alter the longitudinal vs latitudinal expansion of the leaves (Liu *et al.*, 2017). Further studies along with the development of functional tools and genomic resources for this species would be needed to verify these candidate genes.

What have we learned about the mechanistic basis of polygenic glyphosate resistance?

The adaptation of weedy plants to herbicides forms an excellent model system to investigate the genetic basis of adaptation. Plants adapt to herbicides broadly by two mechanisms, commonly known as target-site resistance (TSR; resistance is conferred by changes in the target gene) and nontarget-site resistance (NTSR; resistance is conferred by changes in nontarget genes) (Powles & Yu, 2010; Mithila & Godar, 2013). Although multiple studies have identified genes associated with TSR, a comprehensive study examining the complex polygenic genetic basis and the evolutionary factors that maintain NTSR over generations remain limited (Délye, 2013; Baucom, 2019; Beckie, 2020; Leon *et al.*, 2021). In chapter 3, I performed a multi-level analysis to answer (1) what mechanism and genes underlie NTSR glyphosate resistance in the weed, *Ipomoea purpurea*? (2) how are these genes maintained together over evolutionary time? and (3) what are the putative genes that could explain the cost of resistance (lower germination rate) in this species? I found multiple detoxification and stress signaling genes associated with resistance, confirming the role of detoxification in conferring glyphosate resistance. Further, I found strong interchromosomal linkage disequilibrium (ILD) between detoxification genes present on separate chromosomes indicating the potential role of ILD in maintaining resistance through generations. Additionally, through the multi-layer analysis, I also suggest putative cost loci and the role of genetic hitchhiking in incurring the cost associated with herbicide resistance.

Using a comprehensive multi-layer approach involving whole-genome scan, gene expression study, and a functional assay, I showed that detoxification underlies the polygenic NTSR response to glyphosate in *I. purpurea*. I found multiple genes involved in the detoxification pathway (phosphate transporters, cytochrome P450s, glycosyltransferases) to be under selection, and differentially expressed in resistant individuals. Additionally, I identified almost fixed non-synonymous mutations that could potentially underlie the detoxification functional mechanism. Uniquely, I also found evidence of selection and differential expression on genes related to stress response. Additionally, using a functional assay, I showed that inhibition of cytochrome P450 leads to loss of resistance. Detoxification and stress response both have either been hypothesized or shown to be involved in herbicide resistance (Radwan, 2012; Duhoux & Délye, 2013; Dyer, 2018; Vega *et al.*, 2020). Based on my results, I think that detoxification genes might be the major effect genes, while stress signaling and response genes might contribute minor effects on the resistance phenotype in this species. Follow-up work will be needed to confirm this finding.

Since the detoxification genes under selection were present on separate chromosomes, I tested for the presence of interchromosomal linkage disequilibrium in maintaining these alleles over generations. I indeed found that the regions under selection showed evidence of high ILD as compared to the background ILD and the other highly differentiated regions. Additionally, I found that within the regions under selection, the strongest ILD was present between the putative resistance genes. Although one could argue that linkage would decay in the presence of recombination and gene flow, it has been suggested that linkage in the presence of selection on genes involved in co-adaptation, could become very steep (Lewontin & Kojima, 1960; Nei, 1967) and even could lead to the fixation of alleles (Takahasi, 2007). Thus, ILD along with co-adaption could explain how the NTSR resistance alleles are maintained through generations in natural populations.

While trade-offs are central to the theory of evolution, our understanding of the genetic loci associated with the cost of herbicide resistance remains limited -- currently, no studies in the resistance literature have identified fitness cost loci associated with NTSR. Here, through a combination of selection scan, linkage analysis, and differential expression I identified putative cost loci of glyphosate resistance. *I. purpurea* has been shown to incur resistance cost wherein

highly resistant individuals have lower germination rates (Van Etten *et al.*, 2016). Through selection scan and linkage analysis, I found two genes (*NFD6*, *NAC25*) that were under strong selection in resistant individuals and were further tightly linked to the resistance loci. Both these genes are crucial for normal seed development and maturation. Further, by comparing expression profiles of genes in the resistant vs susceptible individuals both in the absence and presence of the herbicide, I showed that *NFD6* is indeed differentially regulated. This indicates the potential role of genetic-hitchhiking in maintaining resistance cost, but further functional studies would need to be conducted to validate this.

What have we learned about teasing apart the mechanistic basis of trait correlation?

Trait correlations are pervasive in nature and are relevant to the field of ecology and evolutionary biology since they can facilitate or constraint adaptation (Roff, 1996; Conner *et al.*, 2011). Understanding the genetic basis of trait correlations can help understand the causes and consequences of trait correlations. Genetic correlation between traits arises due to either pleiotropy (biological or mediated) or linkage and differentiating between them is of prime interest to evolutionary biologists since it primarily determines the stability and persistence of the association over evolutionary times. Importantly, we can also start to understand how prevalent one mechanism may be over the other in giving rise to the phenotypic diversity we see today. In chapter 4, I reviewed the current methods available to study genetic correlations, discussed the pitfalls of these methods, and outlined an integrative approach that can be used to overcome these limitations.

Traditionally, estimating genetic correlations and their underlying causality has involved using family-level phenotypic data to deconstruct the genetic variance and covariance (Fenster & Carr, 1997; Conner, 2002; Delph *et al.*, 2011; Conner *et al.*, 2011; Bemmels & Anderson, 2019). Multiple family designs exist for this with the simplest ones being parent-offspring, half-sibling, and full-sibling. Each of these has inherent issues in that they are compounded by one or more of the following -- environmental factors, maternal effects, dominance factors (Lande & Price, 1989; Falconer, 1996; Hunt & Simmons, 1998; Åkesson *et al.*, 2007). Additionally, unless large family data is used, the errors associated with the estimated genetic correlations can be really

high, making the estimation unreliable (Robertson, 1959; Roff & Preziosi, 1994). Alternatively, a mix of family designs, like nested full-sib half-sib, can be used which takes into account the compounding factors and also lowers the error. One major caveat with estimating genetic correlations with family phenotypic data is that these can only be applied to species that can be crossed and thus cannot be applied universally. Furthermore, differentiating between pleiotropy and linkage using phenotypic data is theoretically possible by following the genetic correlation through multi-generations (Fenster & Carr, 1997; Beldade *et al.*, 2002; Conner *et al.*, 2003; Hansen Wheat *et al.*, 2019), but again the feasibility of this method remains limited to species with short life-spans and species which are easy to breed. Additionally, a multigenerational study can only differentiate between loose linkage and pleiotropy and cannot make distinctions between tight linkage and biological and mediated pleiotropy. Moreover, in the presence of selection on the correlated traits (correlational selection), even loose linkage might not be identified. Thus, the applicability of current family-level phenotypic methods to study genetic correlations remains questionable.

With the advent of sequencing technologies, the focus has moved to utilizing molecular markers to study genetic correlations. This has mainly constituted of using GWAS data, either individual-level or summary-statistic data to identify the extent of associations between loci associated with the two traits (Korte *et al.*, 2012; Yang *et al.*, 2015; Thoen *et al.*, 2017; Calus *et al.*, 2018; Guo *et al.*, 2021). Although multiple methods exist, these have major limitations in that they require high computational power, and have a high standard error associated with them (Ni *et al.*, 2018; van Rheenen *et al.*, 2019). Importantly, methods estimating genetic correlation from summary statistics GWAS assume that the allele frequency differences due to subpopulations are independent of linkage structure, but this assumption does not hold true in the presence of correlational selection (Berg *et al.*, 2019). Furthermore, although multiple studies have claimed the presence of pleiotropy in maintaining trait correlations using GWAS data, their definition of pleiotropy is problematic (Gardner & Latta, 2007; Watanabe *et al.*, 2019). These studies define pleiotropic genetic correlations when the two traits have overlapping loci as identified by GWAS. But it is commonly known that GWAS hits can constitute multiple genes, and thus overlapping loci could be due to linkage as well. Recently, a couple of approaches have been developed for differentiating linkage and pleiotropy (Davey Smith & Ebrahim, 2003; Bolormaa

et al., 2014). For example, Mendelian Randomization (MR) can be used to test the presence of mediated pleiotropy. This technique requires however extensive a priori knowledge and additionally can be confounded by the presence of indirect genetic effects, population stratification, and assortative mating (VanderWeele *et al.*, 2014). This highlights the caveats associated with studying genetic correlation by solely utilizing GWAS datasets.

To enable a comprehensive dissection of genetic correlations, I outlined an integrative approach that can overcome some of the limitations mentioned above. I suggest using multiple analyses using a combination of family-level phenotypic data, GWAS, fine mapping and colocalization, omics-QTL and colocalization, and molecular tools to disentangle whether biological or mediated pleiotropy or linkage is the mechanism at play. These methods can be applied depending on the species, *a priori* knowledge, and the availability of genome editing tools for the species. If followed serially, the suggested analysis approach not just be able to identify causal loci, but also identify the functional mechanism through which the variant alters the traits. This though could be a potential caveat that if the variant does not alter the trait through commonly used omic-QTL, this might not be able to differentiate between pleiotropy and tight linkage unless molecular validation tools are used. By outlining an integrative approach to comprehensively study genetic correlations, I hope to foster a deeper analysis of genetic causes of trait correlations which would help us understand the stability of trait correlations and predict more precisely how trait correlations will evolve over time.

Conclusion

In this thesis, I performed a series of genetic analyses aimed at identifying the genetic basis of complex adaptive traits. The first two chapters focused on using quantitative and population genetic approaches of studying phenotype to genotype, and identified novel loci and mechanisms underlying the traits. The third chapter was focused on outlining how to comprehensively study genetic correlations, which underlie most adaptive traits. Together, my thesis identified novel loci and mechanisms that underlie complex adaptive traits in the genus *Ipomoea* and outlined approaches to better understand the genetic complexity of trait correlations which are central to complex traits.

References

Aguilar-Martínez JA, Uchida N, Townsley B, West DA, Yanez A, Lynn N, Kimura S, Sinha N. 2015. Transcriptional, posttranscriptional, and posttranslational regulation of SHOOT MERISTEMLESS gene expression in Arabidopsis determines gene function in the shoot apex. *Plant physiology* 167: 424–442.

Åkesson M, Bensch S, Hasselquist D. 2007. Genetic and phenotypic associations in morphological traits: a long term study of great reed warblersAcrocephalus arundinaceus. *Journal of avian biology* 38: 58–72.

Ashby E. 1948. Studies in the morphogenesis of leaves: I. An essay of leaf shape. *The New phytologist* 47: 153–176.

Baucom RS. 2019. Evolutionary and ecological insights from herbicide-resistant weeds: what have we learned about plant adaptation, and what is left to uncover? *The New phytologist* 223: 68–82.

Beckie HJ. 2020. Herbicide Resistance in Plants. *Plants* 9.

Beldade P, Koops K, Brakefield PM. 2002. Modularity, individuality, and evo-devo in butterfly wings. *Proceedings of the National Academy of Sciences of the United States of America* 99: 14262–14267.

Bemmels JB, Anderson JT. 2019. Climate change shifts natural selection and the adaptive potential of the perennial forb Boechera stricta in the Rocky Mountains. *Evolution; international journal of organic evolution* 73: 2247–2262.

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, *et al.* 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* 8: e39725.

Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, Tier B, Savin K, Hayes BJ, Goddard ME. 2014. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS genetics* 10: e1004198.

Calus MPL, Goddard ME, Wientjes YCJ, Bowman PJ, Hayes BJ. 2018. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. *Journal of dairy science* 101: 4279–4294.

Chitwood DH, Kumar R, Headland LR, Ranjan A, Covington MF, Ichihashi Y, Fulop D, Jiménez-Gómez JM, Peng J, Maloof JN, *et al.* 2013. A Quantitative Genetic Basis for Leaf Morphology is Revealed in a Set of Precisely Defined Tomato Introgression Lines. *The Plant cell* 25: 2465–2481.

Chitwood DH, Rundell SM, Li DY, Woodford QL, Yu TT, Lopez JR, Greenblatt D, Kang J, Londo JP. 2016. Climate and developmental plasticity: interannual variability in grapevine leaf morphology. *Plant physiology*.

Conner JK. 2002. Genetic mechanisms of floral trait correlations in a natural population. *Nature* 420: 407–410.

Conner JK, Franks R, Stewart C. 2003. Expression of additive genetic variances and covariances for wild radish floral traits: comparison between field and greenhouse environments. *Evolution; international journal of organic evolution* 57: 487–495.

Conner JK, Karoly K, Stewart C, Koelling VA, Sahli HF, Shaw FH. 2011. Rapid independent trait Evolution despite a strong pleiotropic genetic correlation. *The American naturalist* 178: 429–441.

Davey Smith G, Ebrahim S. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International journal of epidemiology* 32: 1–22.

Delph LF, Steven JC, Anderson IA, Herlihy CR, Brodie ED 3rd. 2011. Elimination of a genetic correlation between the sexes via artificial correlational selection. *Evolution; international journal of organic evolution* 65: 2872–2880.

Délye C. 2013. Unravelling the genetic bases of non-target-site-based resistance (NTSR) to herbicides: a major challenge for weed science in the forthcoming decade. *Pest management science* 69: 176–187.

Duhoux A, Délye C. 2013. Reference genes to study herbicide stress response in Lolium sp.: up-regulation of P450 genes in plants resistant to acetolactate-synthase inhibitors. *PloS one* 8: e63576.

Dyer WE. 2018. Stress-induced evolution of herbicide resistance and related pleiotropic effects. *Pest management science* 74: 1759–1768.

Falconer DS. 1996. *Introduction to Quantitative Genetics*. Pearson Education.

Fenster CB, Carr DE. 1997. Genetics of sex allocation in Mimulus (Scrophulariaceae). *Journal of evolutionary biology*.

Gardner KM, Latta RG. 2007. Shared quantitative trait loci underlying the genetic correlation between continuous traits. *Molecular ecology* 16: 4195–4209.

Guo J, Bakshi A, Wang Y, Jiang L, Yengo L, Goddard ME, Visscher PM, Yang J. 2021. Quantifying genetic heterogeneity between continental populations for human height and body mass index. *Scientific reports* 11: 5240.

Gurevitch J. 1988. Variation in leaf dissection and leaf energy budgets among populations of Achillea from an altitudinal gradient. *American journal of botany* 75: 1298.

Hansen Wheat C, Fitzpatrick JL, Rogell B, Temrin H. 2019. Behavioural correlations of the domestication syndrome are decoupled in modern dog breeds. *Nature communications* 10: 2422.

Harris W, Beever RE, Heenan PB. 1998. Phenotypic variation of leaves and stems of wild stands of cordyline Australis (Lomandraceae). *New Zealand journal of botany* 36: 593–604.

Hilu KW. 1983. The Role of Single-Gene Mutations in the Evolution of Flowering Plants. In: Hecht MK, Wallace B, Prance GT, eds. Evolutionary Biology: Volume 16. Boston, MA: Springer US, 97–128.

Hunt J, Simmons LW. 1998. Patterns of parental provisioning covary with male morphology in a horned beetle (Onthophagus taurus) (Coleoptera: Scarabaeidae). *Behavioral ecology and sociobiology* 42: 447–451.

Kawamura E, Horiguchi G, Tsukaya H. 2010. Mechanisms of leaf tooth formation in Arabidopsis. *The Plant journal: for cell and molecular biology* 62: 429–441.

Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics* 44: 1066–1071.

Lande R, Price T. 1989. Genetic correlations and maternal effect coefficients obtained from offspring-parent regression. *Genetics* 122: 915–922.

Leon RG, Dunne JC, Gould F. 2021. The role of population and quantitative genetics and modern sequencing technologies to understand evolved herbicide resistance and weed fitness. *Pest management science* 77: 12–21.

Lewontin RC, Kojima K-I. 1960. The Evolutionary Dynamics of Complex Polymorphisms. *Evolution; international journal of organic evolution* 14: 458–472.

Liu S, Zainuddin IM, Vanderschuren H, Doughty J, Beeching JR. 2017. RNAi inhibition of feruloyl CoA 6'-hydroxylase reduces scopoletin biosynthesis and post-harvest physiological deterioration in cassava (Manihot esculenta Crantz) storage roots. *Plant molecular biology* 94: 185–195.

Migicovsky Z, Li M, Chitwood DH, Myles S. 2017. Morphometrics Reveals Complex and Heritable Apple Leaf Shapes. *Frontiers in plant science* 8: 2185.

Mithila J, Godar AS. 2013. Understanding Genetics of Herbicide Resistance in Weeds: Implications for Weed Management. *Advances in Crop Science and Technology* 1: 1–3.

Nei M. 1967. Modification of linkage intensity by natural selection. *Genetics* 57: 625–641.

Nicotra AB, Leigh A, Boyce CK, Jones CS, Niklas KJ, Royer DL, Tsukaya H. 2011. The evolution and functional significance of leaf shape in the angiosperms. *Functional Plant Biology*.

Ni G, Moser G, Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, *et al.* 2018. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. *American journal of human genetics* 102: 1185–1194.

Powles SB, Yu Q. 2010. Evolution in Action: Plants Resistant to Herbicides. *Annual review of plant biology* 61: 317–347.

Radwan DEM. 2012. Salicylic acid induced alleviation of oxidative stress caused by clethodim in maize (Zea mays L.) leaves. *Pesticide biochemistry and physiology* 102: 182–188.

van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. 2019. Genetic correlations of polygenic disease traits: from theory to practice. *Nature reviews. Genetics* 20: 567–581.

Robertson A. 1959. The Sampling Variance of the Genetic Correlation Coefficient. *Biometrics* 15: 469–485.

Roff DA. 1996. THE EVOLUTION OF GENETIC CORRELATIONS: AN ANALYSIS OF PATTERNS. *Evolution; international journal of organic evolution* 50: 1392–1403.

Roff DA, Preziosi R. 1994. The estimation of the genetic correlation: the use of the jackknife. *Heredity* 73: 544–548.

Royer DL. 2012. Leaf shape responds to temperature but not CO2 in Acer rubrum. *PloS one* 7: e49559.

Royer DL, Meyerson LA, Robertson KM, Adams JM. 2009. Phenotypic plasticity of leaf shape along a temperature gradient in Acer rubrum. *PloS one*.

Takahasi KR. 2007. Evolution of coadaptation in a subdivided population. *Genetics* 176: 501–511.

Thoen MPM, Davila Olivas NH, Kloth KJ, Coolen S, Huang P-P, Aarts MGM, Bac-Molenaar JA, Bakker J, Bouwmeester HJ, Broekgaarden C, *et al.* 2017. Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *The New phytologist* 213: 1346–1362.

VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. 2014. Methodological challenges in mendelian randomization. *Epidemiology* 25: 427–435.

Van Etten ML, Kuester A, Chang S-M, Baucom RS. 2016. Fitness costs of herbicide resistance across natural populations of the common morning glory, Ipomoea purpurea. *Evolution; international journal of organic evolution* 70: 2199–2210.

Vega T, Gil M, Martin G, Moschen S, Picardi L, Nestares G. 2020. Stress response and detoxification mechanisms involved in non-target-site herbicide resistance in sunflower. *Crop science* 60: 1809–1822.

Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, Posthuma D. 2019. A global overview of pleiotropy and genetic architecture in complex traits. *Nature genetics* 51: 1339–1348.

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, Lifelines Cohort Study, Esko T, *et al.* 2015. Genome-wide genetic homogeneity

between sexes and populations for human height and body mass index. *Human molecular genetics* 24: 7445–7449.

Appendix A

Supplementary Methods, Figures and Tables for Chapter 2

Method S1: RNA-Seq data processing and transcriptomic analysis

Briefly, we performed quality control for the obtained raw reads to trim the adaptors, discard low-quality reads and eliminate poor-quality bases. We used cutadapt v1.4 (Martin, 2011) to remove the adapters, and trimmomatic v0.36 (Bolger *et al.*, 2014) to clean the reads based on length and quality score. Further, we performed error correction of the RNA-Seq data using rcorrecter (Song & Florea, 2015) to retain only high-quality data.

Next, we used filtered reads separately from one entire and one lobed individual, randomly chosen, for *de novo* transcriptome assembly, which served as a reference transcriptome for differential analysis. To get a comprehensive assembly, we used both a single k-mer approach, using Trinity v2.2.0 (Grabherr *et al.*, 2011), with k=25, and multi k-mer approach, using Velvet/Oases v1.2.10 (Zerbino & Birney, 2008; Schulz *et al.*, 2012) with k-mer ranging from 23-41 and 93-99 with a step-size of 2. Next, we used the EvidentialGene tr2aacds pipeline (http://arthropods.eugenes.org/EvidentialGene/trassembly.html) to merge all the assemblies to remove redundancy and to get biologically most useful set of transcripts.

We then evaluated the obtained set of primary transcripts using TransRate v1.0.3 (Smith-Unna *et al.*, 2016) and BUSCO v3- Benchmarking Universal Single-Copy Orthologs (Simão *et al.*, 2015), which reports basic summary statistics (like n50, % reads mapped, etc.) and checks for the completeness of the transcriptome respectively. For annotation of this *de novo* assembled transcriptome, we blasted the transcripts against the NR database with an e-value threshold of 10^{-6} and other default parameters and used only the top 20 hits for annotation. Additionally, we

identified conserved protein domains by searching through the InterPro collection of databases. We used the results from these to functionally annotate using BLAST2GO v4.1.9 (Conesa *et al.*, 2005) by identification of Gene Ontology (GO) Slim terms and KEGG pathways.

References:

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29: 644–652.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28: 1086–1092.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome research* 26: 1134–1144.

Song L, Florea L. 2015. Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* 4: 48.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18: 821–829.

Figures

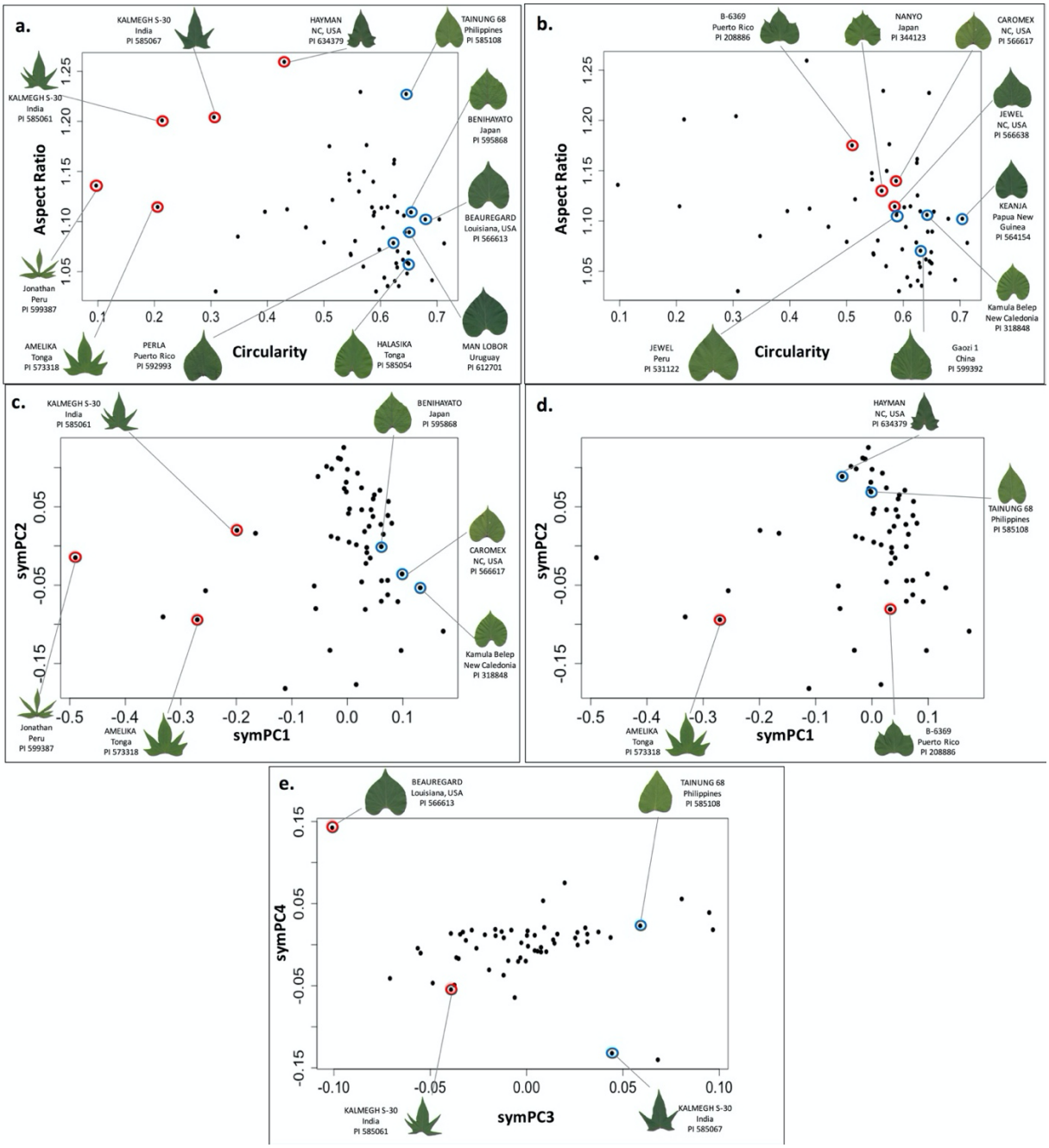

Figure S2-1: Green-house grown accessions selected for transcriptomic analysis for leaf shape traits.
a. Circularity, b. Aspect Ratio, c. symPC1, d. symPC2, and e. symPC3. Red circles represent the accessions chosen for low ends of the trait spectrum and blue circle represents the accessions chosen for high ends of the trait spectrum.

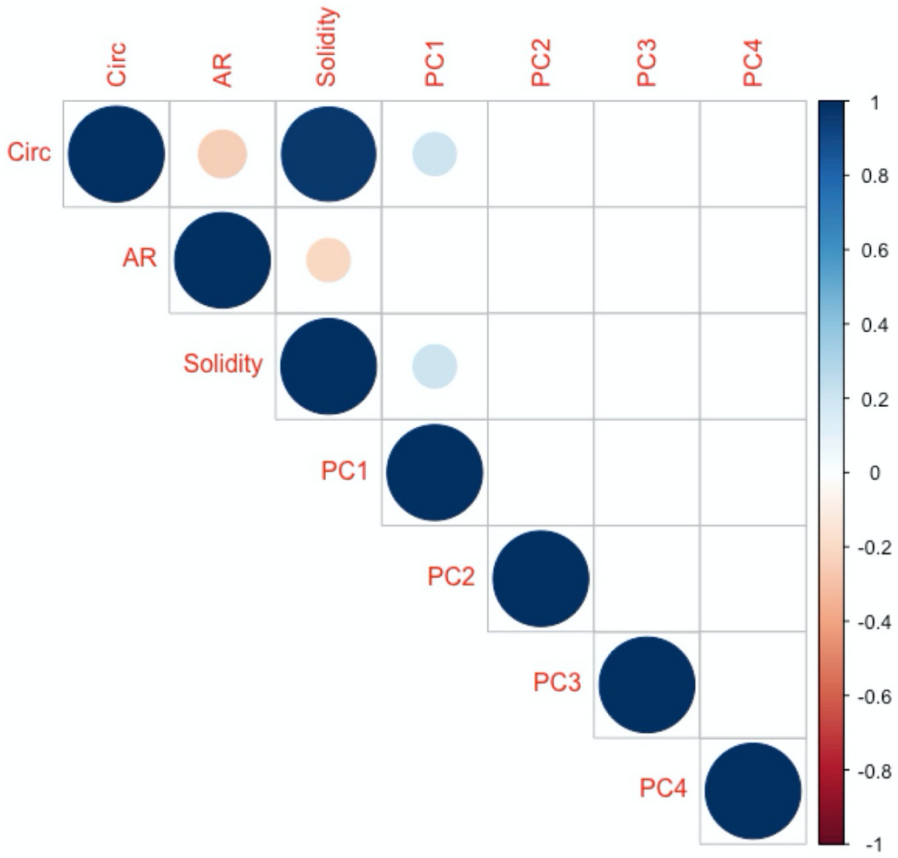

Figure S2-2: Correlation plot between leaf shape traits.
Correlation between traditional and EFD PCs showing that only symPC1 is slightly correlated with circularity and solidity; other traditional leaf shape traits (circularity, aspect ratio and solidity) are not correlated with symPCs.

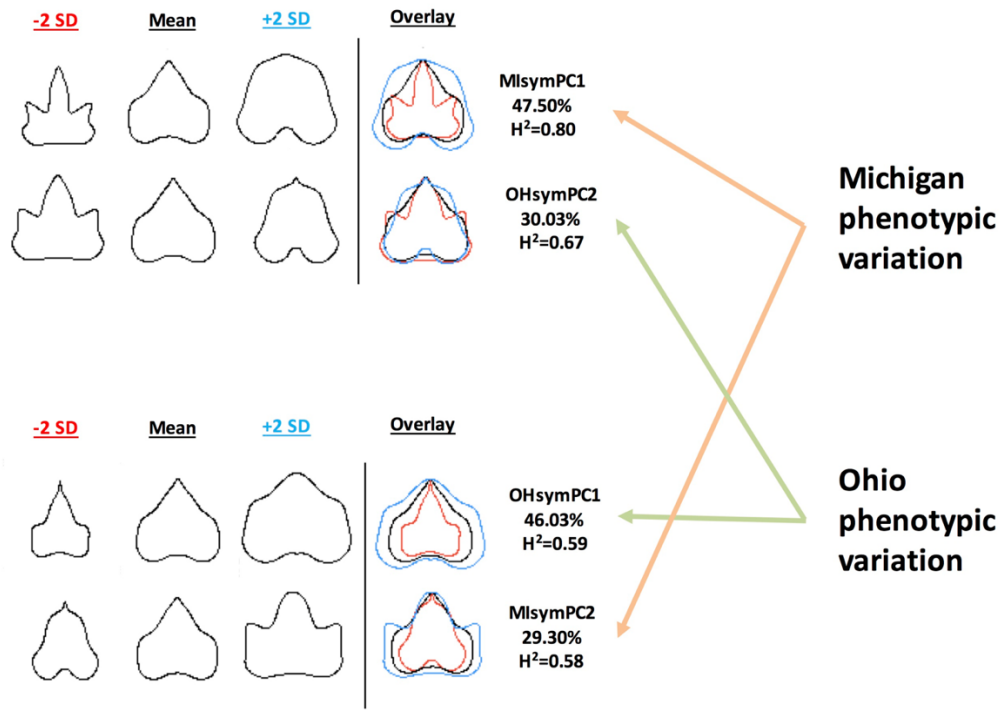

Figure S2-3 Leaf shape variation captured by EFDs from MI and OH differing significantly in their order of variation explained.
MIsymPC1 explains variation in leaf shape that is attributed to lobing, tip and petiolar sinus differences, similar to OHsymPC2 (which only explains ~30% of the variation in OH).

Tables

Table S2-1 Accession IDs with their source and location of origin used in this study.
(as separate file)*

Table S2-2 Differentially expressed transcripts associated with leaf shape traits found in this study
(as separate file)*

Table S2-3 Raw read counts of orthologs of homeobox domain genes within the assembled transcriptomes, for accessions chosen for circularity RNA-Seq analysis
(as separate file)*

**File containing Tables S2-1:S2-3 is linked here*

Appendix B

Supplementary Figures and Tables for Chapter 3

Figures

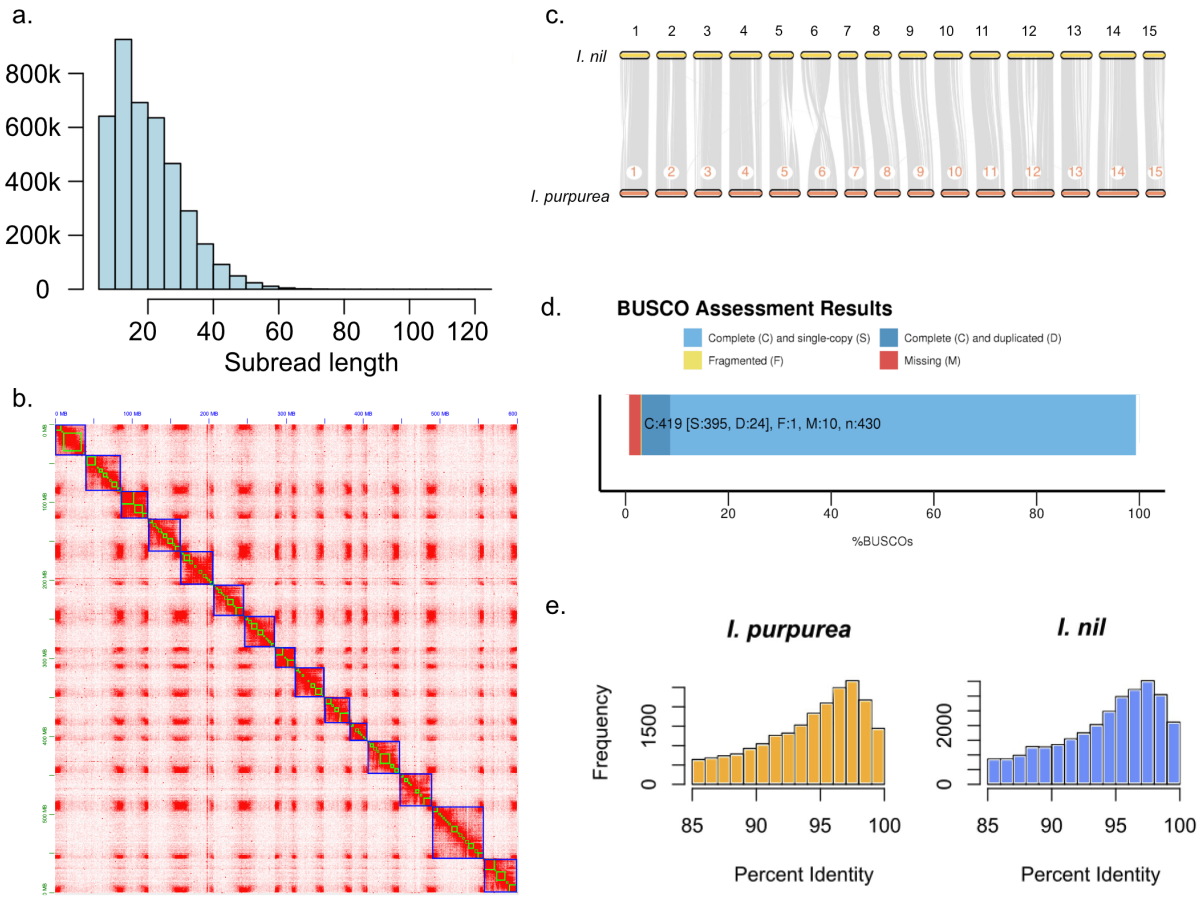

Figure S3-1 Chromosome scaffolding and renaming.
a) Raw PacBio Sequel filtered subread lengths. b) Phase Genomics Hi-C Proximo scaffolding results produces 15 chromosome pseudomolecules. c) Synteny of this *Ipomoea purpurea* assembly against *Ipomoea nil* was used to orient and name *I. purpurea* pseudomolecules. d) BUSCO results against Viridiplantae odb10 indicate high completeness of conserved gene sets in the raw assembly. e) Retrotransposon annotations using LTRharvest were used to compute pairwise LTR identities within LTRharvest for both *I. purpurea* and *I. nil*.

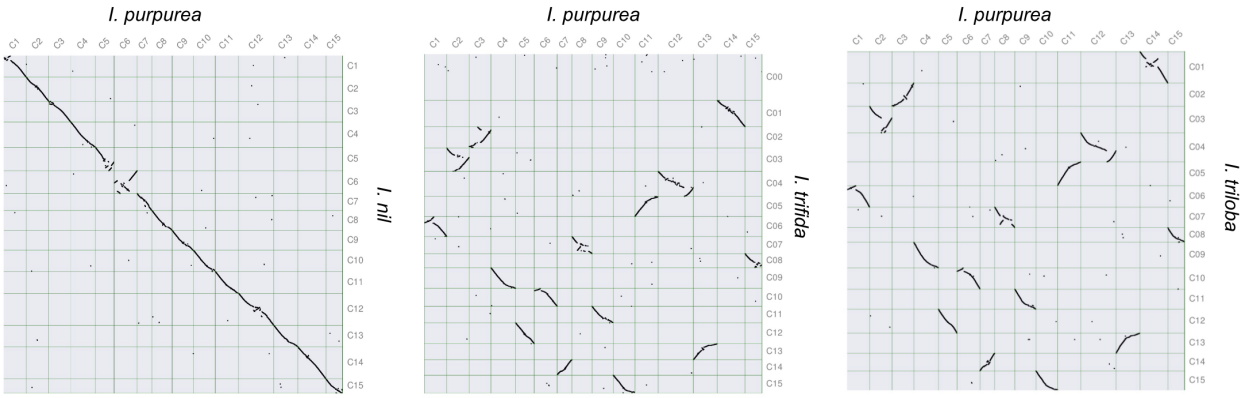

Figure S3-2 Synteny of the *I. purpurea* genome against related Convolvulaceae species, including *I. nil*, *I. trifida*, and *I. triloba*.

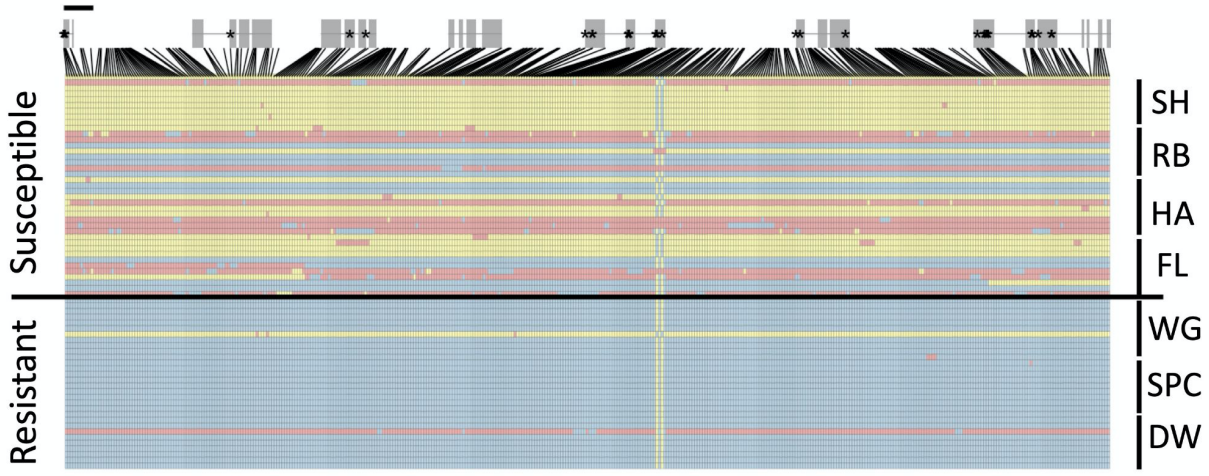

Figure S3-3 Signs of selection across conserved haplotype of multiple glycosyltransferases for each individual on Chromosome10.
Exons are shown in grey. Blue and yellow indicate homozygotes, red indicates heterozygotes; stars indicate non-synonymous substitutions. Black bar above gene models indicate 1kb.

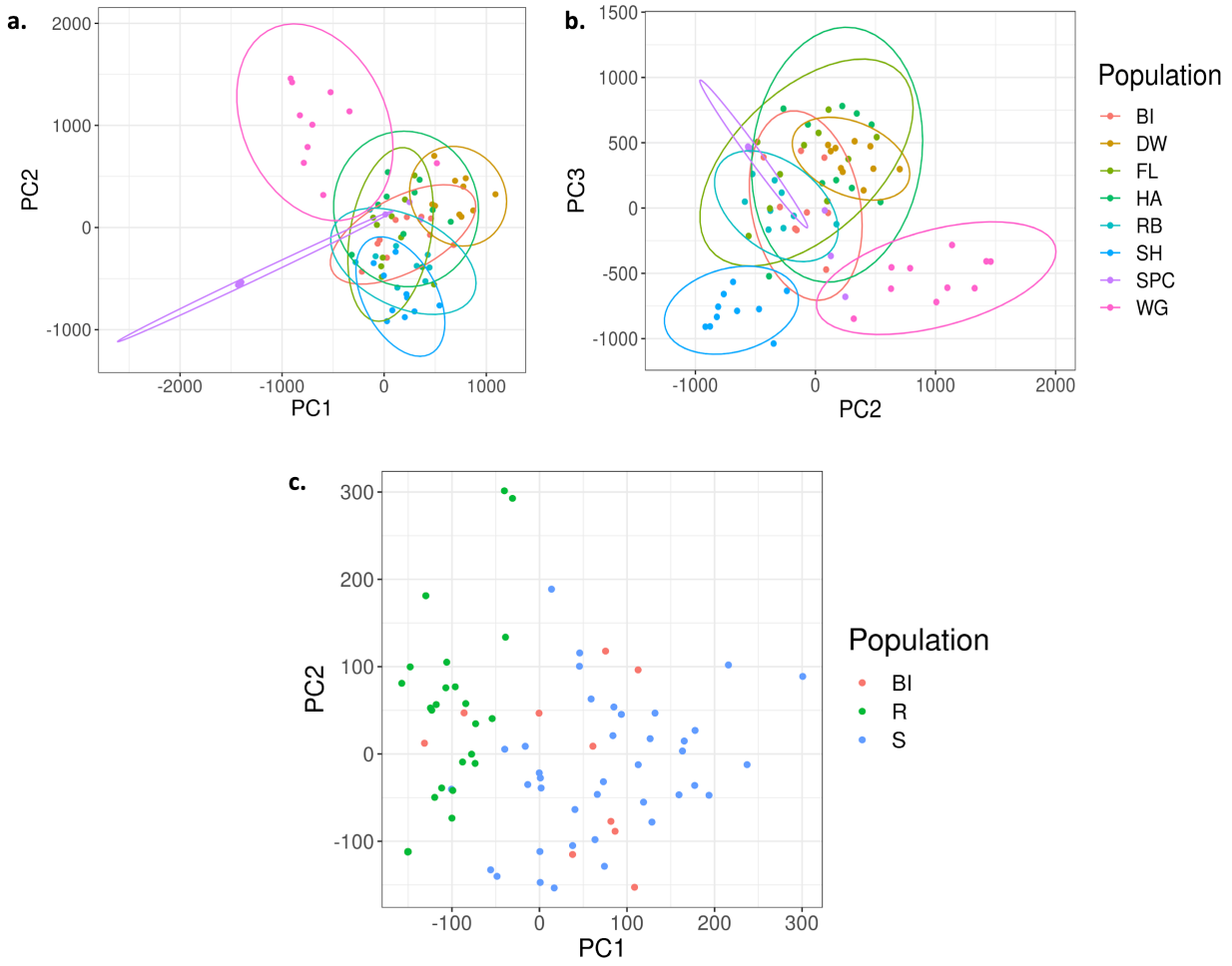

Figure S3-4 PCA of resistant and susceptible populations used.
Individuals from the sampled populations do not cluster into distinct resistant and susceptible groups when using all the SNPs (a and b), but there is some grouping when only considering the SNPs from the regions under selection (c).

Tables

Table S3-1 SNP outliers (in the top 5% BF and 5% Rho) identified via bayenv2
(as separate file)*

Table S3-2 Functional annotation of genes present within +/- 5KB of bayenv2 SNP outliers
(as separate file)*

Table S3-3 Summary of the genome-wide regions under selection via the Md-rank-P
(as separate file)*

Table S3-4 Functional annotations of genes under selection identified via md-rank-P approach
(as separate file)*

Table S3-5 Functional annotations of genes under selection identified via bayenv2 and md-rank-P approach
(as separate file)*

Table S3-6 Mutations and their effects present within the genes of interest
(as separate file)*

Table S3-7 List of differentially expressed genes between treated (herbicide sprayed) resistant vs susceptible individuals
(as separate file)*

Table S3-8 List of differentially expressed genes between control (non-herbicide sprayed) resistant vs susceptible individuals
(as separate file)*

Contrast	Contrast Estimate	t-ratio	P-value
Malathion vs Glyphosate	0.455	3.913	0.0007
Malathion vs Malathion-Glyphosate	0.834	8.200	<0.0001
Malathion vs Control	-0.460	-5.342	<0.0001
Glyphosate vs Malathion-Glyphosate	0.379	2.946	0.0190
Glyphosate vs Control	-0.915	-7.834	<0.0001
Malathion-Glyphosate vs Control	-1.29	-12.654	<0.0001

Table S3-9 Pairwise contrast statistics for normalized above-ground biomass between the four treatment conditions.
These were calculated using the lsmeans function in R, with P-values adjusted for multiple tests using tukey correction.

Table S3-10 ILD summary statistic (99th percentile value and max r2) for the five regions under selection that exhibited GST > 0.39.
The summary is reported only for only the SNPs within regions under selection. The 99th percentile reports the top 1% of r2 values whereas the max r2 value is the highest r2 value in the region
(as separate file)*

Table S3-11 Individual ILD interactions, above the 99 percentile cutoff r2 value, for SNPs within the region under selection
(as separate file)*

**File containing Tables S3-1:S3-8, S3-10:S3-11 is linked here.*

Pop Abbrev	Resistance Type	State	Proportion survival at 1.7	Latitude	Longitude	No of individuals sampled
BI	R	TN	1	35.775	-85.903	10
DW	R	NC	1	34.983	-78.039	10
FL	S	SC	0.20	34.145	-79.865	10
HA	S	NC	0.15	35.424	-77.917	10
RB	S	TN	0.18	35.316	-87.353	9
SH	S	VA	0.1	38.373	-78.662	10
SPC	R	TN	0.71	35.533	-85.951	10
WG	R	TN	0.83	35.099	-86.225	10

Table S3-12 Population information for each population used in the study.
Pop Abbrev = abbreviation for each population as used in Kuester *et al.* 2015, Resistance type = classification of resistance in the population R >0.5 prop. survival S <0.5 prop. survival, State = state where seeds were collected, Proportion survival at 1.7 = proportion of individuals that survived a spray rate of 1.7 kg/ha of glyphosate based on Kuester et al 2015, Latitude and Longitude = location where seeds were collected.

Sample name	Pop Abbrev	Resistance Type	TRT
IP_438	WG	R	Control
IP_447	WG	R	Control
IP_235	WG	R	Herbicide
IP_244	DW	R	Herbicide
IP_247	WG	R	Herbicide
IP_248	WG	R	Herbicide
IP_252	DW	R	Herbicide
IP_261	WG	R	Herbicide
IP_459	RB	S	Control
IP_477	SH	S	Control
IP_177	HA	S	Herbicide
IP_188	RB	S	Herbicide
IP_189	HA	S	Herbicide
IP_232	HA	S	Herbicide
IP_257	RB	S	Herbicide
IP_260	RB	S	Herbicide
IP_497	SH	S	Herbicide

Table S3-13 RNA-Seq sample information used in the study.
Pop Abbrev = abbreviation for each population as used in previous studies, Resistance type = classification of resistance in the population R >0.5 prop. survival S <0.5 prop. survival,TRT = Treatment type based on either herbicide sprayed (Herbicide) or not sprayed (Control).

Pop Abbrev	Resistance Type	TRT	Sample size
BI	R	Malathion	6
DW	R	Malathion	13
WG	R	Malathion	22
HA	S	Malathion	4
IN	S	Malathion	3
RB	S	Malathion	11
SH	S	Malathion	3
BI	R	Glyphosate	3
DW	R	Glyphosate	5
WG	R	Glyphosate	8
HA	S	Glyphosate	3
RB	S	Glyphosate	4
BI	R	Glyphosate + Malathion	2
DW	R	Glyphosate + Malathion	9
WG	R	Glyphosate + Malathion	13
HA	S	Glyphosate + Malathion	4
RB	S	Glyphosate + Malathion	6
BI	R	Control	5
DW	R	Control	12
WG	R	Control	20
HA	S	Control	8
IN	S	Control	3
RB	S	Control	10
SH	S	Control	3

Table S3-14 Sample information used for the malathion assay.
Pop Abbrev = abbreviation for each population as used in previous studies, Resistance type = classification of resistance in the population R >0.5 prop. survival S <0.5 prop. survival,TRT = Treatment type, Sample size = Total number of individuals per population per resistance type per treatment.