

The Evolution of the Glucosinolate Pathway in the Brassicaceae

by

Carrie F. Olson-Manning

Department of Biology
Duke University

Date: _____

Approved:

Thomas Mitchell-Olds, Supervisor

Xinnian Dong

Fred Nijout

Mark Rausher

Jane Richardson

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

2013

ABSTRACT

The Evolution of the Glucosinolate Pathway in the Brassicaceae

by

Carrie F. Olson-Manning

Department of Biology
Duke University

Date: _____

Approved:

Thomas Mitchell-Olds, Supervisor

Xinnian Dong

Fred Nijout

Mark Rausher

Jane Richardson

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

2013

Copyright by
Carrie F. Olson-Manning
2013

Abstract

Understanding the mechanisms that underlie the formation of, and innovation in biochemical pathways is an important goal in evolutionary biology. The following work addresses the problem of biochemical pathway evolution in two ways. In the first chapter, I combine genetic manipulations and population genetic analyses to investigate whether flux control in the aliphatic glucosinolate pathway of *Arabidopsis thaliana* drives evolutionary rate heterogeneity. My results indicate that the first enzyme in the pathway, CYP79F1, has majority flux control and is the only one to show convincing evidence for positive selection. The second chapter builds on the first by asking whether flux control is stable under a variety of environmental conditions. I find that flux control remains with CYP79F1 in all my environmental treatments. In the final chapter, I address the evolution of one enzyme in this pathway from *Boechera stricta* that is responsible for a gain-in-function polymorphism that results in increased fitness in nature. With molecular phylogenetic analysis, site-directed mutagenesis, structural biology and enzymatic assays, I determine what residues are under selection and test their functional effects. I find that just two mutations in this enzyme are responsible for the change in function, and discuss their position within the enzyme with respect to function. Strikingly, the enzyme with majority flux control in *A. thaliana* is homologous to the enzyme responsible for the novel function in *Boechera*. Together these results

suggest that selection may predictably exploit the same small subset of genes to optimize biochemical pathway output and for evolutionary innovation.

Dedication

To my wonderful and supportive parents, Rick and Janet.

Contents

Abstract.....	iv
List of Tables.....	ix
List of Figures.....	xi
Acknowledgements.....	xiii
1. Introduction.....	1
1.1 The glucosinolate pathway.....	3
1.2 Review of Brassicaceae glucosinolate distribution.....	5
2. Evolution of flux control in the glucosinolate pathway in <i>Arabidopsis thaliana</i>	8
2.1 Results and Discussion.....	11
2.1.1 Estimation of flux control in the glucosinolate pathway.....	13
2.1.2 Relative substitution rates.....	17
2.2 Conclusions.....	23
2.3 Materials and Methods.....	24
2.3.1 Insertion lines.....	24
2.3.2 Plant growth conditions:.....	24
2.3.3 Genotyping.....	25
2.3.5 Analysis of GLS Concentration:.....	26
2.3.6 Sequence analysis.....	27
2.3.7 Analysis of flux control.....	30
Statistical Analysis.....	32
3. Flux control under different environmental conditions.....	34

3.1 Results	37
3.2 Discussion.....	50
3.2.1 Flux control under environmental variation.....	50
3.2.2 Genotype by environment interactions.....	51
3.2.3 Influence of flux control on generalist herbivore feeding	52
3.2.4 Gene expression and glucosinolate profile.....	53
3.3 Conclusions	53
3.4 Materials and Methods.....	55
3.4.1 Materials and Methods for Statistical Analyses.....	57
3.4.2 Gene expression meta-analysis	60
4. The evolution of a novel glucosinolate in <i>Boechera stricta</i>	61
4.1 QTL cloning and BCMA function.....	63
4.2 Gene phylogeny and biochemical evolution.....	64
4.3 Discussion and Conclusions	70
4.4 Materials and Methods	72
Appendix A – Derivation of flux control coefficients.....	81
Appendix B – Supplemental Tables and Figures	83
References.....	95
Biography	105

List of Tables

Table 1: Abbreviation for glucosinolates examined in the following studies	5
Table 2: Relative Expression Ratio, R_i , and 95% CI. Bolded values were significantly different than 1.	12
Table 3: Estimated Flux Control Coefficients Calculated Using All Three Lines for WT Individuals and 95% Bootstrap CIs After 1,000 Bootstrapping Runs	15
Table 4: McDonald-Kreitman test of polymorphism in <i>A.thaliana</i> compared with divergence from <i>A.lyrata</i> . ^a Significance values based on permutation test.	18
Table 5: Synonymous and nonsynonymous within species variation of <i>A.thaliana</i> and between species variation of <i>A.thaliana</i> compared to <i>A.lyrata</i> . <i>A.lyrata</i> has no ortholog of GSTU20.	20
Table 6: MANOVA Effect of genotype and environment on glucosinolate concentration	38
Table 7: Estimated flux control coefficients calculated using all three lines for WT individuals and 95% Cis after 1,000 Bootstrapping Runs.	42
Table 8: Univariate estimates of the effect of genotype on glucosinolate concentration. Means and standard errors are reported for untransformed data, while P- are estimated from the log-transformed data. Standard errors are shown in parentheses.	43
Table 9: Univariate estimates of the effect of environmental treatments on glucosinolate concentration. Means and standard errors are reported for untransformed data, while P- are estimated from the log-transformed data. Standard errors are shown in parentheses.	44
Table 10: P-values from log-transformed data derived from ANOVA of amount of leaf area removed with * indicating $P < 0.05$	45
Table 11: Primers used for genotyping tDNA insertion lines	83
Table 12: Primers for qrtPCR	84
Table 13: Mean and standard error for each glucosinolate compound for each insertion line	85

Table 14: Univariate estimates of changes GLS concentration.	86
Table 15: Tajima's D and Normalized Fay and Wu's H.....	88
Table 16: Likelihood ratio tests for hypothesis of accelerated biochemical evolution using PAML. Tree branches are identified by letters in Figure 13.	89
Table 17: Heterologous expression experiments showing in vitro activity of BCMA wild type and mutant enzymes towards alternate substrates. Each cell presents information on mean enzymatic activity (in nmol product/ minute/ nmol cytochrome P450), t-statistic, P-value, and N. Initial fixed effects ANOVA of log expression level showed highly significant interaction between constructs and substrates ($F = 7.36$; $df = 9, 62$; $P < 10^{-6}$; $R\text{-square} = 97.4\%$; $N = 82$), hence individual constructs are compared to controls at $\alpha = 0.05$. Constructs with significantly higher activity than the control are indicated in bold. BCMA2 alleles have identical amino acid sequence, so only one allele was assayed. Cyp79F genes are from <i>Arabidopsis</i> . Constructs with site directed mutagenesis were derived from BCMA2, and are indicated by G134L, P536K, or Both. BCMA genes from the two parental genotypes or <i>Arabidopsis thaliana</i> are compared to the empty vector control, while genes modified by site directed mutagenesis are compared to BCMA2.	90
Table 18: Primer sequences relevant to Chapter 4	91

List of Figures

- Figure 1: Core A) aliphatic and B) indolic pathways in *A.thaliana*. The enzyme that catalyzes each reaction is found to the right of the arrow in bold. The glucosinolate name is in dark grey (Table 1) and the generic compound name is in light grey italics in the center of the pathways. 4
- Figure 2: Structure of the generic glucosinolate A) and the relevant aliphatic amino acid R groups: chain-elongated methionine B), isoleucine C), and valine D). 5
- Figure 3: Unrooted species tree of relevant Brassicaceae. Dark grey circles indicate the production of branched-chain derived GLS (IME and 1MP) and light grey circles indicate methionine derived GLS. 7
- Figure 4: qRT PCR comparison of heterozygous (HET) to wild type (WT) genotypes of the different insertion lines. UBQ10 was used as the reference gene to normalize the glucosinolate gene (abbreviations on horizontal axis) of interest in each RNA extraction. All comparison between the expression of HET (dark grey bars) and WT (light grey bars) are significantly different except GSTF11 and GSTU20. The vertical hashes indicate standard error bars. Asterisks above each pair of bars show relative expression ratio of less than one, as judged by appropriate contrasts in ANOVA. 13
- Figure 5: Heatmap of the univariate estimates of the proportional change in GLS concentration of the HET compared to the WT. Environmental conditions are pooled in this analysis, but the untreated controls exhibit the same pattern of GLS change (data not shown). Cool colors indicate a decrease in concentration in the HET compared to the WT and warm colors indicate an increase. * indicate level of significance of the log transformed data. *** $p < 0.0001$, ** $p < 0.01$, * $p < 0.05$ 46
- Figure 6: The direction of glucosinolate concentration change for the aliphatic and indolic glucosinolate products. The proportional change in glucosinolate concentration of the seven glucosinolate products (abbreviations on the left) following four environmental treatments (W=water deprivation, C=leaf crushing, S=soil nutrient deprivation, J=methyl Jasmonate treatment). Warm colors indicate higher concentration in treatments than controls, and cool colors indicate decreased concentration in treatments. Stars indicate level of significance *** < 0.001 , ** < 0.01 , * < 0.05 . Genes are pooled in this analysis, but all genes produce the same pattern as above (data now shown). 47
- Figure 7: Correlation between the total aliphatic GLS concentration and amount of leaf area removed by *T.ni*. There is a significant correlation ($P=0.01940$, based on log

transformed data damage levels) for the amount of leaf area removed and total aliphatic GLS concentration, but no significant correlation with total indolic concentration ($P=0.3609$). Dotted line is the best-fit line fit to untransformed data. 48

Figure 8: Average relative expression of genes in the A) aliphatic and B) indolic glucosinolate pathways after methyl jasmonate application. Bars are the mean of three different EMBL Gene Expression Atlas experiments and horizontal lines represent standard error. All genes were significantly different from controls in at least two of three experiments. 49

Figure 9: Gene tree of homologs of CYP79F1. Light grey shading is MET-specific BCMA2 clade and dark grey is the BCMA1 and BCMA3 clade (ILE and VAL specific, respectively). The unshaded taxa make only MET aliphatic GLS. 67

Figure 10: In vitro enzyme activity levels (nmol of product per nmol of enzyme per minute) relative to controls; error bars denote SE. Labels indicate CYP79F enzymes from *Arabidopsis* and BCMA1, BCMA2, and BCMA3 from *Boechera*, with alleles from Colorado or Montana. BCMA1 and BCMA3 gained VAL activity (dark grey). BCMA2 alleles encode identical proteins, so one allele was assayed. BCMA2 (light grey) retains the ancestral MET activity and was engineered to change G134L, P536K, or both (white). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ 68

Figure 11: Homology model of BCMA2, a CYP79F4 enzyme responsible for fitness variation in nature. (A) Homology model of BCMA2 with the substrate binding cleft above the heme group (magenta). Amino acid changes G134L and P536K (in yellow) show statistical evidence for accelerated protein evolution, and alter catalytic function when changed by site-directed mutagenesis. G148L and M268V are differences between BCMA1 and BCMA3 (light blue). The location of amino acid 529 is colored, as 530-541 are not depicted in the model. (B) Close up view of substrate binding cleft with mutation G134L residing just above the heme, but G148L and M268V in the substrate recognition regions (in purple). 69

Figure 12: Genome wide distributions of Tajima's D and Fay and Wu's H values for *A.thaliana*. 93

Figure 13: Phylogenetic tree of Cyp79 genes. Shown are sequences from *B. stricta* (Bstr), *B. retrofracta* (Bret), *A. thaliana* (Athal), and *Thellungiella halophila* (Thal). Tree is the consensus from Bayesian inference, maximum likelihood, and maximum parsimony, with bootstrap support indicated from top to bottom, respectively. Grey italics letters denote branches tested in PAML analysis (Table 16). 94

Acknowledgements

First, I offer my deepest thanks to my advisor, Tom Mitchell-Olds, for your intellectual and financial support. I was very lucky to find an advisor who encouraged me pursue my interests and supported my (in retrospect) brash and naïve dabbling in P450s. I cannot count the number of exciting opportunities I've been afforded working you; thank you for the advice and guidance during my training as a scientist.

Thank you to all the members of my committee. I would especially like to thank Mark Rausher, my unofficial co-advisor. He knows how often I've pestered him with my questions, but he may not realize how influential he has been on my thinking as a scientist. Thank you, Fred Nijout for your seemingly endless knowledge of empirical methods that helped give me the confidence to dive into many of my most difficult projects. Thank you, Xinnian Dong for helping me see my project from a different perspective and offering constructive criticism. And finally, thank you, Jane Richardson. While you have only officially been on my committee for a short time, you have been guiding me through the new and exciting world of biochemical evolution since I first began graduate school.

I also need to thank so many at Duke for your support over the years: Anne Lacey, Jim Tunney, Jo Bernhardt, Randy Smith, Jill Foster and so many more have made the administrative part of grad school a breeze.

I want to thank the past and current members of the Mitchell-Olds lab. Cheng-Ruei, Prasad, Bao-Hua and Kathy offered so much help over the years. I'd also like to thank the Schuler and Halkier labs and Carrie Wessinger for helping me get into the wonderful and terrible world of P450s.

I want to especially thank Jill Anderson for our morning water breaks, the wonderful science and life discussions, the jokes that are funny 100% of the time, and for generally being the perfect mentor. To Maggie Wagner, you made an uncountable number of experiences not only bearable, but also fun. To Cathy Rushworth, thank you for our Friday afternoon chats/crafting sessions.

Finally, to my life outside academia I have many people to thank. First I have to thank one animal, my running and spooning partner, Buddy. When I think I'm stressed out or think that my problems are important, he was always there to remind me that the world is really all about him. Silly me.

I want to give a special thanks to the all the Biograds who have made grad school so much fun. Thanks to Katie Ferris, Jessica Selby, Amanda Grusz, Caiti Heil, and Carrie Wessinger. I don't know what I would have done without the themed parties and Shooters II expeditions. I regret nothing. I want to especially thank "the band" and fellow slip-n-slide enthusiasts, Paul Durst, Matt Johnson, Chris Iacoboni, David Rasmussen and relative newcomers Marissa Lee and Amanda Lea. I can't imagine making it through grad school without all The Terrace-induced shenanigans.

My parents Rick and Janet, to whom this dissertation is dedicated, could not have been more supportive of my dreams. Thank you to my siblings Amanda, Emma and Joey; your incredible passion for your chosen fields inspires me every day and makes me work harder.

And finally to my wonderful husband, Andy. I feel privileged that someone who is as brilliant and loving as you would choose someone like me. Thank you gently (or not so gently) coaxing me through the roughest times in grad school. Thank you for too often taking on too many of the responsibilities around the house. I'm so happy I get to share this with you.

1. Introduction

Most cellular processes and organismal phenotypes rely on metabolic pathways and regulatory networks. The genomics era has resulted in a slew of molecular sequence patterns that seem to be common in these pathways and networks, (as reviewed in (Olson-Manning *et al.*, 2012)). For example, genes central in regulatory networks tend to evolve more slowly than genes on the periphery (Casals *et al.*, 2011, Jovelin & Phillips, 2011, Luisi *et al.*, 2012, Alvarez-Ponce *et al.*, 2009, Alvarez-Ponce *et al.*, 2011). In biochemical pathways, upstream genes tend to be the targets of selection rather than downstream genes (Livingstone & Anderson, 2009, Lu & Rausher, 2003, Ma *et al.*, 2010, Rausher *et al.*, 2008, Rausher *et al.*, 1999, Ramsay *et al.*, 2009). However, there is a dearth of empirical evidence that explains whether certain genes are more likely to experience adaptive evolution and the molecular constraints and mechanisms that may lead to these patterns. In this thesis, I study the evolution of the genes in the glucosinolate (GLS) pathway in the Brassicaceae, first at the whole-pathway level, and then a gain-of-function in single enzyme in the pathway.

The remainder of Chapter 1 describes the structure of the glucosinolate pathway and the distribution of glucosinolates in the Brassicaceae.

Chapter 2 describes how flux control, a property of an enzyme in a biochemical pathway, influences rate variation in the aliphatic glucosinolate pathway in *Arabidopsis thaliana*. Many studies have found that upstream genes in pathways are more often the

targets of natural selection (Livingstone & Anderson, 2009, Lu & Rausher, 2003, Ma et al., 2010, Rausher et al., 2008, Rausher et al., 1999, Ramsay et al., 2009), often invoking flux control or pleiotropy to explain this pattern. Here we present evidence that the first step in the aliphatic GLS pathway has majority flux control and that natural selection has produced a strong signature of positive in this gene.

Chapter 3 asks whether flux control in the aliphatic glucosinolate pathway remains in the same enzyme under a range of environmental conditions. A variety of environmental treatments (water and soil nutrient deprivation, leaf wounding and methyl jasmonate treatments) are known to influence the quantity and quality of glucosinolate profiles in close relatives of *A. thaliana*. We find that while each of the environmental treatments has a significant impact on the quantity of glucosinolates and the proportions of each type that are produced, majority flux control remains with the first enzyme in the pathway. Taken together, Chapters 2 and 3 suggest that mutations that occur in the beginning of a biochemical pathway are most likely to have an influence on phenotype regardless of environment and that natural selection will preferentially act on these genes.

Finally, Chapter 4 delves into the evolution of a novel glucosinolate in *Boechera stricta*, a close relative of *A. thaliana*. Natural populations of *B. stricta* segregate for different types of aliphatic glucosinolate, and this polymorphism has a large impact on fitness in natural habitats. Using quantitative trait locus (QTL) mapping, our lab

determined the region responsible, named *BCMA*, and subsequently the gene responsible for this polymorphism. Field studies that show that this locus is under selection in the field are summarized. I also describe in detail the molecular signatures of selection acting on this gene, the heterologous expression of the enzyme and biochemical characterization and the mutations responsible for the gain in function of this enzyme.

1.1 The glucosinolate pathway

Glucosinolates (GLS), or mustard oil glycosides, are biologically active secondary compounds found in *Arabidopsis* and its relatives (Halkier & Gershenzon, 2006, Fan *et al.*, 2011, Hopkins *et al.*, 2009). They are derived from amino acid precursors (methionine [depicted in Figure 1], isoleucine and leucine as part of the aliphatic GLS pathway [Figure 2], and tryptophan as part of the indolic pathway GLS [Figure 1]) and are stored in the vacuole (Grubb *et al.*, 2004). Intact GLSs have low toxicity until a leaf is damaged and they come in contact with the hydrolytic enzyme myrosinase, resulting in the production of compounds (isothiocyanates, nitriles, thiocyanates, epithionitriles, oxazolidine-2-thiones) that slow growth and development of herbivores (Blau *et al.*, 1978, Kliebenstein *et al.*, 2005). GLSs influence oviposition and feeding by many insect herbivores (Hopkins *et al.*, 2009), function in defense against microbial pathogens (Brader *et al.*, 2006), and affect the composition of associated microbial communities (Bressan *et al.*, 2009). Typically, generalist insects are sensitive to glucosinolate-based

plant defenses, whereas specialists may be able to cope with these compounds, which can even serve as oviposition cues and feeding stimulants (Hopkins et al., 2009).

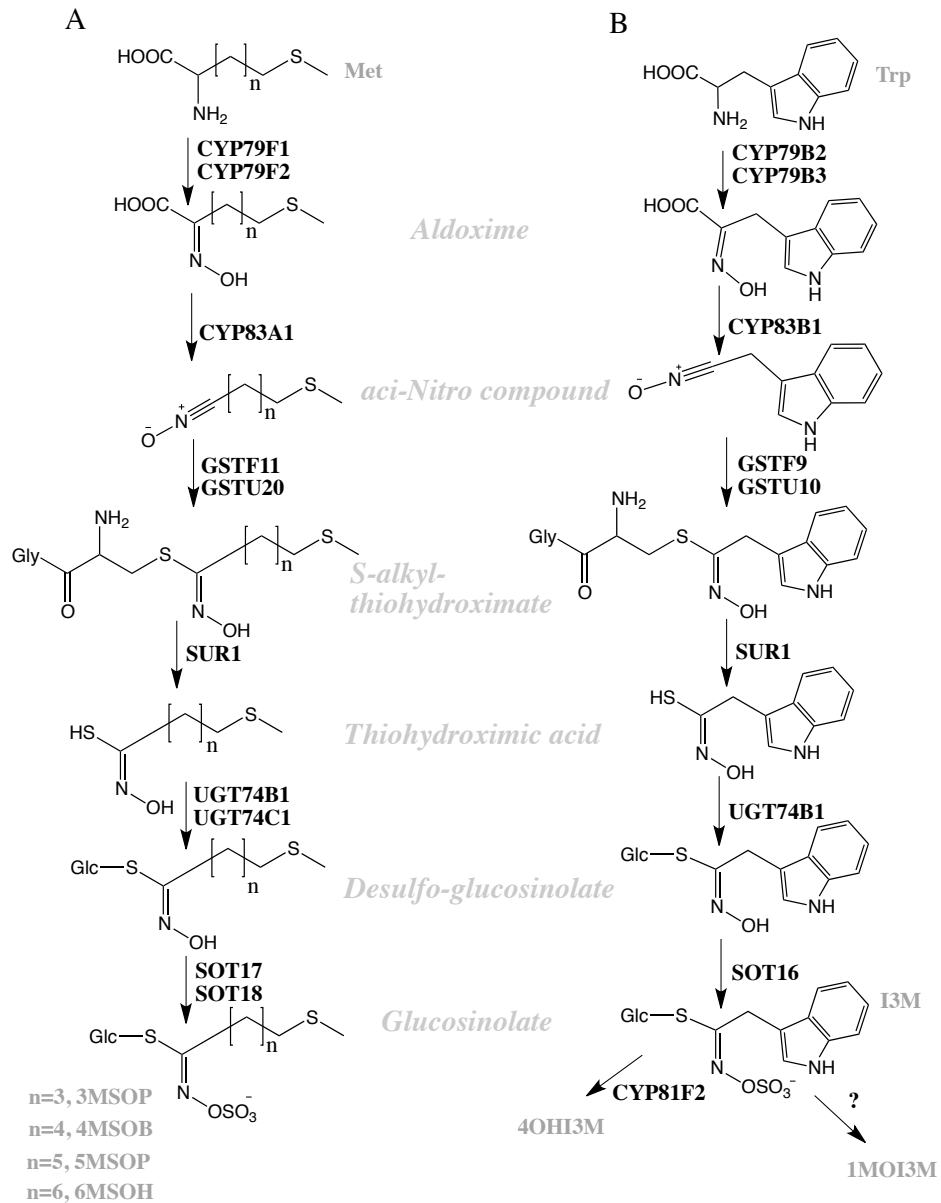


Figure 1: Core A) aliphatic and B) indolic pathways in *A.thaliana*. The enzyme that catalyzes each reaction is found to the right of the arrow in bold. The glucosinolate name is in dark grey (Table 1) and the generic compound name is in light grey italics in the center of the pathways.

Table 1: Abbreviation for glucosinolates examined in the following studies

Abbreviation	Glucosinolate	Pathway	Amino acid precursor
3MSOP	3-methylsulfinylpropyl	Aliphatic	Homomethionine
4MSOB	4-methylsulfinylbutyl	Aliphatic	Dihomomethionine
5MSOP	5-methylsulfinylpentyl	Aliphatic	Trihomomethionine
6MSOH	5-methylsulfinylhexyl	Aliphatic	Tetrahomomethionine
1ME	1-methylethyl	Aliphatic	Valine
1MP	1-methylpropyl	Aliphatic	Isoleucine
I3M	Indolylmethyl	Indolic	Tryptophan
4OHI3M	4-hydroxy-3-indolylmethyl	Indolic	Tryptophan
1MOI3M	4-methoxy-3-indolylmethyl	Indolic	Tryptophan

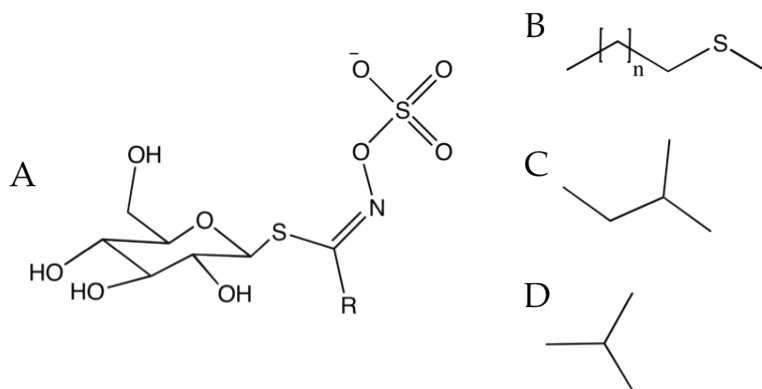


Figure 2: Structure of the generic glucosinolate A) and the relevant aliphatic amino acid R groups: chain-elongated methionine B), isoleucine C), and valine D).

1.2 Review of Brassicaceae glucosinolate distribution

Glucosinolates are found exclusively in order Brassicales (and in one unrelated genus, *Drypetes*, family Euphorbiaceae), especially in the families Brassicaceae and Capparaceae (Rodman *et al.*, 1998). Hundreds of GLS have been characterized

(Daxenbichler *et al.*, 1991, Fahey *et al.*, 2001), but they all derive from seven amino acids (PHE, TYR, ALA, MET, ILE, VAL, TRP). The majority of the species in the Brassicaceae produce mostly MET, TRP and some PHE derived GLS.

However, we found that some *Boechera stricta*, a close relative of *Arabidopsis thaliana*, produce uncommon GLSs derived from the branched chain (BC) amino acids, ILE and VAL (1MP and 1ME, respectively). To better understand the distribution of glucosinolates in the *Boechera* clade, I used high-pressure liquid chromatography (HPLC), to survey the glucosinolate profile for many species in the Brassicaceae, with a focus on species in the genus *Boechera*. The unrooted species phylogeny (Alexander *et al.*, 2010) depicts the relationship of the species studied here, and the aliphatic glucosinolates they produce as either light grey (MET-derived) or dark grey (BC-derived) circles (Figure 3). I find that only species within the *Boechera* clade produce substantial amounts of BC-derived glucosinolates. In Chapter 4, I discuss the evolution of the BC-GLS, but it is important to note that most species in the Brassicaceae produce only MET-derived aliphatic GLS. In Chapter 4 I discuss in detail the evolution of novel BC-amino acid GLS in *Boechera*.

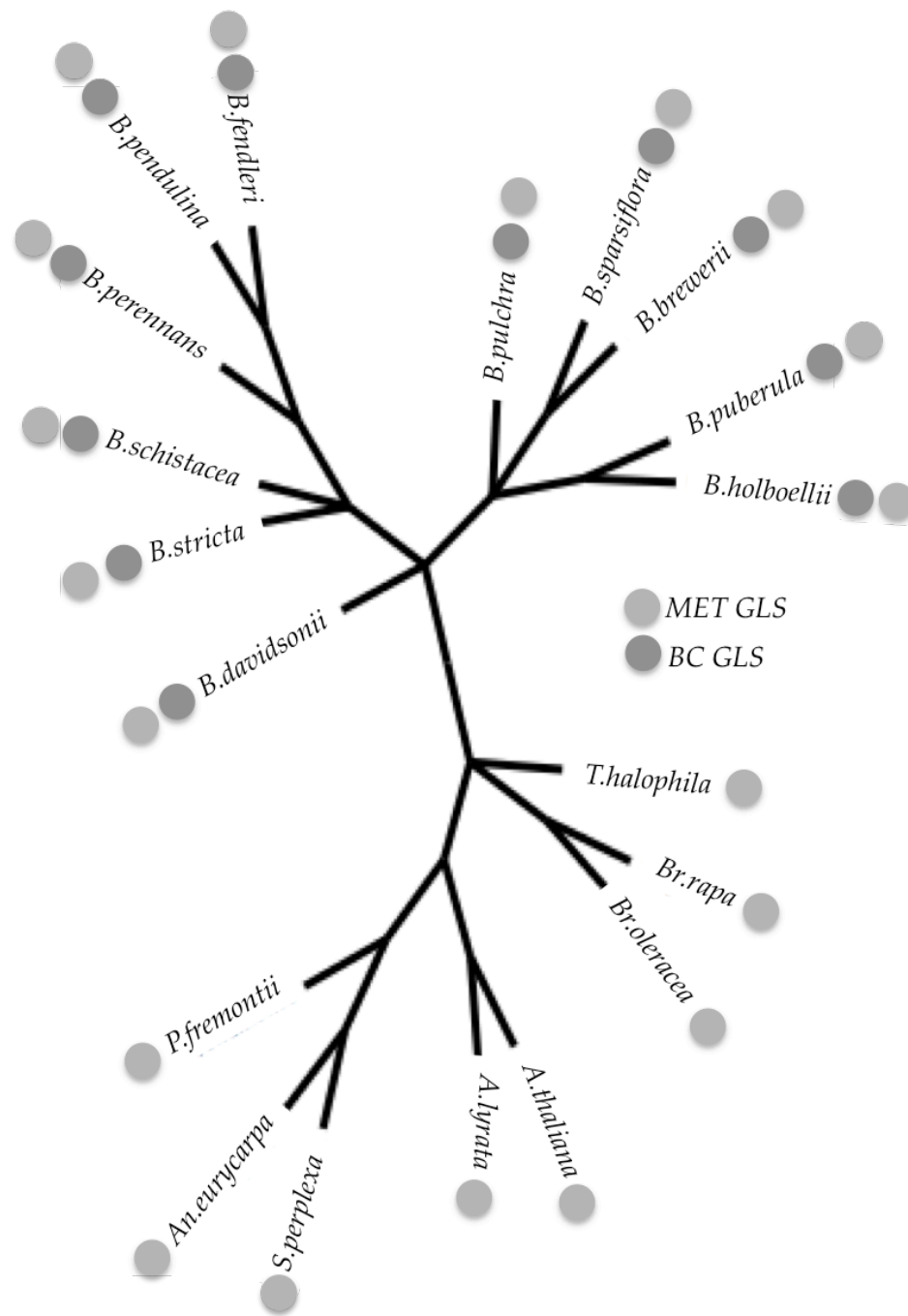


Figure 3: Unrooted species tree of relevant Brassicaceae. Dark grey circles indicate the production of branched-chain derived GLS (IME and 1MP) and light grey circles indicate methionine derived GLS.

2. Evolution of flux control in the glucosinolate pathway in *Arabidopsis thaliana*

Several recent investigations have documented correlations between network properties of genes and selective constraint or rates of substitution, suggesting that network characteristics of enzymes influence their rates of evolution (Flowers *et al.*, 2007, Zera, 2011, Cork & Purugganan, 2004, Slotte *et al.*, 2011, Eanes, 2011). Such properties include the number of connected enzymes (Pfeiffer *et al.*, 2005), enzyme expression (Yang & Gaut, 2011) or flux control (Wright & Rausher, 2010, Flowers *et al.*, 2007). While some of these correlations likely represent differences in the neutral substitution rate due to differences in constraint (Rausher *et al.*, 2008), others represent differences among pathway genes in rates of adaptive substitution (Flowers *et al.*, 2007). If natural selection preferentially acts on certain genes or pathway positions, this may result in repeated or parallel evolution, and contribute to extensive heterogeneity of evolutionary rates among enzymes. Recent studies have considered sequence signatures of selection on enzymes in their network context as a way of understanding what forces influence their evolutionary fate (Gaut *et al.*, 2011) and several patterns have emerged.

One pattern is that the rate of evolution is often correlated with enzyme position in a metabolic pathway. Several studies have found, for example, that genes at the beginning of a pathway are under greater selective constraint, as reflected in the dN/dS ratios than coding for downstream genes (Livingstone & Anderson, 2009, Lu & Rausher,

2003, Ma et al., 2010, Rausher et al., 2008, Rausher et al., 1999) (Ramsay et al., 2009). Differences in pleiotropy among genes at different pathway positions might cause this pattern, as genes in the beginning of a pathway may influence a larger number of downstream products than downstream genes (Ramsay et al., 2009, Rausher et al., 1999). However, such differential pleiotropy has not been demonstrated for any pathway. Moreover, Wright and Rausher (2010) provide an alternative theoretical explanation for this pattern: in linear metabolic pathways, genes coding for upstream enzymes tend to evolve the largest control over flux, which means that slightly deleterious substitutions are more likely to occur in genes coding for downstream enzymes. Their model also predicts that in linear pathways, adaptive substitutions will tend to be concentrated in upstream enzymes because (1) their greater control over flux means that on average mutations in these genes will experience stronger selection; and (2) the probability an advantageous mutation will be fixed is proportional to its selection coefficient. The model thus provides theoretical support for the oft-expressed expectation that adaptive substitutions will be concentrated in enzymes that exert the most control over flux (Eanes, 1999, Hartl *et al.*, 1986, Watt & Dean, 2000).

Despite these expectations, we are unaware of any investigations that have tested them by both estimating the magnitude of flux control for enzymes in a metabolic pathway and also assessing patterns of substitution in the genes coding for those enzymes. In an attempt to bridge this gap between theory and evidence, we have

conducted an explicit test of whether these expectations are met by the glucosinolate pathway in *Arabidopsis*.

Due to the ease of quantifying total pathway outputs, glucosinolate (GLS) production in *Arabidopsis thaliana* provides an excellent pathway for addressing the relationship between flux control and patterns of substitution.

The biosynthesis of methionine (MET) and tryptophan (TRP) -derived GLSs occurs through a series of reactions, with the initial reactions occurring in the chloroplast, and subsequent reactions in the cytosol. In the chloroplast reactions methionine undergoes side-chain elongation. These products are then transferred to the cytosol, where they undergo a series of reactions (Figure 1). Reactions occurring in the cytosol will be examined as the “core” glucosinolate pathway (Figure 1 adapted from Sonderby *et al.*, 2010).

Two features of this core pathway are relevant. First, several steps are catalyzed by two different enzymes. These pairs have resulted from gene duplication and are typically co-expressed. The first pair, CYP79F1 and CYP79F2 has slightly different substrate specificities, with CYP79F1 using both short- and long-chain substrates, while CYP79F2 tends to use only long-chain substrates. The second important feature is that the pathway leading to the synthesis of glucosinolates from methionine shares one enzyme, SUR1, with a parallel pathway that synthesizes glucosinolates derived from tryptophan (Figure 1). Although each pathway is linear, this shared enzyme could

provide some cross talk between them. In particular, if SUR1 is saturated or nearly saturated *in vivo*, reduction in flux down one pathway could increase flux down the other pathway. This effect is not expected, however, if SUR1 is relatively unsaturated, since there would be little competition among the precursors for access to that enzyme. Thus, a test of whether enzymes with predominant flux control in one pathway also exert control over flux down the other pathway provides information about SUR1 saturation and cross talk.

In this analysis of *A. thaliana*, we asked whether the enzymes in the core glucosinolate pathway show differential flux control resulting in heterogeneous substitution patterns. In particular, we address the following questions: (1) Do upstream enzymes exert the majority of control over flux? (2) Are adaptive substitutions concentrated in enzymes that exert strong flux control? (3) Do upstream enzymes exhibit the greatest selective constraint?

2.1 Results and Discussion

To confirm that enzyme expression levels are reduced in heterozygotes and to estimate the relative expression in heterozygotes (C_i), we used quantitative real-time PCR (qPCR) to estimate expression levels for heterozygotes of knockdown lines with insertions in the promoter or exon for each gene, as well as for wild-type homozygotes. Two of the enzymes, GSTF11 and GSTU20, do not exhibit significant reduction in expression levels in heterozygotes (Figure 4, Table 2; unrelativized expression levels), as

judged by contrast statements in ANOVA. For the remainder of the enzymes, the relative expression ratio (R_i) was significantly less than 1 (Figure 4, Table 2), indicating that having just one functional copy of the gene results in substantially reduced expression. For five of these enzymes, the R_i was not distinguishable from the expected value of 0.5, as judged by confidence intervals overlapping 0.5 (Table 2). However, for SUR1 and SOT18, R_i was significantly less than 0.5, indicating that expression in heterozygotes was reduced by more than 50% compared to wild-type individuals.

Table 2: Relative Expression Ratio, R_i , and 95% CI. Bolded values were significantly different than 1.

Gene	R_i	95% CI
CYP79F1	0.3664	(0.2420,0.5550)
CYP83A1	0.4894	(0.3274,0.7315)
GSTF11-1	0.7583	(0.4295,1.3387)
GSTF11-2	1.0826	(0.6806,1.7219)
GSTU20	0.8281	(0.5468,1.2542)
SUR1	0.2698	(0.1843,0.3950)
UGT74B1	0.4011	(0.2684,0.5995)
UGT74C1	0.4667	(0.3082,0.7069)
SOT17	0.4041	(0.2618,0.6238)
SOT18	0.2535	(0.1696,0.3789)

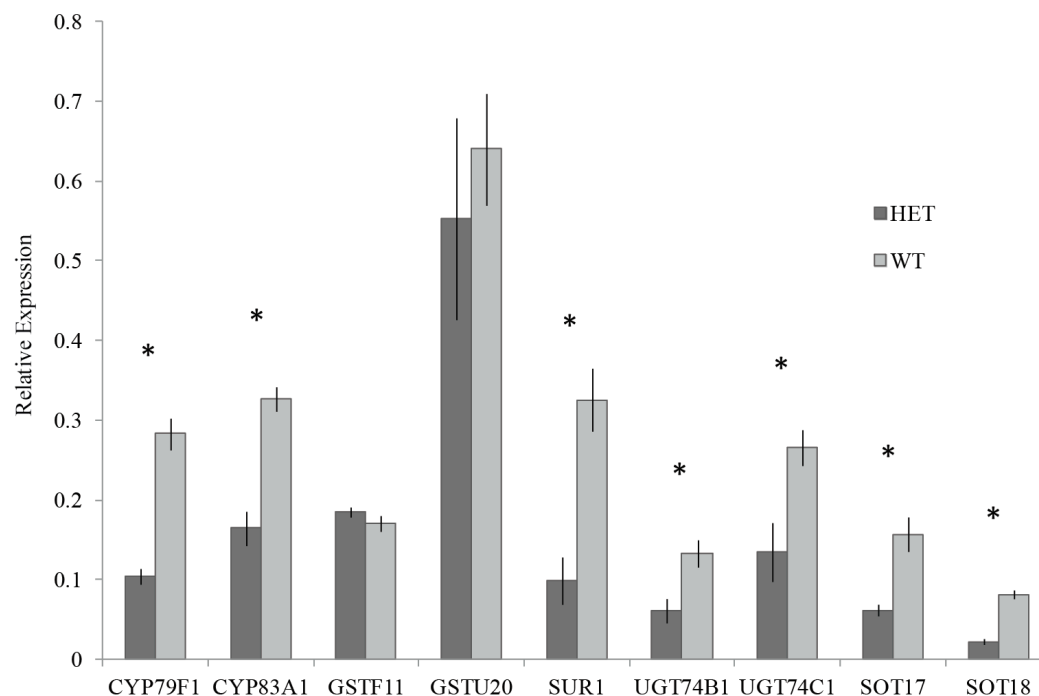


Figure 4: qRT PCR comparison of heterozygous (HET) to wild type (WT) genotypes of the different insertion lines. UBQ10 was used as the reference gene to normalize the glucosinolate gene (abbreviations on horizontal axis) of interest in each RNA extraction. All comparison between the expression of HET (dark grey bars) and WT (light grey bars) are significantly different except GSTF11 and GSTU20. The vertical hashes indicate standard error bars. Asterisks above each pair of bars show relative expression ratio of less than one, as judged by appropriate contrasts in ANOVA.

2.1.1 Estimation of flux control in the glucosinolate pathway

To determine whether differences in enzyme concentrations between heterozygotes and wild-type individuals, as reflected in differences in expression level, affect flux, we compared glucosinolate production in 3-week-old rosette leaves. We first compared wild-type lines. ANOVA indicated that the lines do not differ significantly in glucosinolate production for any of the aliphatic or indolic glucosinolates (Table 13),

indicating no detectable effect of genetic background on glucosinolate production. For subsequent comparisons between HET and WT individuals, we therefore pooled WT individuals from the three lines.

We first examined flux control through the aliphatic pathway. We did not compare HET and WT glucosinolate concentrations for enzymes GSTF11 or GSTU20 because they showed no evidence that expression levels were reduced in heterozygotes. Of the remaining enzymes, only the first enzyme in the pathway, CYP79F1, exhibited a significant reduction in glucosinolate production in heterozygotes as judged by t-test after correction with sequential Bonferroni (Table 14). This reduction was confined to the two short-chain aliphatic glucosinolates, 3MSOP and 4MSOB, and remained significant even after a Bonferroni correction. Estimated flux control coefficients were high for these two compounds ($\lambda = 0.8394$, CI = (0.4873, 1.5947), and $\lambda = 0.4016$, CI = (0.1573, 0.9556) for the two compounds respectively). In agreement with the statistical analysis, all other enzymes exhibited low (all $\lambda < 0.27$) and non-significant flux control coefficients for all four glucosinolates (Table 3). It thus appears that for at least the short-chain aliphatic glucosinolates 3MSOP and 4MSOB, flux control is primarily vested in the first enzyme of the pathway. These results thus support the prediction (Wright & Rausher, 2010) that flux control should evolve to be vested mainly in the first enzyme of a pathway.

Table 3: Estimated Flux Control Coefficients Calculated Using All Three Lines for WT Individuals and 95% Bootstrap CIs After 1,000 Bootstrapping Runs

Gene	GLS			
	3MSOP	4MSOB	5MSOP	6MSOH
CYP79F1	0.8394 (0.4873,1.5947)	0.4016 (0.1573,0.9556)	-0.0238 (-0.1933,0.2934)	-0.1124 (-0.2486,0.0486)
CYP83A1	-0.0411 (-0.2265,0.1172)	-0.0808 (-0.2410,0.0536)	-0.1667 (-0.4194,0.0647)	-0.3315 (-0.6369,-0.1405)
SUR1	0.0673 (-0.0182,0.2436)	0.0726 (-0.0161,0.2639)	-0.0560 (-0.3199,0.6351)	0.2607 (-0.0114,1.5369)
UGT74B1	0.0023 (-0.1781,0.2325)	0.0002 (-0.1829,0.1870)	0.2042 (-0.1094,1.3732)	-0.0604 (-0.4579,0.4414)
UGT74C1	0.0061 (-0.1741,0.2416)	0.0246 (-0.1574,0.2832)	-0.3603 (-1.0650,0.1197)	0.2183 (-0.2090,1.4085)
SOT17	0.0352 (-0.0526,0.1522)	-0.0413 (-0.1529,0.0793)	0.0617 (-0.2282,0.6739)	-0.1483 (-0.4511,0.3360)
SOT18	-0.0148 (-0.0733,0.0542)	-0.0291 (-0.0832,0.0265)	-0.0600 (-0.1442,0.0196)	-0.1194 (-0.2086,-0.0532)

An apparent exception to this principle is the absence of flux control over the production of the long-chain aliphatic glucosinolates 5MSOP and 6MSOH, either in the first enzyme or any of the other enzymes. One possible explanation for this is that flux control associated with the production of these compounds is vested in CYP79F2, an enzyme that was not examined in this investigation because the same kind of loss-of-function mutants are not available. Previous investigations suggest that activity of the CYP79F1 and CYP79F2 enzymes is redundant, but only for long-chain precursors. In a tandem deletion study of *CYP79F1* and *CYP79F2* (Tantikanjana *et al.*, 2004) the authors found that both CYP79F1 and CYP79F2 produce long-chain aliphatic GLS but only CYP79F1 produced short-chain aliphatic GLS. Likewise, biochemical studies also

indicate that the synthesis of long-chain aliphatic GLS by CYP79F1/CYP79F2 is redundant, since both enzymes catalyze tri-, tetra- penta- and hexahomomethionine, but only CYP79F1 can catalyze short-chain aliphatic amino acids to their corresponding aldoximes (Chen *et al.*, 2003) (Hansen & Wagner, 2001). Since both CYP79F1 and CYP79F2 make long-chain aliphatic compounds, it is unclear whether changing CYP79F2 should also change the amount of long-chain glucosinolates produced. Given that both enzymes make long-chain aliphatic compounds, it is possible that the majority of flux control over these compounds is vested in CYP79F2 rather than in CYP79F1. However, because CYP79F2 cannot metabolize short-chain precursors, it is not likely to exert significant flux control over their production. If this hypothesis is true, the separation of flux control for short- vs. long-chain aliphatic glucosinolates between two different enzymes could allow for evolutionary flexibility in the relative concentrations of these two classes of glucosinolates.

A pathway parallel to the aliphatic pathway produces indolic glucosinolates and the two are largely independent (Figure 1). However, the two pathways share one enzyme, SUR1. This sharing creates the possibility that altering the concentrations of enzymes in the aliphatic pathway could affect the flux through the indolic pathway if there is strong competition among aliphatic and indolic precursors for access to SUR1. In this situation, which reflects near saturation of SUR1, control coefficients of aliphatic-pathway enzymes for the production of indolic products would be negative: a reduction

in enzyme concentration would reduce flux through the aliphatic pathway, reducing competition of substrates for SUR1, and increasing indolic flux.

Given that only CYP79F1 exerts detectable control over aliphatic flux, we would *a priori* expect only this enzyme to exhibit this type of negative flux coefficient. In fact, neither it, nor any of the other aliphatic enzymes exhibit detectable control over indolic glucosinolate production: indolic glucosinolate concentrations do not differ significantly between HET and WT for any of the aliphatic enzymes (Table 3). This result indicates that despite the potential for interaction between the two pathways, there is little effect of aliphatic flux level on indolic flux level, implying that SUR1 is far from saturation *in vivo*.

2.1.2 Relative substitution rates

We first examine the prediction that adaptive substitutions are expected to be more frequent in upstream genes, particularly in the first enzyme of the pathway. Repeated adaptive substitutions in a gene can be detected with a MacDonal-Kreitman test (McDonald & Kreitman, 1991). We applied this test to each gene, and corrected significance levels for multiple comparisons. None of the genes exhibited a significant excess of nonsynonymous substitutions, which would be reflected in a DoS greater than 0 (Table 4). In fact, DoS was positive only for GSTF11, and this was not significant. However, both CYP79F1 and SUR1 showed statistically significant negative DoS values based on a permutation test. We used Fisher's exact test to perform MacDonal-Kreitman tests on 10 glucosinolate biosynthetic genes, with a permutation procedure to

account for multiple tests on these loci. These results found $P = 0.0029$ for *CYP79F1*, and $P = 0.0001$ for *SUR1*, showing that both loci show significantly negative DoS.

Table 4: McDonald-Kreitman test of polymorphism in *A.thaliana* compared with divergence from *A.lyrata*. ^aSignificance values based on permutation test.

Gene	P _N	P _S	D _N	D _S	Dos	Permutation test ^a
CYP79F1	13	5	20	42	-4.46	0.0029
CYP79F2	8	8	39	50	-0.28	n.s.
CYP83A1	2	14	6	26	0.62	n.s.
GSTF11	6	5	12	13	-0.3	n.s.
SUR1	14	13	4	31	-7.35	0.0001
UGT74B1	18	25	16	37	-0.67	n.s.
UGT74C1	18	25	19	34	-0.29	n.s.
SOT17	6	8	19	32	-0.26	n.s.
SOT18	8	3	86	80	-1.48	n.s.

One interpretation of these results is that balancing selection has operated at both of these loci. If this were true, we would expect to find positive Tajima's D values. Only *UGT74B1* has a slightly positive D (Table 15), which is not significant. Indeed, none of the genes show Tajima's D values in the top 5% of the genome-wide distribution. However, two alternative interpretations are also consistent with the data. The first is that the recent population expansion (Beck *et al.*, 2008, Sharbel *et al.*, 2000) in *A. thaliana* has allowed the accumulation of deleterious alleles and thus has led to negative values of DoS at most loci in the genome and negative Tajima's D values. The observation that nine of the ten genes exhibit negative DoS, a pattern that is significantly different from chance based on a binomial test ($P < 0.01$), is consistent with this interpretation. In this situation, the significant negative DoS for SUR1 might simply reflect the most extreme

example of this stochastic process. Alternatively, the combination of negative Tajima's D values and positive Fay and Wu's H (Table 15) is consistent with an old population bottleneck that is regaining neutral variation (Haddrill *et al.*, 2005). These measures are influenced by demographic factors and the genes do not deviate from the genome-wide distribution (Table 15).

By contrast, at *CYP79F1*, the π_N/π_S ratio is substantially and significantly > 1 (bootstrap confidence 99% interval: 1.15 – 8.02, 1,000 replicates). While a π_N/π_S ratio greater than one can reflect repeated adaptive substitution (Kryazhimskiy & Plotkin, 2008), our failure to detect any adaptive substitutions with an M-K test strongly suggests that this locus is subject to strong balancing selection. At *SUR1*, however, the π_N/π_S ratio is substantially less than one and is comparable to that of the other pathway genes, providing no indication that this gene has historically been subject to balancing selection. Under this interpretation, which we favor, *CYP79F1* is the only gene subject to selection. The observation that this gene codes for the first enzyme in the pathway, which is also the only enzyme with demonstrable control over flux, is consistent with expectations.

Table 5: Synonymous and nonsynonymous within species variation of *A.thaliana* and between species variation of *A.thaliana* compared to *A.lyrata*. *A.lyrata* has no ortholog of GSTU20.

Gene	Site Type	Polymorphism		Divergence	
		π	π_N/π_S	Proportion site type	dN/dS
79F1	SYN	0.0007	2.4054	0.1075	0.1709
	NONSYN	0.0018		0.0184	
79F2	SYN	0.0043	0.2189	0.1327	0.2699
	NONSYN	0.001		0.0358	
83A1	SYN	0.0158	0.0095	0.1358	0.02
	NONSYN	0.0002		0.0027	
GSTF11	SYN	0.0036	0.3609	0.0975	0.2668
	NONSYN	0.0013		0.026	
GSTU20	SYN	0.0249	0.083	-	n.a.
	NONSYN	0.0021		-	
SUR1	SYN	0.0086	0.2733	0.095	0.0651
	NONSYN	0.0024		0.0062	
B1	SYN	0.0158	0.2212	0.1492	0.1215
	NONSYN	0.0035		0.0181	
C1	SYN	0.0047	0.2201	0.1332	0.1532
	NONSYN	0.001		0.0204	
SOT17	SYN	0.003	0.1	0.143	0.1701
	NONSYN	0.0003		0.0243	
SOT18	SYN	0.0012	0.9741	0.1298	0.1212
	NONSYN	0.0011		0.0157	

A third possible interpretation is a variant of the preceding one: what we have taken for excess nonsynonymous polymorphism may be attributable to divergence among *A. thaliana* populations. The accessions used for our M-K analysis were collected from widespread geographic localities (Cao *et al.*, 2011). Differences among accessions thus might reflect fixation of different alleles at different locations, *i.e.*, between-population substitution rather than within-population variation. If so, the apparent

excess of nonsynonymous polymorphisms and $\pi_N/\pi_S \gg 1$ at *CYP79F1* could represent repeated episodes of positive selection at these two loci. In that case, the theoretical prediction that adaptive substitutions should be concentrated in the first pathway enzyme would be upheld for *CYP79F1*. Indeed, the site-frequency spectrum of derived nonsynonymous polymorphisms in *CYP79F1* (Olson-Manning *et al.*, 2013) finds some high-frequency polymorphisms. The geographic distributions of these polymorphisms show that they are widespread (Olson-Manning *et al.*, 2013) and might represent between-population substitutions. This is consistent with the F_{st} values we find for the genes in the GLS pathway (Horton *et al.*, 2012) (Table S8) where none deviate from the genome-wide distribution.

Although the last two interpretations seem more compelling to us than the inference that both *CYP79F1* and *SUR1* have experienced balancing selection, we cannot rule out that interpretation. If that interpretation is true, balancing selection at *SUR1* is difficult to account for by the principle that selection preferentially targets enzymes with high flux control. Instead, the elevated nonsynonymous variation would presumably reflect some other phenomenon, for example if *SUR1* displays greater pleiotropy than other genes in the pathway, which has been implicated as a possible enzyme property that correlates with rate of evolution. For example, Ramsay *et al.* (2009) found that the rate of evolution in the plant terpenoid biosynthetic pathway was correlated with inferred levels of pleiotropy (Ramsay *et al.*, 2009). We note that as the only enzyme

involved in the production of both aliphatic and indolic glucosinolates, *SUR1* has the potential to incur greater pleiotropy than the other pathway enzymes.

Given the evidence for positive or balancing selection in the flux-controlling enzyme, we expect that this signal will overwhelm any signature of stronger purifying selection and indeed, our results do not find stronger purifying selection in the flux controlling enzyme *CYP79F1*. The dN/dS ratio is much lower for genes *CYP83A1* and *GSTU20* than it is for *CYP79F1*. The site-frequency spectrum of *CYP79F1* has many high-frequency derived amino acid substitutions, which is not expected of an enzyme under strong purifying selection.

In summary, the data for evaluating the relationship between flux control and patterns of selection are equivocal but suggestive. The observation of elevated nonsynonymous polymorphism indicates that the first pathway enzyme with the greatest control over flux is subject to selection, as is expected. This may be due to several adaptive substitutions in portions of the species range, since there is no clear sequence signature of balancing selection. However, it is not clear whether this pattern reflects balancing selection or repeated adaptive substitution. With the exception of *SUR1*, the other pathway genes exhibit no detectable flux control and no evidence of either balancing or positive selection, as expected. The expectation of lack of selection on *SUR1* is consistent with the data, but we cannot completely rule out the possibility that it is also subject to balancing selection. If so, this is not explained by the expected

relationship between flux and selection, but possibly by greater pleiotropy because of its operation in both the aliphatic and indolic pathways.

2.2 Conclusions

While there is growing evidence from signatures of selection that flux control may be unevenly distributed and focused at the beginning of pathways, we do not know how general our results may be. For one, it is conceivable that flux control could change quickly on evolutionary time scales, or even during development. The simulations of Wright and Rausher (2010) actually showed that flux control can evolve to be centered in different enzymes, although the probability is high that flux control will be vested in the first enzyme. However, studies in anthocyanin (Lu & Rausher, 2003, Rausher et al., 1999) and plant terpenoid (Ramsay et al., 2009) pathways suggest that patterns of sequence rate variation have persisted over very long evolutionary timescales (since the divergence monocots and dicots). Flux control may play some part in shaping these patterns of sequence rate variation, but this possibility remains to be demonstrated.

The results of this study are consistent with theoretical predictions of flux control and sequence rate variation, in that flux control is unevenly distributed and that the beginning of a pathway often shows majority flux control. Although we do not know the range of environmental and physiological conditions to which these results may be generalized, studies of pathway rate variation in plants and animals suggest that these

patterns may be common. We test the theory that flux control changes due to environment in the following chapter.

2.3 *Materials and Methods*

2.3.1 Insertion lines

To determine whether differences in the pattern of selection on different pathway enzymes is correlated with the magnitude of flux control, we approximated flux control of each step in the pathway by perturbing the amount of each enzyme using *Agrobacterium* TDNA insertion lines (Sussman *et al.*, 2000, Alonso *et al.*, 2003) (Alonso *et al.*, 2003, Sussman *et al.*, 2000). These TDNA insertions disrupt the gene of interest and in heterozygous form may substantially decrease the amount of mRNA produced, and thus the total activity of that enzyme available in the cell. *Arabidopsis thaliana* insertion lines Wisconsin (Sussman *et al.*, 2000) background Wassilewskija (WS), SALK Institute Genomic Analysis Laboratory (Alonso *et al.*, 2003) Columbia-0 (Col-0, CS60000), Syngenta *Arabidopsis* Insertion Library (SAIL) were collected for as many genes in the aliphatic GLS pathway as were available (Table S1). Each of these lines contained an insertion causing loss-of-function (LOF) in one aliphatic GLS pathway gene, either in heterozygous or homozygous form.

2.3.2 Plant growth conditions:

Seeds from each line were grown for one generation to determine whether they were heterozygous or homozygous for the LOF allele. For each line, approximately

twenty seeds from heterozygous maternal individuals were placed on soil in a randomized complete block design in a 24-cell flat. Seeds were allowed to imbibe, and then were stratified for 3 days at 4° C to overcome dormancy. Plants were maintained under long day conditions (16 hours) at 18°C for three weeks, when tissue was harvested. One true rosette leaf was collected over two consecutive mornings for each of the following analyses: insertion genotyping, mRNA analysis, and glucosinolate (GLS) quantification. The tissue for RNA analysis was flash frozen, and the tissue for GLS quantification was stored in 2 ml 70% methanol for glucosinolate analysis.

2.3.3 Genotyping

Frozen tissue was ground with ball bearings on liquid nitrogen in a Geno/Grinder (SPEX, CentriPrep 2000). DNA was extracted with a modified CTAB protocol (Rogers & Bendich, 1988) and resuspended in TE buffer. Primers were designed using the SALK SIGnAL iSect primer design tool (<http://signal.salk.edu/tdnaprimers.2.html>) (Table S1) and genotyped according to (Alonso et al., 2003) with Go Taq (Promega) Taq. Genotypes were scored on a 0.7% agarose gel.

2.3.4 Quantitative real-time PCR:

To quantify the relative amounts of mRNA transcripts, duplicate samples of three-week old rosette leaves were collected and flash frozen on liquid nitrogen from heterozygous (HET) or wild-type (WT) plants. Total RNA was extracted from with the

SV RNA kit (Sigma). cDNA synthesis was performed with the DynamoTMcDNA synthesis kit (Finnzymes). If possible, primers were designed to span the intron of the gene of interest (MWG Operon) (Table S3). For GSTF11, two different primer sets were designed that produced the same result. Glucosinolate genes were normalized to the transcript levels of the reference gene *UBQ10* (At4g05320) (Czechowski *et al.*, 2005). Duplicate qRT PCR reactions were performed for each primer pair. Quantitative reverse transcriptase PCR was performed with the DynamoTMSybr® Green qPCR kit (Finnzymes). Data were analyzed on the Mastercycler ep realplex 2 (Eppendorf). The reactions were carried out at 95° C for 2 min, and 40 cycles of 95° C for 15 sec, 56° C for 15 sec and 68° C for 20 sec.

2.3.5 Analysis of GLS Concentration:

Leaf tissue was first weighed, and was then leached in 2 ml of 70% methanol for 3 weeks at 4° C and one week at room temperature. Sinigrin (Sigma) was added to each sample to 1ug/ml to each well to serve as an internal reference. Then the entire 2 ml leaching volume was added to a 96-well plate containing equilibrated DEAE Sephadex and cleaned (Mikkelsen *et al.*, 2009). We cleaved the glucosinolates into desulfo-glucosinolates with 30 ug sulphatase. The desulfo-glucosinolates were run on high-pressure liquid chromatography (Kliebenstein *et al.*, 2001) on an Agilent 1100 high-pressure liquid chromatography (HPLC) with 96-cell autoloader. Separation of glucosinolates was carried out on a C-18 column (Zorbax Eclipse XDB C18, 4.6 x 150 mm

and 5 micron) and peaks were called manually based on retention time and UV absorption spectra at $A_{229\text{nm}}$ (Windsor *et al.*, 2005). Seven glucosinolates were quantified: four from the aliphatic pathway and three from the indolic pathway (Table S2). The area of each peak was calculated and normalized by the weight of the tissue collected; the area of the sinigrin peak and molar concentrations were calculated given the pre-determined calibration curve from pure desulfo-glucosinolates (Brown *et al.*, 2003). Approximately 30 replicates were run for each WT line and 10 for each heterozygote type.

2.3.6 Sequence analysis

To investigate the pattern of natural selection on glucosinolate pathway genes, we used the polymorphism data of *A. thaliana* with *A. lyrata* as outgroup. Eighty-one *A. thaliana* genomes (including the Col-0 reference genome) were downloaded from the MPICao2010 subset (Cao *et al.*, 2011) of the 1001 genome project (<http://1001genomes.org/index.html>), and coding sequences (CDS) of *A. lyrata* genes and their orthology information with *A. thaliana* were downloaded from a recently published dataset (Hu *et al.*, 2010). Based on the annotation from TAIR10 release of The Arabidopsis Information Resource (Lamesch *et al.*, 2011), we extracted CDS from both species (Cao *et al.*, 2011). Alignment was performed by the codon model in PRANK (Goldman *et al.*, 2000, Loytynoja & Goldman, 2005), and GSTU20 (AT1G78370) was

excluded from the following interspecific analyses due to the lack of orthologs in *A. lyrata*.

To examine the pattern of selective constraint or positive selection within each enzyme of the aliphatic GLS pathway in *A.thaliana*, we calculated the average pairwise difference per site (π) and the ratio of synonymous and nonsynonymous π (π_N/π_S) in DnaSP (Rozas *et al.*, 2003). Significance of deviation of this ratio from 1 was tested by bootstrapping (1,000 replicates) using a program written in APL by one of the authors (MDR).

We used DnaSAM (Eckert *et al.*, 2010) to calculate additional population genetic statistics based on the site frequency spectrum (Tajima's D and normalized Fay & Wu's H). The statistics of genes in the glucosinolate pathway were then compared to the distribution of all *A. thaliana* genes with known *A. lyrata* orthologs. Additionally, we plotted the site-frequency spectrum of the derived amino acid polymorphisms in all the genes in the glucosinolate pathway. The derived amino acid was determined based on either the ortholog in *A.lyrata*, or, in the case of SOT18 and GSTU20 with no *A.lyrata* ortholog, the ancestral reconstruction software ANCESCON (Cai *et al.*, 2004). To explore the geographic distribution of the amino acid polymorphisms we plotted the geographic distribution of ancestral and derived polymorphisms for CYP79F1 and SUR1 with the R package chplot (Vidmar & Pohar, 2005). Finally, we searched the worldwide *A.thaliana*

Regmap panel (Horton et al., 2012) to search for departures from genome-wide values of F_{st} in the glucosinolate pathway.

To address constraints on divergence between-species (*A.thaliana* and *A.lyrata*) we calculated the dN/dS ratio in DnaSP. Additionally, a Perl script (Holloway *et al.*, 2007) was used to perform MK tests (McDonald & Kreitman, 1991) for possible adaptive substitution. As an index of whether deviations from neutrality are due to excess adaptive substitutions or an excess of non-synonymous variation, we calculated the statistic DoS (direction of selection), a variant of the McDonald Kreitman test. This statistic is zero under neutrality (Stoletzki & Eyre-Walker, 2011). Positive DoS indicates an excess of nonsynonymous substitutions between species, or a deficiency of nonsynonymous polymorphisms within species. Negative DoS indicates an excess of nonsynonymous polymorphism within species, or a deficit of nonsynonymous divergence.

We used a permutation procedure to control levels of statistical significance for multiple tests. With 11 loci, here we focused on two loci that were potentially significant based on univariate tests. Under the null hypothesis that synonymous/nonsynonymous status is independent of whether variation is polymorphic within species or fixed between species, we computed the null distribution for Fisher's exact test (FET) for the most extreme locus, and the second most extreme locus, in a sample of 11 N loci. We permuted synonymous/nonsynonymous status across SNPs, and calculated FET for each

locus in each permutation. The most extreme and second most extreme FET values were identified among these loci, and saved to a null distribution based on 50,000 permutations. Finally, the two loci (*CYP79F1* and *SUR1*) with most extreme actual FET values were compared to these statistical null distributions. Calculations employed a Python program written by one of the authors (TMO).

2.3.7 Analysis of flux control

The magnitude of flux control exerted by an enzyme is typically assessed in one of three ways: (1) by perturbing enzyme concentration or activity; (2) by calculation from elasticity coefficients; or (3) by quantifying transient metabolite concentrations (Delgado, 1992). We present here a novel approach that is based on comparing the glucosinolate concentration of individuals heterozygous for a loss-of-function mutant in a gene in the glucosinolate pathway with that of wild-type individuals. A reduction in glucosinolate concentration in heterozygous individuals implies that reducing enzyme concentration reduces glucosinolate production and thus that the enzyme has substantial control over flux. The magnitude of the flux control coefficient for enzyme i , λ_i (sensitivity coefficient of Kacser & Burns, 1973) was estimated from the equation

$$\lambda_i = [C_i(1 - (F_R)_i)] / [(F_R)_i(1 - C_i)]$$

where C_i is the relative concentration of the enzyme i in heterozygotes (*i.e.* concentration in heterozygotes divided by concentration in wild type) and $(F_R)_i$ is the relative flux of enzyme i in heterozygotes (flux in heterozygotes divided by flux in wild type) (see Appendix for derivation).

As an approximation, we estimated C_i by the relative expression, R_i , the ratio of heterozygote expression level of enzyme i divided by wild-type expression level. Although we did not measure enzyme concentrations directly, it is expected that concentrations in heterozygotes for a loss-of-function allele will be approximately half that in wild-type homozygotes unless there is feedback regulation of transcription. In some cases mRNA expression level may be an imperfect indicator of enzyme activity due to possible post-translational regulation, which might reduce the correlation between mRNA and protein concentration (Vogel & Marcotte, 2012). However, a recent large scale analysis in humans showed that the correlation between mRNA and protein abundance is strong (Schwanhausser *et al.*, 2011). Additionally, our study compares the relative expression between isogenic WT and HET lines, so differential patterns of post-translational modification between WT and HET lines should be minimal. $(F_R)_i$ was estimated indirectly by the relative glucosinolate level in heterozygotes (concentration in heterozygotes divided by concentration in wild type). This approach assumes that final glucosinolate production is proportional to flux, as would be the case if glucosinolate

production occurred over a fixed period of time. Confidence intervals for the estimates of λ_i were calculated by bootstrapping (1,000 replicates).

Statistical Analysis

Comparison of expression levels in heterozygotes and corresponding wild-type homozygotes was performed using appropriate contrast statements in an analysis of variance. Expression levels were log-transformed before analysis. Relative expression levels, R_i , were calculated by first estimating the difference in log (expression) between heterozygotes and wild type, D_i then calculating $R_i = e^{D_i}$. Confidence intervals for R_i were calculated by first calculating the confidence intervals for D_i from their standard errors and then transforming by exponentiation.

Comparison of glucosinolate production between wild-type and heterozygous plants was conducted using multivariate analysis of variance (MANOVA). Dependent variables were concentrations of different glucosinolates. When the multivariate effect of treatment (WT vs. HET) was significant, as judged by Wilk's λ statistic (Timm, 1975) univariate analyses were performed to determine which glucosinolates were affected by treatment. Two types of analysis were performed. In one analysis, we pooled all WT individuals of different lines and compared these individuals to HET individuals. This pooling was justified because preliminary analyses did not show a significant line effect for WT individuals at any of the genes (Table S4). However, we also performed a second analysis for each gene in which HET individuals were compared to only WT individuals

from the same line. Because the results of the two analyses were fully concordant, we report only the results of the analyses with WT lines pooled. Analyses were performed using PROC GLM in the SAS statistical software package (SAS 9.3, Cary, NC).

3. Flux control under different environmental conditions

The productivity of a metabolic pathway is largely influenced by the enzyme(s) with the highest flux control. However, the environment in which an individual finds itself dictates what resources are available, other metabolic or regulatory pathways that may be activated and subsequently the flux through a pathway. Given such changes, it is possible that under stressful or limiting conditions, flux control may change among enzymes in a pathway. If flux control is stable, the enzymes that are exploited by natural selection may be predictable. However, if majority control changes among enzymes depending on the conditions, there would be a variety of ways in which selection could to adjust pathway output, and where selection acts will be unpredictable.

The productivity of the aliphatic glucosinolate pathway in *A. thaliana* is determined by the activity of the enzyme with the most flux control, CYP79F1, as discussed in the previous chapter. However, different environmental conditions (nutrient availability, water stress, and damage) are known to change the amount and proportion of the different aliphatic and indolic glucosinolates produced (Bodnaryk, 1994, Koritsas *et al.*, 1991). This raises the question of how these changes in glucosinolates occur. For example, majority flux control switching among enzymes under different conditions might explain the proportional increase or decrease of different types of glucosinolates.

Previous studies of evolutionary rate variation in biochemical pathways implicitly assume that flux control remains relatively constant over long evolutionary times. The pattern that upstream enzymes experience stronger purifying selection in the glucosinolate (Olson-Manning et al., 2013), anthocyanin (Lu & Rausher, 2003, Rausher et al., 1999) and plant terpenoid (Ramsay et al., 2009) pathways is consistent with the idea that mutations in these enzymes have the largest effect on phenotype, and are selected most efficiently. In the previous chapter and (Olson-Manning et al., 2013) we showed that the first step in the aliphatic glucosinolate pathway has majority flux control under benign greenhouse conditions, and that it is the only gene to exhibit a convincing evidence of positive selection (with a significant excess in nonsynonymous polymorphism based on the McDonald Kreitman test and a π_N/π_S ratio much greater than 1). However, even weak selection can produce strong patterns in sequences (Nielsen, 2005).

To test whether environmental conditions alter the distribution of flux control among enzymes in a biochemical pathway, we studied the effects of environmental treatments on the aliphatic glucosinolate pathway in *A. thaliana*. Nutrient limitation and water limitation were studied because of their importance in organism growth and glucosinolate concentration (Bouchereau *et al.*, 1996, Mailer & Cornish, 1987), while methyl jasmonate and mechanical wounding treatments have been shown to change the proportion of aliphatic and indolic glucosinolates in some *Brassica* species (Bodnaryk,

1994, Koritsas et al., 1991). Jasmonate and methyl jasmonate are phytohormones that influence a variety of cellular processes including senescence, growth inhibition and the induction of secondary metabolism (Pauwels *et al.*, 2009). They have been shown to increase the amount of indolic glucosinolates relative to aliphatic GLS (Bodnaryk, 1994). Mechanical crushing has similar effects to jasmonate, but also increases the expression of both aliphatic and indolic compounds (Koritsas et al., 1991). We hypothesized that subjecting plants to these treatments should change the quantity or quality of the glucosinolate profiles produced and might reveal possible changes in pathway flux control.

We chose three genes in the glucosinolate pathways and measured their flux control. *Cyp79f1* encodes the enzyme responsible for the first step in the pathway, and was previously shown to have majority flux control (Olson-Manning et al., 2013) under typical greenhouse conditions. Two other genes, *Cyp83a1* and *Sur1*, did not show significant flux control, but have other interesting properties and were also chosen for this study. *Cyp83a1* shows a pattern of selective constraint that is 10x higher than other genes in the pathway; a pattern that would be expected for a gene under strong purifying selection. The SUR1 enzyme functions in both aliphatic and indolic glucosinolate pathways (Figure 1). If the SUR1 enzyme is saturated *in vivo* and shows catalytic preference towards either aliphatic or indolic side-chains, under some

environmental conditions, it could divert flux down one of the two pathways in which it is involved.

In addition, it is unknown whether changes in flux may influence whole-organism phenotypes that determine fitness in the field. For example, damage by herbivores may cause drastic reductions in plant performance, and many herbivores are sensitive to the concentration and type of glucosinolates produced by Brassicaceae (Schranz *et al.*, 2009). Specifically, the herbivore *Trichoplusia ni* is highly sensitive to the isothiocyanate glucosinolate breakdown products of many *A. thaliana* accessions (Lambrix *et al.*, 2001), so this insect may respond to slight changes in glucosinolate concentration, and thus to pathway flux.

Here we ask the following questions: Does flux control shift among enzymes in the aliphatic glucosinolate pathway of *A. thaliana* under different environmental conditions? Is feeding by the generalist herbivore *Trichoplusia ni* affected by changes in flux in the glucosinolate pathway? And finally, do environmental effects on gene expression predict the resulting glucosinolate profiles?

3.1 Results

We measured glucosinolate concentration in *A. thaliana* on plants that were either heterozygous (HET) for gene insertion lines, or homozygous wild type (WT) at these insertion sites. Plants were subjected to four treatments and glucosinolate concentration was measured on 3-week-old rosette leaves. The overall effects of genotype and

environments on glucosinolate concentration were analyzed using MANOVA (Table 6), with genotype (HET vs. WT) and the four environments (low water (W), leaf crushing (C), low nutrient soil (S) and methyl jasmonate (J) treatments) contrasted to the control treatment (standard greenhouse conditions). In the full model, the interaction of genotype by environment and environment by environment were also tested.

Table 6: MANOVA Effect of genotype and environment on glucosinolate concentration

	<i>Cyp79f1</i> p-value	<i>Cyp83a1</i> p-value	<i>Sur1</i> p-value
Geno	1.37*10 ⁻²⁰ ***	0.0237 *	0.0216 *
W	5.14*10 ⁻⁷ ***	0.0008 **	8.08*10 ⁻⁵ ***
C	1.11*10 ⁻⁷ ***	6.82*10 ⁻⁷ ***	4.74*10 ⁻⁵ ***
S	4.90*10 ⁻⁸ ***	3.73*10 ⁻⁵ ***	9.57*10 ⁻⁸ ***
J	3.87*10 ⁻¹¹ ***	7.46*10 ⁻¹⁴ ***	2.79*10 ⁻¹¹ ***
Geno x W	0.9397	0.8735	0.7093
Geno x C	0.3199	0.4382	0.7466
Geno x S	0.8302	0.3913	0.1797
Geno x J	0.8355	0.5039	0.935
W x C	0.1989	0.4219	0.9822
W x S	0.083	0.0376 *	0.1614
W x J	0.3688	0.7558	0.9404
C x S	0.1257	0.297	0.1088
C x J	0.0538	1.22*10 ⁻⁵ ***	0.0002 **
S x J	0.4732	0.5571	0.8802

For all genes tested, MANOVA finds that genotype and each of the four environments are statistically significant. None of the genotype by environment interactions are significant. Several of the environment-by-environment interactions were significant, but the majority of these interactions are not.

We performed univariate tests to examine the effect of genotype on concentration of each glucosinolate compound. Because no significant genotype by environment interactions were detected in the MANOVA, for the univariate analyses we pooled environmental treatments. The effect of genotype is highly significant for *Cyp79f1*, as was found in our previous work (Olson-Manning et al., 2013). The univariate analyses show that *Cyp79f1* has a much greater effect on glucosinolate concentration than either of the other genes (Figure 5, Table 8). 3MSOP and 4MSOB were significantly decreased in the HET compared to the WT (as previously reported, (Olson-Manning et al., 2013)) but 6MSOH, I3M and 1MOI3M were increased. *Cyp83a1* and *Sur1* also have significant genotype effects in the MANOVA, but the changes are of a much lower magnitude than *Cyp79f1*. None of the univariate comparisons were significant for *Cyp83a1*, but *Sur1* had a significant increase in 3MSOP and 4MSOB in the HET genotype.

We calculated flux control coefficients (Table 7) on the pooled environmental treatment as in Appendix 1. We used the relative expression ratios for *Cyp79f1*, *Cyp83a1* and *Sur1* determined in Chapter 2 (Table 2). The ratio of HET to WT expression and confidence interval was calculated by bootstrapping for 1,000 iterations. We find results qualitatively similar to our previous study. CYP79F1 has majority flux control for the compounds 3MSOP ($\lambda=1.1179$, CI=1.0057-1.2474) and 4MSOB ($\lambda=0.7396$, CI=0.5568-1.0183), but not for any other GLS. The other genes have much lower estimated control coefficients ($\lambda<0.47$).

Each of the environmental treatments had a highly significant effect on the quantity and spectrum of glucosinolates produced (Figure 6, Table 9). However, there are no significant genotype by treatment interactions, indicating the response to environmental treatment is homogenous in the HET and WT, and therefore flux control is robust across environmental treatments.

There were no significant differences in the glucosinolate profiles among the different TDNA insert lines, so we pooled the WT lines for our estimates of environmental influences on glucosinolate concentration. Figure 6 and Table 9 show the directions of glucosinolate concentration changes as a result of the four treatments.

The effect of genotype and the four treatments on the amount of leaf area removed was analyzed with ANOVA (Table 10). *Cyp79f1* genotype had a significant effect on amount of leaf area removed, as did soil and methyl jasmonate treatments. While the glucosinolate profiles do not differ between Col (from which the *Cyp79f1* and *Sur1* mutants are derived) vs. *Ws* (*Cyp83a1*), these accessions may differ in other herbivore feeding cues. Therefore in this ANOVA each line is compared only with others from the same line.

To determine if the change in gene expression following methyl jasmonate treatment, we performed a meta-analysis with data available from three experiments on the EMBL Gene Expression Atlas. All of the genes in the aliphatic and indolic pathways

were significantly up-regulated in at least two of the three EMBL available experiments after application except *cyp81f2*, which was significantly down-regulated (Figure 8).

Table 7: Estimated flux control coefficients calculated using all three lines for WT individuals and 95% Cis after 1,000 Bootstrapping Runs.

	CYP79F1	CYP83A1	SUR1
3MSOP	1.1179 (1.0057,1.2474)	0.2052 (0.1474,0.3157)	0.0055 (-0.0135, 0.0311)
4MSOB	0.7396 (0.5568,1.0183)	0.3278 (0.1856,0.6260)	0.0520 (0.0321,0.0745)
5MSOP	0.0397 (0.0101,0.0761)	0.2033 (0.1675,0.2409)	0.0673 (0.0474,0.0884)
6MSOH	0.0612 (0.0117,0.1253)	0.4486 (0.3045,0.8142)	0.1801 (0.1392,0.2315)
I3M	0.0870 (0.0468,0.1328)	0.4610 (0.3127,0.7195)	0.1121 (0.0873,0.1392)
4OHI3M	0.1289 (0.0914,0.1721)	0.3420 (0.2640,0.4599)	0.1117 (0.0868,0.1407)
1MOI3M	0.2000 (0.1508,0.2569)	0.3718 (0.2950,0.4940)	0.2508 (0.2137,0.2929)

Table 8: Univariate estimates of the effect of genotype on glucosinolate concentration. Means and standard errors are reported for untransformed data, while P-values are estimated from the log-transformed data. Standard errors are shown in parentheses.

gene	compound	Mean HET	Mean WT	P-value
<i>cyp79f1</i>	3MSOP	0.173 (0.009)	0.266 (0.009)	2.10*10 ⁻¹⁷
	4MSOB	1.580 (0.092)	1.892 (0.096)	1.14*10 ⁻⁶
	5MSOP	0.068 (0.004)	0.056 (0.004)	0.053
	6MSOH	0.325 (0.015)	0.202 (0.016)	1.80*10 ⁻⁷
	I3M	7.904 (0.341)	6.391 (0.358)	0.005
	4OHI3M	0.793 (0.034)	0.731 (0.036)	0.118
	1MOI3M	2.049 (0.112)	1.703 (0.118)	0.031
	<i>cyp83a1</i>	3MSOP	0.250 (0.011)	0.266 (0.009)
4MSOB		1.910 (0.108)	1.892 (0.085)	0.124
5MSOP		0.047 (0.005)	0.056 (0.004)	0.157
6MSOH		0.176 (0.013)	0.202 (0.010)	0.475
I3M		5.966 (0.387)	6.391 (0.303)	0.683
4OHI3M		0.631 (0.041)	0.731 (0.032)	0.129
1MOI3M		1.721 (0.131)	1.703 (0.102)	0.895
<i>sur1</i>		3MSOP	0.307 (0.013)	0.266 (0.009)
	4MSOB	2.100 (0.124)	1.892 (0.089)	0.029
	5MSOP	0.057 (0.005)	0.056 (0.004)	0.737
	6MSOH	0.201 (0.016)	0.202 (0.012)	0.976
	I3M	5.870 (0.419)	6.391 (0.301)	0.660
	4OHI3M	0.725 (0.050)	0.731 (0.036)	0.915
	1MOI3M	1.594 (0.158)	1.703 (0.113)	0.018

Table 9: Univariate estimates of the effect of environmental treatments on glucosinolate concentration. Means and standard errors are reported for untransformed data, while P- are estimated from the log-transformed data. Standard errors are shown in parentheses.

Treatment	Compound	Mean Untreated	Mean Treated	P-value
Water deprivation	3MSOP	0.238 (0.007)	0.238 (0.007)	0.880
	4MSOB	1.802 (0.070)	1.848 (0.072)	0.045
	5MSOP	0.060 (0.003)	0.058 (0.003)	0.526
	6MSOH	0.223 (0.011)	0.256 (0.012)	0.003
	I3M	6.443 (0.256)	7.085 (0.264)	0.019
	4OHI3M	0.832 (0.026)	0.629 (0.026)	5.10*10 ⁻⁷
	1MOI3M	1.676 (0.092)	1.952 (0.095)	0.081
Crushing	3MSOP	0.223 (0.007)	0.253 (0.007)	0.077
	4MSOB	1.636 (0.069)	2.031 (0.072)	0.005
	5MSOP	0.052 (0.003)	0.066 (0.003)	0.800
	6MSOH	0.221 (0.011)	0.258 (0.012)	0.566
	I3M	6.093 (0.252)	7.482 (0.264)	4.78*10 ⁻⁵
	4OHI3M	0.709 (0.026)	0.760 (0.027)	0.671
	1MOI3M	1.939 (0.092)	1.668 (0.096)	0.004
Soil nutrient deprivation	3MSOP	0.240 (0.007)	0.236 (0.007)	0.653
	4MSOB	1.975 (0.071)	1.676 (0.070)	0.001
	5MSOP	0.062 (0.003)	0.056 (0.003)	0.081
	6MSOH	0.254 (0.011)	0.224 (0.011)	0.063
	I3M	7.580 (0.258)	5.945 (0.255)	1.07*10 ⁻⁵
	4OHI3M	0.661 (0.026)	0.805 (0.026)	1.27*10 ⁻⁵
	1MOI3M	1.837 (0.094)	1.783 (0.094)	0.868
Methyl Jasmonate	3MSOP	0.242 (0.007)	0.233 (0.007)	0.596
	4MSOB	1.874 (0.070)	1.772 (0.072)	1.000
	5MSOP	0.051 (0.003)	0.067 (0.003)	0.260
	6MSOH	0.209 (0.011)	0.271 (0.011)	1.71*10 ⁻⁷
	I3M	5.877 (0.252)	7.681 (0.259)	5.18*10 ⁻¹²
	4OHI3M	0.769 (0.026)	0.696 (0.027)	2.819
	1MOI3M	1.427 (0.090)	2.215 (0.093)	2.22*10 ⁻¹⁶

Table 10: P-values from log-transformed data derived from ANOVA of amount
of leaf area removed with * indicating $P < 0.05$.

Gene	genotype	W	C	S	J
<i>cyp79f1</i>	0.0339*	0.1200	0.7615	0.0383*	0.0381*
<i>cyp83a1</i>	0.9106	0.9810	0.8604	0.4399	0.2932
<i>sur1</i>	0.4609	0.9613	0.5532	0.1431	0.0620

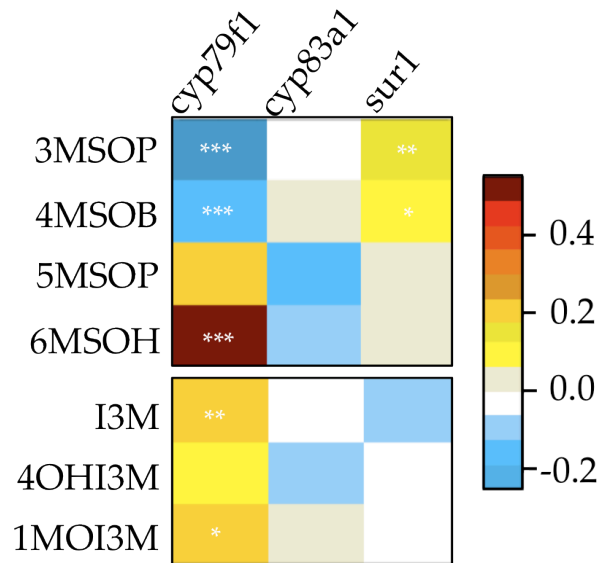


Figure 5: Heatmap of the univariate estimates of the proportional change in GLS concentration of the HET compared to the WT. Environmental conditions are pooled in this analysis, but the untreated controls exhibit the same pattern of GLS change (data not shown). Cool colors indicate a decrease in concentration in the HET compared to the WT and warm colors indicate an increase. * indicate level of significance of the log transformed data. *** $p < 0.0001$, ** $p < 0.01$, * $p < 0.05$.

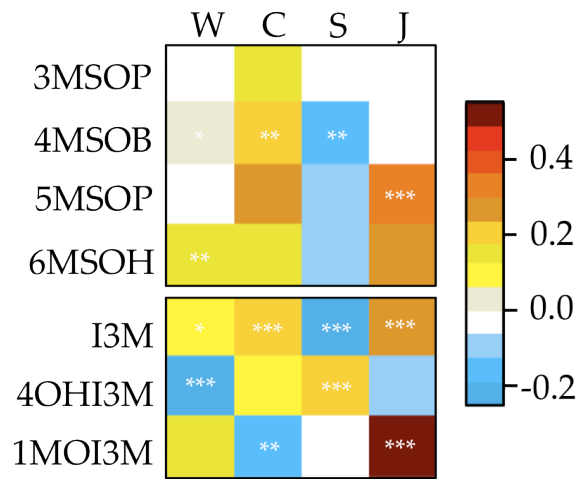


Figure 6: The direction of glucosinolate concentration change for the aliphatic and indolic glucosinolate products. The proportional change in glucosinolate concentration of the seven glucosinolate products (abbreviations on the left) following four environmental treatments (W=water deprivation, C=leaf crushing, S=soil nutrient deprivation, J=methyl Jasmonate treatment). Warm colors indicate higher concentration in treatments than controls, and cool colors indicate decreased concentration in treatments. Stars indicate level of significance *** <0.001, **<0.01, *<0.05. Genes are pooled in this analysis, but all genes produce the same pattern as above (data now shown).

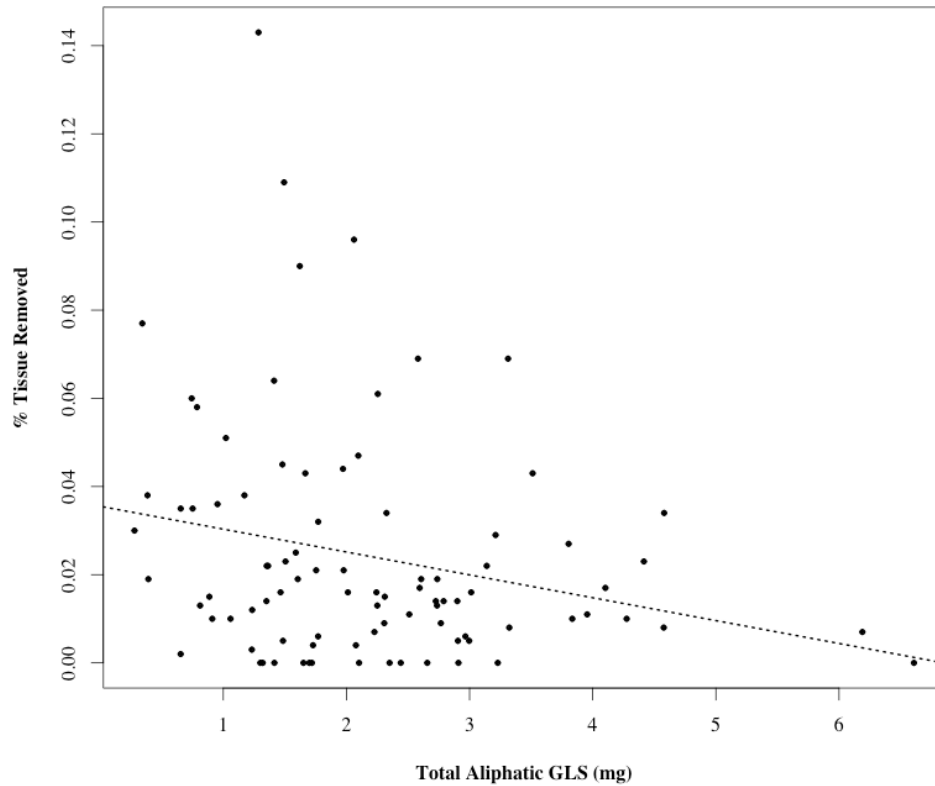


Figure 7: Correlation between the total aliphatic GLS concentration and amount of leaf area removed by *T.ni*. There is a significant correlation ($P=0.01940$, based on log transformed data damage levels) for the amount of leaf area removed and total aliphatic GLS concentration, but no significant correlation with total indolic concentration ($P=0.3609$). Dotted line is the best-fit line fit to untransformed data.

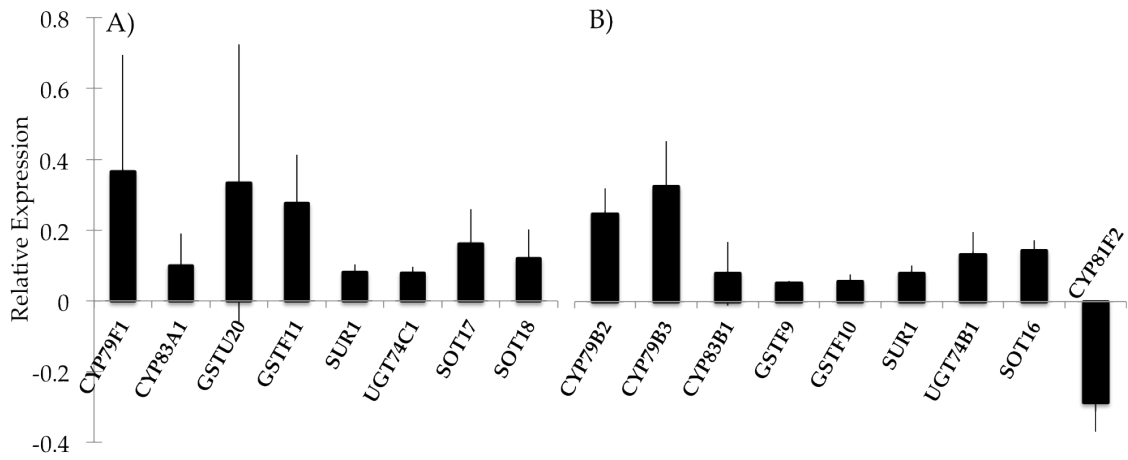


Figure 8: Average relative expression of genes in the A) aliphatic and B) indolic glucosinolate pathways after methyl jasmonate application. Bars are the mean of three different EMBL Gene Expression Atlas experiments and horizontal lines represent standard error. All genes were significantly different from controls in at least two of three experiments.

3.2 Discussion

3.2.1 Flux control under environmental variation

Flux control in the aliphatic glucosinolate pathway is consistently in the first enzyme in the pathway, CYP79F1, under a variety of environmental conditions (Table 6). With our substantially increased sample size in this study, we detected minor flux control for CYP83A1 and SUR1, but the changes they show are much less than those caused by CYP79F1 (Table 8). We find, as in our previous study, that CYP79F1 has much higher control coefficients for 3MSOP and 4MSOB ($\lambda=1.1179$ and $\lambda=0.7396$, respectively) than either CYP83A1 or SUR1 (Table 7). We also find that no enzymes tested have majority control over the other GLS tested.

While we find that HET *Cyp79f1* produce less short-chain aliphatic GLS (3MSOP and 4MSOB), surprisingly, this genotype also has significantly increased quantities of long-chain aliphatic GLS (5MSOP and 6MSOH) (Figure 5). Before the MET-derived products reach the core glucosinolate pathway in the cytosol, they go through chain elongation in the chloroplast (Sonderby et al., 2010). If MET goes through the chain-elongation pathway twice, 3MSOP is produced. If MET goes through three times, 4MSOP is produced, and so on. An abundance of short-chain precursors in the *Cyp79f1* HETs that are not immediately used by the core GLS pathway may incur additional processing in the chain-elongation pathway before entering the core GLS pathway. Our

result of excess long-chain MET-derived GLS in the *Cyp79f1* HETs are consistent with additional processing in the chain-elongation steps.

Cyp79f1 HETs produce more indol GLS than the WT, and, to a lesser extent, *Sur1* HETs produce more short-chain aliphatic GLS than the WT (Figure 5) suggesting extensive crosstalk between the indolic and aliphatic pathways. We first considered the known transcription factors that influence these pathways. The indolic and aliphatic pathways are regulated by different MYB transcription factors; yet, the same environmental cues produce the same effect on MYBs for both pathways. The MYBs are activated under methyl jasmonate and wounding, and repressed under limiting sulfur conditions (Gigolashvili *et al.*, 2009). Thus, based on the environments tested in this study, the regulation of transcription factors does not explain the crosstalk. It is also possible that the expression level of one enzyme could influence the expression of different genes in the pathway. For example, knocking down expression of *Cyp79f1* may increase expression at *Sur1* and that would explain our observed pattern. However, further studies are needed to fully understand how perturbing gene expression influences crosstalk between these pathways.

3.2.2 Genotype by environment interactions

If the level of flux control at a given step changes based on environmental condition, this would cause a significant genotype by environment interaction. We found no evidence for such interactions in the MANOVA (Table 6). Flux control remains

with CYP79F1 in all environmental treatments, and the magnitude of the change in GLS concentration is always strongest in *Cyp79f1* HETs (Figure 5, Table 8). Under more severe conditions, or conditions found in nature, it is possible that predominant flux control might change among these enzymes. However, the molecular signature of positive selection in *Cyp79f1* is consistent with this enzyme having majority control under most conditions and we conclude that flux control lies at CYP79F1 in a broad range of biologically relevant conditions.

3.2.3 Influence of flux control on generalist herbivore feeding

The concentration of aliphatic glucosinolates has a significant negative correlation with amount of damage (linear regression analysis, $P=0.0194$, Figure 7) indicating that the generalist herbivore *T. ni* is sensitive to small changes in glucosinolate concentration. Preliminary ANOVA results suggest that there is a marginally significant ($P=0.0339$) increase in damage in the HETs compared to the WT and that some environmental conditions increase the amount of leaf area removed ($P<0.04$, Table 10). However, these ANOVA results are no longer significant following sequential Bonferoni correction for multiple tests. Further investigation will be needed to determine whether there is indeed more damage in HET over WT *Cyp79f1* individuals.

3.2.4 Gene expression and glucosinolate profile

In three of the four environments tested, the indolic compounds I3M or 1MOI3M change in the same direction, but 4OHI3M changes in the opposite direction (Figure 6). This effect is particularly strong in the methyl jasmonate treatment. Using the EMBL Gene Expression Atlas we harvested data from three microarray experiments conducted on the Affymetrix GeneChip Platform (Accessions E-GEOD-17464 and E-GEOD-18667, E-MEXP-883) that measured gene expression in *A. thaliana* after methyl jasmonate treatment. These experiments find that most genes in the aliphatic and indolic GLS pathways increase in expression, consistent with the nearly ubiquitous increase in GLS concentration in our methyl jasmonate treated plants. One notable exception is *cyp81f2*, which decreases its expression after methyl jasmonate treatment (Figure 8). I3M is converted to 4OHI3M by the enzyme CYP81F2 (Figure 1). Under methyl jasmonate treatment with less CYP81F2 present, we find a decrease in 4OHI3M and an increase in I3M and 1MOI3M, as expected. Although not tested here, it is possible that CYP83F2 has majority control over the expression of these indolic compounds under some conditions and thus, fine-tuning of the specific glucosinolate profile is possible and occurs under some environmental conditions.

3.3 Conclusions

Our results suggest that flux control is robust under a variety of environmental conditions known to alter GLS concentration. If these results are general, natural

selection may be constrained to one, or a few, enzymes to optimize biochemical pathway output making the location of adaptive substitutions predictable. However, this constraint seems to be lessened for the relative proportion of the different indolic products. Under the conditions tested, the proportion of the different indolic GLS is determined by the expression of a gene late in the pathway, *Cyp81f2*. Thus, fine-tuning of the specific glucosinolate profile produced under different conditions may depend on modifier genes that selection could act on separately.

We also find, preliminarily, that herbivores are sensitive to changes in flux. These results could be of great value to agriculture for simple traits. If a single enzyme under most environmental conditions consistently controls flux through an important pathway, engineering efforts can be targeted more effectively.

In the following chapter, I discuss the evolution of a novel glucosinolate in the aliphatic GLS pathway in *Boechera stricta*, a relative of *A. thaliana*. [I'm not sure if I need a transition paragraph.]

3.4 Materials and Methods

To determine if flux control of three genes involved in glucosinolate (GLS) production is influenced by environment, offspring of heterozygous (HET) parents for each of three different genes were grown under various environmental conditions. Briefly, altering the quantity of each enzyme approximated flux control for three genes in the GLS production pathway was accomplished by using *Agrobacterium* TDNA insertion lines that contain a loss-of-function insertion for a single gene and result in a decrease in the amount of mRNA produced for that gene (Olson-Manning et al., 2013). The three genes with a loss-of-function insertion were *Cyp79f1*, *Cyp83a1*, and *Sur1*.

Plants of each of the three lines were subject to all possible combinations of five treatments. These treatments included control conditions, reduced water availability (W), reduced soil nutrient availability (S), mechanical leaf crushing (C), and methyl jasmonate (J) application (detailed below). Treatments were overlapping, with a total of sixteen conditions. For each of the sixteen environmental conditions, each line contained approximately 8 WT and 15 HET replicates, producing a total of 368 individuals per insertion line. All environments were randomized within 24 hours of planting and flats were shuffled weekly.

Seeds from each of the three lines were planted directly on the soil in 16.5 cm² single cell cone-tainers. Seeds were placed in the dark at 4 C for 72 hours after planting

to overcome dormancy. Plants were then grown under long day conditions (16 hours light) at 18 C for 28 days, at which point leaf tissue was harvested.

Randomized racks of ninety-six cone-tainers were placed in standing water from time of planting to tissue collection. To restrict water availability, the bottoms of cone-tainers were allowed to dry and then the bottom of the cone-tainers of low-water treatment individuals were sealed with plastic sleeves (fingers cut from nitrile gloves) eight days prior to tissue collection. This restricted water access of plants in the low water treatment, while control plants for this treatment still received bottom watering.

Plants requiring methyl jasmonate application were separated from the controls during hormone application. Each plant was misted with approximately 0.45 ml of 1:1000 methyl jasmonate (Bodnaryk, 1994) (4.6 molar stock; Sigma-Aldrich) 24 hours prior to tissue collection. Treated plants were covered with clear plastic wrap for one hour then uncovered and returned to randomized racks.

Seeds grown in standard levels of nutrients were planted in Fafard 4P Mix covered with approximately 1 cm of Sunshine #1 Natural & Organic mix. Seeds grown in low nutrient conditions were planted in soil consisting of one part perlite: two parts sand: two parts Fafard 4P Mix. The low nutrient soil was also covered with approximately 1 cm of Sunshine #1 Natural and Organic Mix to allow seeds to germinate and begin growth with minimal mortality.

Mechanical wounding was accomplished 24 hours prior to tissue collection by crushing a single leaf with a corrugated refrigerator clip. A leaf other than the damaged leaf was collected for GLS analysis.

Twenty-five days after planting, tissue was collected for GLS quantification, herbivory analysis, and insertion genotyping. A single true leaf was removed for GLS quantification and for herbivory analysis within the same 24 hour period. For herbivory analysis, the leaf was placed on a moist paper towel (to prevent desiccation) in an 2.5 cm petri dish for 20-24 hours at 23 C with two, second instar *Trichoplusia ni* larvae, (a generalist herbivore). Photographs of each leaf were taken before and after exposure to *T. ni* and percent area consumed by herbivores was calculated using *ImageJ* (NIH).

To determine GLS concentration, each leaf was weighed and stored for 21 days in 2ml 70% methanol at 4 C and then for 7 days at room temperature. GLS quantification and mutant genotyping were performed exactly as in Chapter 2.3.5 and (Olson-Manning et al., 2013).

3.4.1 Materials and Methods for Statistical Analyses

Analysis of glucosinolates produced between wild-type and heterozygous plants under different environmental conditions was conducted with multivariate analysis of variance (MANOVA). GLS concentrations were log transformed to improve normality. MANOVA was performed for each gene in JMP with the concentrations of the seven glucosinolate products as dependent variables and genotype and all four environmental

contrasts as fixed effects (Table 6). We performed two types of analyses in which we either used all the WT individuals from all three lines, or compared each WT to the HET from the same line. The results of both of these analyses agreed for the MANOVA and thus we report only the pooled WT analysis. The full model included the interaction of genotype by each environment (E) and all pairwise interactions of environment by environment.

$$[3MSOP][4MSOB][5MSOP][6MSOH][I3M][4OHI3M][1MOI3M] = genotype + W + C + S + J + genotype * E + E * E$$

When MANOVA was statistically significant, subsequent univariate analyses were used to test the significance of treatment on each glucosinolate compound, using alpha = 0.05 (Scheiner, 2001). As none of the genotype by environment interactions were significant in MANOVA, we pooled wild-type and heterozygous individuals when testing for the effects of treatment. The heat maps were generated with the *heatmap* package in R (Figure 6). The proportional amount of change was calculated by,

$$(Tc - Uc) / Uc$$

where Tc is the concentration of glucosinolate in the treated and Uc is the concentration in the untreated. We found the same qualitative pattern of increase, decrease or no change when analyzed only the untreated individuals, as we would expect from our lack of genotype by environment interaction in the MANOVA. Thus, the environments were pooled for Figure 5 and the genotypes were pooled for Figure 6.

We used linear regression to test whether total aliphatic or total indolic GLS predicted the amount of leaf area that the herbivore *T. ni* removed from a single leaf. We used ANOVA to test whether genotype or treatment had an influence on the amount of leaf area removed.

$$[\text{Total Aliphatic or Indolic GLS}] = \text{genotype} + W + C + S + J$$

We did not have sufficient sample size for the herbivory treatment to test any of the interaction terms. As in the MANOVA, we did two analyses, comparing treatments to the pooled WT, and by comparing each line to its own WT. Surprisingly, we found statistically significant effects of genotype, soil nutrient and jasmonate treatments when we compare only within a line, but a non-significant result when we pool WT lines. These lines are derived from different accessions of *A. thaliana*, and may make a slightly different combination of other compounds to which the herbivores may be sensitive. Therefore, for this analysis it is most appropriate to compare each line to its own WT and those are the results we report here.

Control coefficients were calculated as in Chapter 2, with the same relative expression ratios as determined in that chapter. The ratio of GLS concentration in the HET compared to the WT was estimated 1,000 times in a Python program written by one of the authors (Olson-Manning).

3.4.2 Gene expression meta-analysis

We used available data for a meta-analysis of the effect of methyl jasmonate on the expression of genes in the aliphatic and indolic glucosinolate pathways. Data was harvested from the EMBL Gene Expression Atlas experiment numbers E-GEOD-17464 and E-GEOD-18667 and E-MEXP-883. Each of these experiments was performed with the Affymetrix GeneChip Platform on rosette leaf tissue. Methyl jasmonate significantly changed all the genes analyzed in at least two of the three experiments. Mean values of the control and methyl jasmonate treated individuals were averaged to obtain the proportional change in expression (Figure 8). Negative values indicated a decrease in expression of the treated compared to the untreated plants.

4. The evolution of a novel glucosinolate in *Boechera stricta*

A detailed understanding of how novel functions arise and are maintained in nature has been central to evolutionary biology since its inception. However, few studies have identified the genes that underlie complex trait variation in nature, the evolutionary processes that influence these polymorphisms and the evolutionary steps required to gain novel function. Most work to-date has focused on loss of function mutations that lead to adaptive phenotypes (Barrett & Hoekstra, 2011), likely because gain of function changes occur infrequently and require persistent natural selection to be maintained in populations (Rogers & Hartl, 2012). Nevertheless, development of new functional mechanisms may be crucially important for adaptive evolution. Thus, studies of the origins and selective pressures on gain of function variants are critical for understanding how complex phenotypes evolve in nature.

To understand the adaptive consequences of complex trait variation we must establish a direct relationship from genetic polymorphisms to phenotypic traits, and further to their causal effects on fitness in natural environments (Barrett & Hoekstra, 2011). This requires interdisciplinary studies of ecology, evolution, biochemistry, and genomics.

Work completed in the Mitchell-Olds lab identified a locus that controls a GLS polymorphism in populations of *Boechera stricta*. Briefly, the causal gene was mapped to the first step in the GLS pathway and this locus controls GLS production, damage by

herbivores, and components of fitness in the natural populations that segregate for this polymorphism (Prasad *et al.*, 2012). Here, I will outline the biochemical and structural evidence that led to a gain of function at this locus.

Using the ecological model organism *B. stricta* (Brassicaceae), we mapped a QTL that contributes to insect resistance and controls allocation to glucosinolates derived from branched chain amino acids or methionine (the Branched Chain Methionine Allocation locus: BCMA) (Schranz *et al.*, 2009). Although most Brassicaceae produce GLS from MET or TRP, among the genera closely related to *Arabidopsis*, only *Boechera* and other close relatives produces GLS from VAL or ILE (Windsor *et al.*, 2005) (Figure 2).

We study *B. stricta* because it brings genetic tractability to wild populations where we can measure natural selection in largely undisturbed habitats, in which current environments are similar to historical conditions that have existed for ~3,000 years (Brunelle *et al.*, 2005). *B. stricta* is a native, short-lived perennial (Rushworth *et al.*, 2011), and its close phylogenetic relationship to *Arabidopsis* facilitates genomic analysis, and enabled positional cloning of the *BCMA* locus. Here I examine the specific mutations that led to the evolution of a novel glucosinolate phenotype *Boechera stricta* with a brief mention of the ecological and fitness consequences of this *BCMA* polymorphism.

4.1 QTL cloning and BCMA function

The BCMA polymorphism (Schranz et al., 2009) was mapped to a large region. A probe from this region was prepared and hybridized to *B.stricta* Bacterial Artificial Chromosome (BAC). Positive BAC pools were sequenced from both parents that identified molecular markers for fine scale mapping, as well as several candidate genes in the glucosinolate pathway. Among these were orthologous sequences to *Cyp79f* and *Cyp83a*, the first and second step in the aliphatic GLS pathway (Figure 1).

Fine scale mapping resulted in 9 polymorphic markers in the 1 cM interval containing the BCMA QTL. For markers in this region we used ANOVA to test for marker-trait cosegregation, calculating $LOD = 0.217 \times F\text{-ratio}$ (Haley & Knott, 1992). *Cyp83a* is 0.33 cM from the LOD peak for the BCMA QTL (Prasad et al., 2012). Maximum $LOD = 365.3$ at the *Cyp79f* polymorphism (and two very tightly linked markers), with a 10 LOD confidence interval < 0.1 cM wide. *Cyp79f* had much greater statistical support than the tightly linked *Cyp83a* candidate (Prasad et al., 2012), so we performed transformation experiments in *Arabidopsis* for each locus and allele of the *Cyp79f* gene family.

By transformation (Prasad et al., 2012) and heterologous expression (below) we verified that BCMA belongs to a gene family with three expressed copies, encoding the CYP79F enzymes which catalyze the first step in glucosinolate biosynthesis, orthologous to the *Cyp79f1* - *Cyp79f2* gene pair in *Arabidopsis* (Figure 9). BCMA2 is syntenic with the

cyp79f1 (*At1g16410*) region in *A. thaliana*, but is not linked to the *BCMA* QTL that controls BC-RATIO. *BCMA3* and *BCMA1* are tightly linked at the LOD peak of the *BCMA* QTL on LG7. Alleles of *BCMA3* are present in both parental genotypes, while *BCMA1* is present in the MT genotype, but absent from the CO parent.

We expressed these *BCMA* sequences or empty vector constructs, with 130 independent *A. thaliana* transformants to control for possible position effects and number of insertions. We compared foliar glucosinolates derived from MET, VAL, or ILE in transgenic plants (Prasad et al., 2012), and found highly significant differences in biochemical function of these genes. Briefly, transgenic results show that *BCMA2* and *BCMA1* cause modest increases in Met-GLS, both *BCMA3* copies cause increased production of VAL-GLS (1ME) and *BCMA1* causes an increase in ILE-GLS (2MP).

These analyses show that the *BCMA2* locus retains the ancestral MET activity and a genomic location syntenic with *A. thaliana*, while the *BCMA1* and *BCMA3* loci have novel catalytic activity towards branched chain amino acid substrates, and are found elsewhere in the genome.

4.2 Gene phylogeny and biochemical evolution

Transgenic results in *A. thaliana* show that the enzymes encoded by a clade of *Boechera* genes (*BCMA1* and *BCMA3*, but not *BCMA2*) had acquired catalytic activity towards branched chain amino acid precursors (Windsor et al., 2005). Therefore, we tested for accelerated biochemical evolution in this clade, comparing the rate of

nonsynonymous substitution (d_N) versus synonymous substitution (d_S) with the maximum-likelihood method of (Yang, 2007). Analysis of the branch leading to the *BCAM1,3* clade (dark grey, Figure 9) showed significantly accelerated biochemical evolution ($P = 0.036$). In addition, two amino acid sites (134 and 536) showed strong statistical evidence for rapid evolution (Figure 13, Table 16). Therefore, we mutated these two sites from their ancestral to derived states in order to verify this statistical signal of adaptive molecular evolution.

Heterologous expression of BCMA allowed us to assay catalytic activity towards VAL and ILE. These estimates were largely concordant with results from transgenic analysis of glucosinolate production *in planta*. The BCMA1-MT enzyme had significantly elevated activity towards isoleucine (Table 17). For valine, we found significant catalytic activity for BCMA1, BCMA3, and a modest increase for BCMA2. In contrast to results from transgenic plants, heterologous expression did not detect a significant difference in the rate of valine catalysis of the two alleles of *BCMA3* ($t = 1.09$, $df = 6$, $P = 0.32$). This may reflect differences in experimental variation between transgenic plants and *in vitro* assays, glucosinolate turnover *in vivo*, or between enzyme function *in vitro* vs. *in vivo*.

The overall tertiary structures of eukaryotic P450 proteins are highly conserved in spite of significant divergence in their primary structures (Poulos & Johnson, 2005) allowing us to predict the structures for the BMCA proteins using a chimeric modeling process and alignments with multiple mammalian P450 templates (Baudry *et al.*, 2006).

The predicted structure of BCMA2 (Figure 11) was used to visualize the locations of variations in the BCMA1 and BCMA3 proteins that might explain their altered catalytic functions. In this, G134L, one of the two residues showing the highest statistical evidence for accelerated molecular evolution, occurs in SRS1 near the heme (highlighted in yellow in Figure 11) and is predicted to alter the catalytic site space in the region closest to the heme. The other, P536K, occurs just five amino acids upstream from their carboxy-termini and is predicted to alter electrostatic interactions of this highly flexible tail region. These results provide biochemical verification of the statistical signal of adaptive molecular evolution.

Similar mapping of the two positions varying between the BCMA3-MT and BCMA3-CO alleles indicates that V148L occurs in a region potentially affecting interactions with electron transfer partners and that M268V (highlighted in light blue, Figure 11) occurs in a SRS3 region (highlighted in dark blue) predicted to affect the volume of the upper catalytic site and/or substrate access. The function of these substitutions was beyond the scope of this study, but will be addressed in future work.

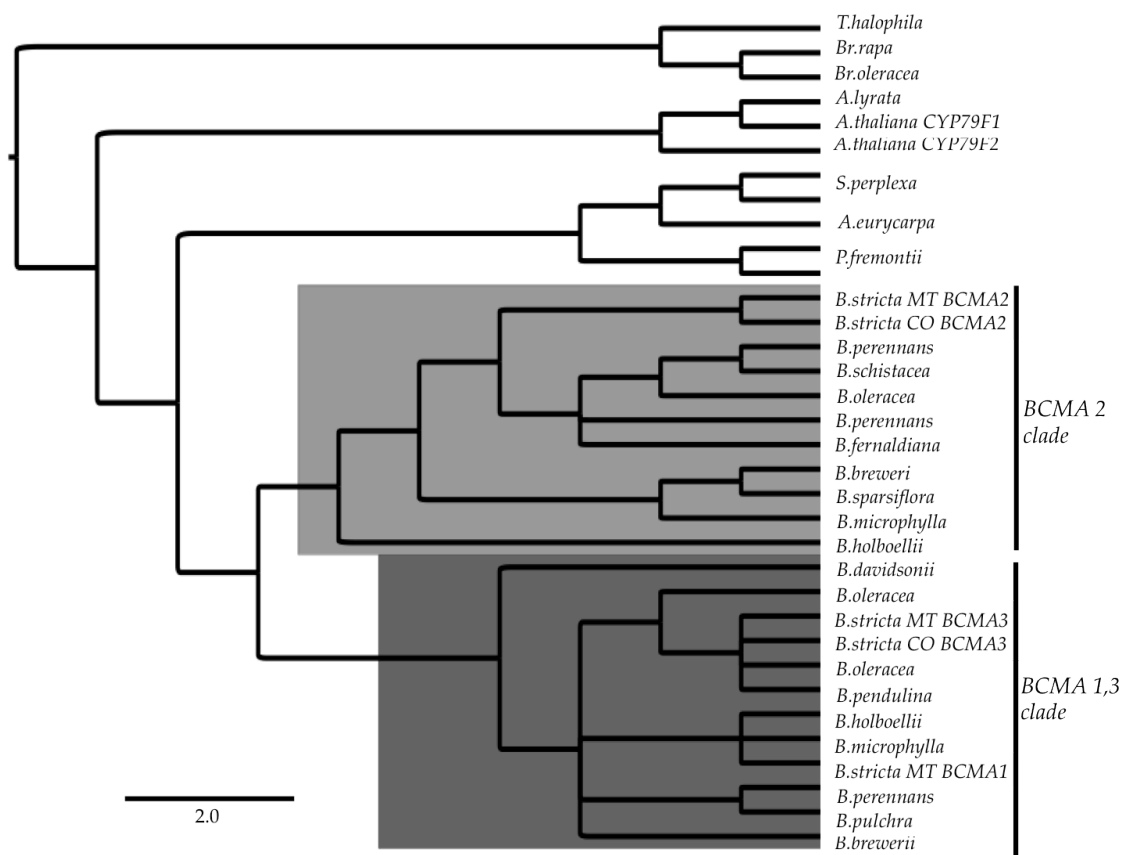


Figure 9: Gene tree of homologs of CYP79F1. Light grey shading is MET-specific BCMA2 clade and dark grey is the BCMA1 and BCMA3 clade (ILE and VAL specific, respectively). The unshaded taxa make only MET aliphatic GLS.

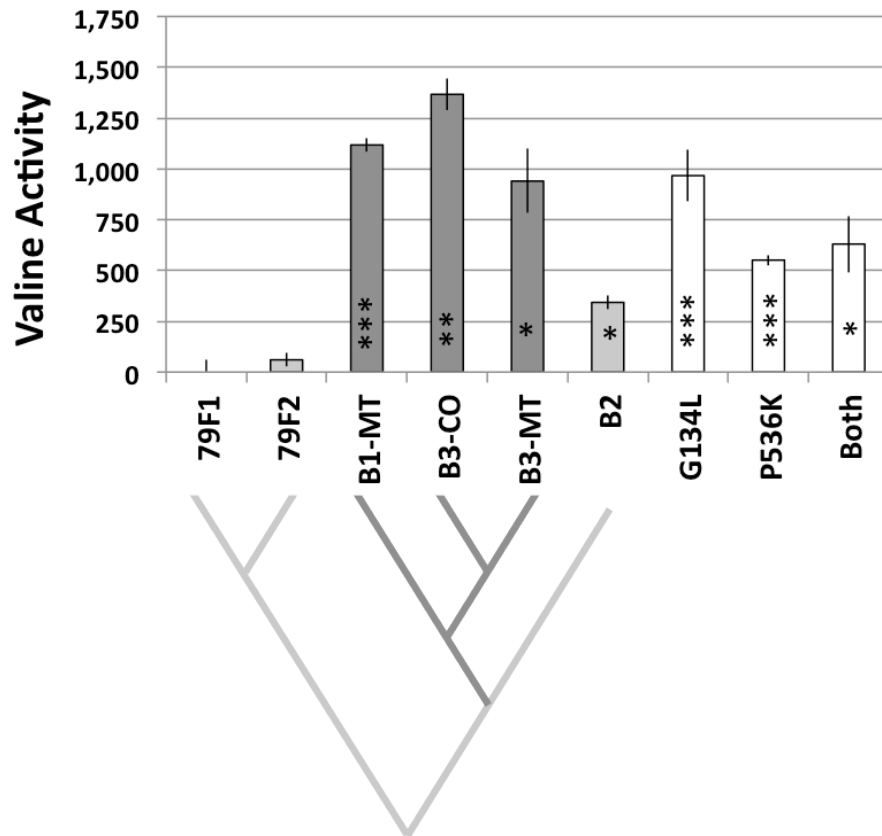


Figure 10: In vitro enzyme activity levels (nmol of product per nmol of enzyme per minute) relative to controls; error bars denote SE. Labels indicate CYP79F enzymes from *Arabidopsis* and BCMA1, BCMA2, and BCMA3 from *Boechera*, with alleles from Colorado or Montana. BCMA1 and BCMA3 gained VAL activity (dark grey). BCMA2 alleles encode identical proteins, so one allele was assayed. BCMA2 (light grey) retains the ancestral MET activity and was engineered to change G134L, P536K, or both (white).

*P < 0.05, **P < 0.01, ***P < 0.001.

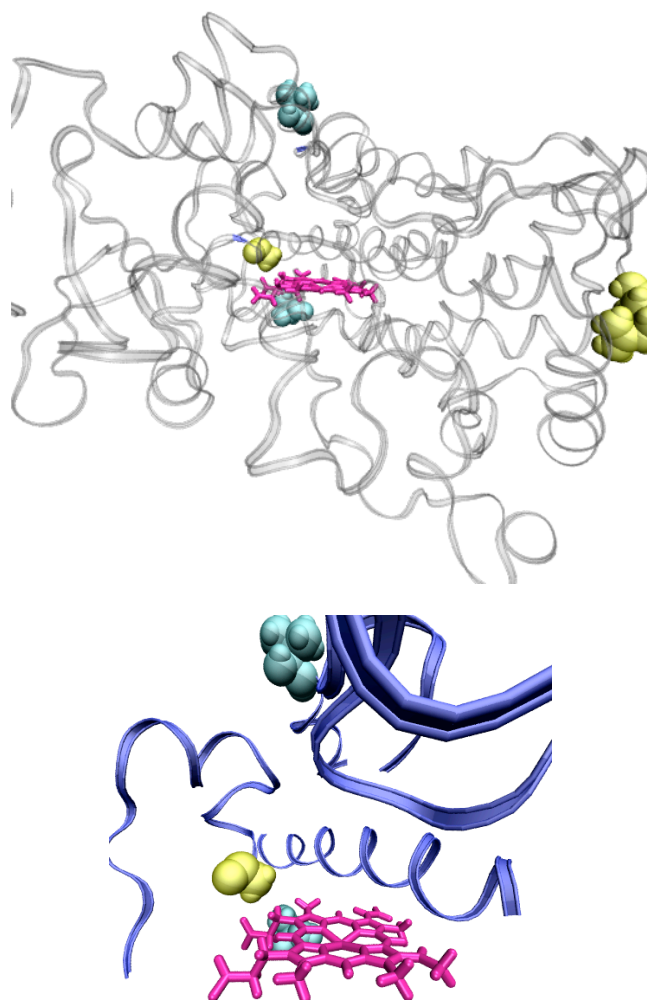


Figure 11: Homology model of BCMA2, a CYP79F4 enzyme responsible for fitness variation in nature. (A) Homology model of BCMA2 with the substrate binding cleft above the heme group (magenta). Amino acid changes G134L and P536K (in yellow) show statistical evidence for accelerated protein evolution, and alter catalytic function when changed by site-directed mutagenesis. G148L and M268V are differences between BCMA1 and BCAM3 (light blue). The location of amino acid 529 is colored, as 530-541 are not depicted in the model. (B) Close up view of substrate binding cleft with mutation G134L residing just above the heme, but G148L and M268V in the substrate recognition regions (in purple).

4.3 Discussion and Conclusions

We cloned an ecologically important QTL and show that it controls defensive chemistry, damage by insect herbivores, survival, and reproduction in the natural populations where this polymorphism evolved. Although the *BCMA* QTL has large effects on plant chemistry, its effects on insect resistance and components of fitness are quantitative, as expected for complex traits. This complex trait variation reflects a gain of novel function in enzyme catalysis following gene duplication, which is modulated by allelic differences in catalytic rate and gene copy number. This new catalytic function is controlled by two amino acid changes that were identified by molecular evolutionary analysis and verified by site directed mutagenesis and heterologous expression.

By integrating approaches drawn from ecology, quantitative genetics, genomics, and biochemistry, we were able to link natural allelic variation at a causal gene to enzymatic function, phenotype and fitness under laboratory and field conditions. That one of the two positively-selected amino acid changes leading to synthesis of new glucosinolates maps in the P450 substrate-binding site reinforces the importance of heme-proximal residues in allowing for the evolution of new biosynthetic activities in the on-going war between insects and plants. That the other positively-selected change maps, quite unexpectedly, near the P450 carboxy-terminus highlights the importance of natural variation for detecting novel changes in catalytic function. Future

interdisciplinary studies of this sort will elucidate the molecular basis of phenotypic evolution and local adaptation.

4.4 Materials and Methods

Full-length cDNA was obtained from *Theellungiella halophila* grown from seeds provided by Dr. Ray Bressan, Purdue University. Total RNA was isolated with the SV Sigma RNA extraction kit with an on-column digestion with DNase1 (Ambion, USA). cDNA was synthesized with the M-MLV reverse transcriptase (Promega) with a poly-T primer. PCR was performed with the PrimeStar Polymerase (Takara) with primers designed from CDS from *T. halophila* from Genbank (BY803083 and BY827035). The PCR products were A-tailed (below) and ligated into the pGEM-T easy vector (Promega). At least 5 clones were sequenced from two separate PCR reactions.

RACE (Rapid Amplification of cDNA ends): For determination of 5' and 3' ends of *CYP79F1* orthologous gene transcripts from LTM, SAD12, and ES913, 2 µg of total RNA was used. Isolated RNA was treated with DNase1 (Ambion, USA) to eliminate genomic DNA contamination.

5' RACE: Gene Racer™ kit (Invitrogen, USA) was used to obtain full-length 5' ends of cDNA sequences according to the manufacturer's protocol. In brief, total RNA was treated with Calf Intestinal Phosphatase (CIP) to remove the 5' phosphates. Dephosphorylated RNA was treated with tobacco acid pyrophosphatase (TAP) to remove the 5' cap structure from the intact full length mRNA. The Gene Racer™ RNA oligo (having a primer site) was ligated to the 5' end of the mRNA using T4 RNA ligase. The RNA oligo provided a priming site for forward Gene Racer PCR Primers

(GeneRacer 5' or GeneRacer 5'nested primer) after the cDNA was synthesized.

Superscript III reverse transcriptase module (Invitrogen, USA) was used for cDNA synthesis. GeneRacer 5' or GeneRacer 5'nested primer in combination with gene specific reverse primers (GSP primer or GSP nested primer) ~350-460bp downstream of the ATG codon were used for amplification of 5' end of the transcripts of *CYP79F1* orthologs from both LTM and SAD12 parents.

3' RACE: In order to obtain 3' ends of the transcripts of *CYP79F1* orthologs, first strand cDNA synthesis was carried out using forward gene specific primer (which are ~350 bp upstream from the stop codon) and GeneRacer Oligo dT primer (Invitrogen, USA). Those mRNAs with a poly-A tail are reverse transcribed. Superscript III reverse transcriptase module (Invitrogen, USA) was used for cDNA synthesis. Gene specific reverse primers (GSP primer or GSP nested primer), upstream of the TAA codon were used for amplification of 3' end of the transcripts of *CYP79F1* from both LTM and SAD12 parents.

Cloning of RACE products: For the amplification of 5' and 3' cDNA ends, high fidelity hot start *Taq* polymerase was used (HS Prime Star *Taq* polymerase, Takara USA) as recommended by the manufacturer. The blunt ends of the resulting amplicons were A-tailed by incubating them in a reaction mixture to a final concentration of 1X PCR buffer with MgCl₂, 0.2 mM dATP along with 5U *Taq* polymerase (Go *Taq* Flexi polymerase, Promega, USA) at 70°C for 30 min. The A-tailed amplicons were ligated into

pGEMT-EASY using pGEMT-EASY cloning kit as per manufacturer instructions (Promega, USA). About 4µl of ligation mixture was used for transformation of *E. coli* (DH5α) chemical competent cells. The transformants were selected on LB medium supplemented with 80µg/ml X-Gal, 0.5 mM IPTG and LB ampicillin (100 mg/L) plate. Two independent PCR reactions were performed for amplification of each RACE product and two independent ligation reactions were carried out with each of the PCR products. About ten single colonies from each ligation reaction were used for further analysis. Plasmid from overnight cultures of each of these colonies was isolated using Plasmid Mini prep kit (Qiagen, USA). Sequencing primers that were specific to T7 promoter and Sp6 promoters (Promega, USA) in the plasmid were used for sequencing of the cloned insert. Sequencing reactions were carried using Big Dye Terminator v3.1 cycle sequencing kit (Applied Biosystems, USA).

Cloning of BCMA:

Total RNA was extracted from LTM and SAD12 parents using the RNeasy total RNA isolation kit (Qiagen, USA) according to the manufacturer's recommendations. Total RNA was treated with RNase free DNase (Invitrogen, USA). First strand cDNA was synthesized with Superscript III reverse transcriptase module (Invitrogen, USA) using 1µg of DNase-treated RNA in a reaction volume of 20 µl. Each preparation of the cDNA was tested for possible contamination with genomic DNA by using the primers flanking the intron regions of known genes. A full length cDNA corresponding to

different copies/alleles of *BCMA* was amplified using 1 µl of the first strand product synthesized from each of the RNA preparations using very high fidelity Taq polymerase (HS Prime Star Taq polymerase, Takara, USA) as per manufacturer's instructions. The primers used for amplification are specified below. Both forward and reverse primers were designed to incorporate *BamHI/KpnI* restriction sites for cloning of *BCMA1* and *BCMA2*. However, to facilitate cloning of *BCMA3*, *BamHI/SmaI* restriction sites were included in the primers. Each of the PCR products was run on a separate Agarose (1%) gel to avoid contamination. Amplicons were excised from the gel, and purified using a gel extraction kit (Qiagen, USA) prior to A-tailing, as described previously. A-tailed PCR products were subsequently cloned into a sequencing vector pGEMT-Easy using the pGEMT-EASY (Promega, USA) cloning kit as per manufacturer instructions. Two independent PCR reactions were used, two independent ligation reactions were performed, and ten independent colonies from each ligation reaction were sequenced. Sequencing was performed on ABI 3700 DNA sequencer with Big dye terminators (PE Applied Biosystems) as described above. The sequencing results were assembled and analyzed using LASERGENE software (DNASTAR, Inc., Madison, WI) using coding sequence of *CYP79F1* from *Arabidopsis thaliana* as template.

Molecular evolution of BCMA:

Based on the cDNA from LTM, SAD12, *Boechera retrofracta*, *Arabidopsis thaliana*, and *Thellungiella halophila*, the coding sequences were aligned with minor manual

adjustment. Stop codon and two codons in the alignment gaps were deleted. There are no gaps in the resulting alignment.

Three methods, maximum parsimony, maximum likelihood, and Bayesian inference, were used to construct the consensus gene tree. PAUP* 4.0b10 (Swofford, 2002) was used to generate maximum parsimony and likelihood trees. Branch support was obtained by bootstrapping 10,000 times for parsimony and 100 times for likelihood. In Bayesian analysis (Huelsenbeck & Ronquist, 2001), two independent tests were performed, each with two million iterations. Trees were sampled every 1,000 generations, and the first 500 sampled trees were discarded as burn-in, retaining the final 1,500. All three methods produced congruent results, and branch length of the consensus tree was calculated with maximum likelihood in PAUP*.

In order to investigate the rate of nonsynonymous substitution versus synonymous substitution, we employed the maximum-likelihood-based method developed by Yang (2007). Based on the consensus gene tree with branch length estimated from model M0 of PAML, the branch-site model was implemented for every branch in the tree. For each focal branch (foreground branch), the branch-site model estimates whether there is any codon with d_N/d_S value greater than 1.0, indicating positive selection. Codons are assigned into different groups based on the d_N/d_S values on the foreground branch and the average of all other branches (background branches). While the null model forces the d_N/d_S value of all codons to be ≤ 1.0 , the alternative

model allows the d_N/d_S values of some codons on the foreground branch to exceed 1.0. Statistical significance is determined by likelihood ratio test with one degree of freedom. All model runs were independently repeated three times to ensure the convergence of maximum likelihood iterations.

NdeI and *XbaI* restriction sites were introduced to the 5' and 3' ends, respectively (underlined) and cloned into the pCWori+ expression vector (Barnes, 1996). Each construct was confirmed by sequencing. Several modifications were made to the 5' end of the protein to increase functional expression of the cytochrome P450s (Duan & Schuler, 2006). Each construct had the first thirty-two amino acids excluded from the 5' end (Duan & Schuler, 2006). The thirty-third amino acid was changed to a start codon and the thirty-fourth was changed from LEU to ALA (Rupasinghe & Schuler, 2006). The next eleven amino acids were changed from plant to *E.coli*-specific codon bias (Gustafsson *et al.*, 2004).

Heterologous Expression of CYP79F Homologs in Escherichia coli: Expression vectors were transformed into the *E.coli* strain DE3(C43). For expression of the *BCMA* constructs a single transformed colony was grown at 37°C overnight in terrific broth medium with ampicillin (50 µg/ml), and 750 µl was used to inoculate 75 ml terrific broth with ampicillin and shaken at 220 rpm at 37°C. After the culture reached approximately OD₆₀₀ 0.7, 1 mM δ -isopropylthio- β -galactoside (IPTG) and 1mM δ -amiolevulinic acid

(ALA) were each added to a final concentration of 1 mM, and 25µg chloramphenicol was added and the cells were grown at 24-26°C shaking at 180 rpm for 20 h.

Isolation of Spheroblasts and NADPH-dependent P450-oxidoreductase: Spheroplasts were made following (Halkier *et al.*, 1995) and were used directly after homogenization for subsequent assays. *Arabidopsis thaliana* ATR1 (NADPH P450 reductase) was prepared and quantified as previously described (Benveniste *et al.*, 1986).

Measurement of Carbon Monoxide Spectrum: Spheroplast lysates were diluted in 50 mM KPi, pH 7.9 with a few mg of sodium thiosulfite and distributed into two cuvettes. A baseline was recorded from 400-500 nm, the sample cuvette was bubbled with CO for 45 s and spectral changes were recorded on a Shimadzu UV-2401PC UV-VIS spectrophotometer at room temperature. The amount of expression functional cytochrome P450 was quantified as the Fe²⁺ CO-bound versus Fe²⁺ difference spectroscopy and quantified using an extinction coefficient of 91 mM⁻¹ cm⁻¹ (Omura & Sato, 1964).

Reconstitution and Activity Measurement of CYP79F Homologs: Spheroplast lysates containing individual recombinant BCMA protein preparations were reconstituted with purified *A.thaliana* ATR1 and catalytic activity was determined *in vitro*. The reaction mixtures (total volume, 30 µl) containing 1.5pmol of CYP79F protein (or an equivalent volume of an identical preparation except the vector had no P450 insert, hereafter vector only control), 0.05 U of *A.thaliana* ATR1, 10.6 mM L-alpha-dioleoyl phosphatidylcholine,

100 μ M of U-¹⁴C-labeled L-amino acid (L-Ile, L-Val; MP Bio), 0.1 M NaCl in 20 mM KPi, pH 7.9 and started with 1mM NADPH. After incubation (5 min at 30 C) the products formed were extracted into ethyl acetate (160 μ l) and added to 10 ml scintillation fluid and quantified on the scintillation counter. Control assays that excluded NADPH did not significantly differ from the vector only controls. Expression towards methionine was not evaluated because elongated-methionine substrates were not available.

Site-directed mutagenesis: From the Bayes Empirical Bayes test (Yang, 2007) result for the branch-site model in PAML, two codons were detected to be under positive selection in the branch leading to BCMA1/BCMA3 in *Boechera* (Fig. S3; Tables S12 & S13). To narrow down the choices for the sites responsible, we compared these sites to regions known to be important in P450s known as substrate recognition regions (SRS) (Gotoh, 1992). Both of the sites with the highest posterior probability of having d_N/d_S greater than one are near these SRS regions.

Site directed mutagenesis was carried out with the Invitrogen GENEART Site-Directed Mutagenesis System according to manufacturers instructions and the recombinant enzymes were expressed and analyzed as above.

Sequence Alignments and Homology Models: Homology models for the predicted CYP79F structures were built using the MOE modeling program (Chemical Computing Group, Montreal, Canada) with the CHARMM22 force field (MacKerell Jr *et al.*, 1998). Using a chimeric modeling approach described in (Baudry *et al.*, 2006) and (Rupasinghe

and Schuler, 2006), the predicted BCMA structures were generated with CYP1A2 (PDB 2HI4) structure for the main template and three different templates (CYP2D6 (PDB 2F9Q), CYP2C9 (PDB 1OG2), CYP3A4 (PDB 1TQN)) for the highly variable B-region that includes SRS1, the FG region that includes SRS2 and SRS3, and the β 4-region that includes SRS6, respectively.

Appendix A – Derivation of flux control coefficients

We show here how the flux control coefficient was estimated for each enzyme. A standard result from metabolic control theory is that the flux control coefficient for enzyme i in a linear pathway with n enzymes, λ_i , is

$$\lambda_i = (1/E_i) / (1/E_1 + 1/E_2 + \dots + 1/E_n) \quad (\text{A.1}),$$

where

$1/E_i = M_i K_{1i} / V_i$, M_i is the Michaelis-Menten parameter for enzyme i , K_{1i} is the equilibrium constant for the initial substrate and the product of enzyme i , and V_i is the V_{\max} parameter for enzyme i (Kacser and Burns 1973).

Equation A.1 can be re-expressed as

$$\lambda_i = (1/E_i) / (1/E_i + \Phi_i) \quad (\text{A.2}),$$

where $\Phi_i = (\sum_{j=1,n} 1/E_j) - 1/E_i$. Dividing A.2. by Φ_i yields

$$\lambda_i = (M_i K_{1i} / V_i \Phi_i) / [(M_i K_{1i} / V_i \Phi_i) + 1] \quad (\text{A.3})$$

and rearrangement yields

$$(M_i K_{1i} / \Phi_i) = V_i [\lambda_i / (1 - \lambda_i)] \quad (\text{A.4}).$$

Also from metabolic control theory, flux for a pathway in which the final product is sequestered is given by

$$\begin{aligned}
 F &= [S K_{1n}] / [(M_i K_{1i} / V_i) + \Phi_i] \\
 &= [S K_{1n} V_i / \Phi_i] / [(M_i K_{1i} / \Phi_i) + V_i]
 \end{aligned}
 \tag{A.5},$$

Where S is the concentration of the initial substrate (Kacser and Burns 1973).

Substituting A.4 into A.5 yields

$$F = [S K_{1n} / \Phi_i] / [\lambda_i / (1 - \lambda_i) + 1]
 \tag{A.6}.$$

Suppose the amount of enzyme i is reduced in heterozygotes to a fraction C_i of that in wild-type individuals. Then the new V_{\max} of the reaction catalyzed by enzyme i is hV_i (Siegel 1975). The flux of this reaction is then

$$\begin{aligned}
 F_{\text{HET}} &= [S K_{1n} C_i V_i / \Phi_i] / [(M_i K_{1i} / \Phi_i) + C_i V_i] \\
 &= [S K_{1n} C_i / \Phi_i] / [\lambda_i / (1 - \lambda_i) + C_i]
 \end{aligned}
 \tag{A.7}.$$

The relative flux for heterozygotes is then

$$F_R = F_{\text{HET}} / F = C_i [[\lambda_i / (1 - \lambda_i) + 1] / [[\lambda_i / (1 - \lambda_i) + C_i]]
 \tag{A.8}.$$

Given that F_R and C_i are known, A.8 can be solved for λ_i to yield

$$\lambda_i = [C_i (1 - F_R)] / [F_R (1 - C_i)]
 \tag{A.9}$$

Appendix B – Supplemental Tables and Figures

Table 11: Primers used for genotyping tDNA insertion lines

TAIR gene ID	Line alias	Line origin	Insertion placement ^a	LP primer / RP primer
AT1G16410	CYP79F1	(SALK) 011806	Exon	CTAGGTCCAAATATTTCCGCC / CACAAGCCTGTCTCTTCCAAC
At1g16400	CYP79F2	(SALK) 129669	Exon*	GCATCTCTGTTTCGACCAGAG / TTTAGTCCCGTGGATTACGTG
AT4G13770	CYP83A1	(WISC) CS856556	Promoter	CACGCCGATGATGATATCTTC / AGTTTCATCCATCAGCAATGG
AT3G03190	GSTF11	(SALK) 014567	Promoter	TTCATGAATCCCTTGTGCTTC / TGCCCATATACTTTGACCACC
At1g78370	GSTU20	(SALK) 080514	Promoter	AAGCTTATCACGCCAATGTTG / TTTTTGTTGTGCGTGAACAAG
AT2G20610	SUR1	(SAIL) CS25414	Exon	AACAACCACACGCCAATCTAG / CCATTATCCCGAGCTTCCTAG
AT1G24100	UGT74B1	(SAIL) CS875580	Promoter	AGAAACCAATGGTGTGAGCAC / GTTGTTACGTTTTGCGTTG
AT2G31790	UGT74C1	(SAIL) CS842120	Promoter	ACTTCCGAAATGGTGAACC / TTGTATGCAATGCGTGAGAG
AT1G18590	SOT17	(SALK) 057288	Exon	GGTTTGCTTGGTAAGCTTCC / CAGGGATAATTGGTGACCTCTG
AT1G74090	SOT18	(WISC) CS853748	Promoter	ACCTTCGAGGAGAGACGGTAG / CTGCTCTCAAAGCTGCAATC

Table 12: Primers for qrtPCR

TAIR gene <i>Abbreviation*</i>	qRT PCR primer
AT1G16410 <i>CYP79F1</i>	TATGTCCCTTCCCATCTTGC GACGCTCCGGTTTGTATACC
At1g16400 <i>CYP79F2</i>	TATGTCCCACCTCATGTTGC GACGCTCCGGTTCGTATGCT
AT4G13770 <i>CYP83A1</i>	CCTTTCGCTTCTGAGTTTACTG AGATACGTCATCCCCACAC
AT3G03190 <i>F11</i>	ACAAAGTATGCGGACCAAGG CTCAACTTCCACCCACTGGT
AT1G78370 <i>U20</i>	CCTTACGGGAGAGCTCAGG GCCTGCTTCTTGTTCCTCAC
AT2G20610 <i>SUR1</i>	TAAGGGATGGGTTGTTCTG GCAGGGTCAGGAGTTACGTC
AT1G24100 <i>B1</i>	TAACCATGAAAATGCTGATTGG TCTTCCATCCGATCATCAAGA
AT2G31790 <i>C1</i>	TTTCCACATGAACACCCTCA AATGCAAAGGGCATAAATGG
AT1G18590 <i>SOT17</i>	GGAAACACGCTTTTCTCGAC CAAACGTGTCCTTTGGGTCT
AT1G74090 <i>SOT18</i>	GTCCTGGTGTTTACGCGAAT GCCATCTCCGGAGTCAGATA
AT4G05320 <i>UBQ10</i>	GGCCTTGATAATCCCTGATGAATAAG AAAGAGATAACAGGAACGGAAACATAGT

Table 13: Mean and standard error for each glucosinolate compound for each insertion line

Compound	Insertion Line	Mean	Standard Error	One-Way ANOVA F	
				Ratio	P-value
3MSOP	SAIL	0.2737	0.0172	1.0498	0.3535
	SALK	0.2982	0.014		
	WISC	0.3133	0.024		
4MSOB	SAIL	1.5919	0.1003	0.1788	0.8365
	SALK	1.664	0.0816		
	WISC	1.6687	0.1401		
5MSOP	SAIL	0.0583	0.0085	0.2604	0.7713
	SALK	0.0503	0.0069		
	WISC	0.0536	0.0118		
6MSOH	SAIL	0.3572	0.0404	1.6495	0.1969
	SALK	0.3151	0.0329		
	WISC	0.231	0.0564		
1MOI3M	SAIL	3.9245	0.6927	0.8547	0.4282
	SALK	4.9589	0.5631		
	WISC	3.9045	0.9667		
4OHI3M	SAIL	0.4394	0.05188	1.3356	0.2673
	SALK	0.4981	0.04217		
	WISC	0.3652	0.07241		
I3M	SAIL	5.477	0.6061	1.533	0.2205
	SALK	6.108	0.4927		
	WISC	4.415	0.8458		

Table 14: Univariate estimates of changes GLS concentration.

Gene	Compound	Mean ($\mu\text{mol/g}$)	Standard Error	HET/WT	P-value
79F1	3MSOP	0.1194	0.0293	0.4075	0.0001*
	4MSOB	0.9683	0.1745	0.5901	0.0006*
	5MSOP	0.0558	0.0145	1.0333	0.5788
	6MSOH	0.3907	0.0703	1.2403	0.8833
	1MOI3M	3.6932	1.1673	0.8322	0.1174
	4OHI3M	0.5167	0.1071	1.1331	0.6111
	I3M	5.8618	1.0892	1.0445	0.5735
83A1	3MSOP	0.305	0.0297	1.0410	0.6838
	4MSOB	1.7819	0.1705	1.0859	0.8551
	5MSOP	0.064	0.0143	1.1852	0.9185
	6MSOH	0.466	0.069	1.4794	0.9993
	1MOI3M	5.0167	1.1749	1.1304	0.7728
	4OHI3M	0.5152	0.0878	1.1298	0.8935
	I3M	6.124	1.0345	1.0912	0.7725
F11	3MSOP	0.2952	0.0315	1.0075	0.5308
	4MSOB	1.6982	0.1825	1.0349	0.617
	5MSOP	0.0284	0.0153	0.5259	0.0461
	6MSOH	0.2555	0.0741	0.8111	0.2129
	1MOI3M	8.2927	1.3426	1.8686	0.9581
	4OHI3M	0.3001	0.0934	0.6581	0.0249
	I3M	6.4538	1.173	1.1500	0.6838
U20	3MSOP	0.3191	0.0329	1.0891	0.7972
	4MSOB	1.8204	0.1901	1.1093	0.8309
	5MSOP	0.0687	0.0179	1.2722	0.7332
	6MSOH	0.2408	0.0788	0.7644	0.2106
	1MOI3M	3.8637	1.3123	0.8706	0.3101
	4OHI3M	0.3806	0.0983	0.8346	0.177
	I3M	4.5259	1.172	0.8065	0.1983
SUR1	3MSOP	0.2533	0.0344	0.8645	0.092
	4MSOB	1.405	0.1992	0.8562	0.0897
	5MSOP	0.0615	0.0177	1.1389	0.6243
	6MSOH	0.1963	0.0817	0.6232	0.0672
	1MOI3M	4.0407	1.4373	0.9105	0.4192

B1	4OHI3M	0.4436	0.1061	0.9728	0.4584
	I3M	8.2179	1.3008	1.4643	0.883
	3MSOP	0.2919	0.0239	0.9962	0.4873
	4MSOB	1.6406	0.1369	0.9998	0.4987
	5MSOP	0.0432	0.0114	0.8000	0.1486
	6MSOH	0.3388	0.0572	1.0756	0.6443
	1MOI3M	5.792	1.0043	1.3051	0.8491
	4OHI3M	0.5886	0.0759	1.2908	0.9041
C1	I3M	7.5729	0.9217	1.3494	0.9229
	3MSOP	0.291	0.2369	0.9932	0.4703
	4MSOB	1.6025	0.1573	0.9765	0.3797
	5MSOP	0.0824	0.0145	1.5259	0.9104
	6MSOH	0.2596	0.0656	0.8241	0.2032
	1MOI3M	5.5436	1.1272	1.2491	0.8119
	4OHI3M	0.5306	0.0826	1.1636	0.851
	I3M	7.7079	0.994	1.3735	0.9644
SOT17	3MSOP	0.2776	0.0191	0.9474	0.2266
	4MSOB	1.7527	0.1198	1.0681	0.761
	5MSOP	0.0489	0.0103	0.9056	0.3668
	6MSOH	0.408	0.0735	1.2952	0.7482
	1MOI3M	3.7544	0.7605	0.8460	0.1924
	4OHI3M	0.4503	0.061	0.9875	0.469
	I3M	7.2385	0.8382	1.2898	0.8871
	SOT18	3MSOP	0.2948	0.0332	1.0061
4MSOB		1.6554	0.1903	1.0088	0.5309
5MSOP		0.0381	0.0159	0.7056	0.1074
6MSOH		0.2813	0.0773	0.8930	0.3166
1MOI3M		4.6607	1.31	1.0502	0.5788
4OHI3M		0.5303	0.1013	1.1629	0.728
I3M		5.7067	1.1528	1.0169	0.5385

Table 15: Tajima's D and Normalized Fay and Wu's H

Gene	Tajima's D	Normalized Fay and Wu's H
CYP79F1	-0.9896	0.55557
CYP79F2	-0.5553	0.46564
CYP83A1	-1.4963	-2.85289
GSTF11	-1.3102	-0.27478
SUR1	-0.1404	0.24017
UGT74B1	0.2093	-1.49408
UGT74C1	-2.2313	-2.36059
SOT17	-1.6899	-0.19733
SOT18	-1.2617	0.48494

Table 16: Likelihood ratio tests for hypothesis of accelerated biochemical evolution using PAML. Tree branches are identified by letters in Figure 13.

branch	lnL-H0	lnL-H1	2lnL	P	Note
A	-4498.77	-4498.77	0.00	1.00	
B	-4498.79	-4498.79	0.00	1.00	
C	-4497.63	-4497.54	0.19	0.66	
D	-4498.77	-4498.77	0.00	1.00	
E	-4498.70	-4498.70	0.01	0.93	
F	-4497.49	-4495.30	4.38	0.04	Including ES913, BCMA1, BCMA3
G	-4498.76	-4498.76	0.00	1.00	
H	-4498.79	-4498.79	0.00	1.00	
I	-4498.79	-4498.79	0.00	1.00	
J	-4493.35	-4493.29	0.12	0.73	
K	-4498.79	-4498.79	0.00	1.00	
L	-4498.79	-4498.79	0.00	1.00	
M	-4496.41	-4485.65	21.51	0.00	Atha-Cyp79F2
N	-4498.79	-4498.79	0.00	1.00	
O	-4498.73	-4492.42	12.63	0.00	Thelungiella

Table 17: Heterologous expression experiments showing in vitro activity of BCMA wild type and mutant enzymes towards alternate substrates. Each cell presents information on mean enzymatic activity (in nmol product/ minute/ nmol cytochrome P450), t-statistic, P-value, and N. Initial fixed effects ANOVA of log expression level showed highly significant interaction between constructs and substrates ($F = 7.36$; $df = 9, 62$; $P < 10^{-6}$; $R\text{-square} = 97.4\%$; $N = 82$), hence individual constructs are compared to controls at $\alpha = 0.05$. Constructs with significantly higher activity than the control are indicated in bold. BCMA2 alleles have identical amino acid sequence, so only one allele was assayed. Cyp79F genes are from *Arabidopsis*. Constructs with site directed mutagenesis were derived from BCMA2, and are indicated by G134L, P536K, or Both. BCMA genes from the two parental genotypes or *Arabidopsis thaliana* are compared to the empty vector control, while genes modified by site directed mutagenesis are compared to BCMA2.

	BCMA3 -CO	BCMA3 -MT	BCMA 1-MT	BCMA2	Cyp79F1	Cyp79F2	G134L	P536K	Both	Empty vector control
VAL	2,450	2,023	2,201	1,426	1,059	1,143	2,393	1,977	2,056	1,083
	7.26	2.81	9.58	2.54	0.14	0.48	4.15	8.24	2.84	
	0.0019	0.0484	<0.0001	0.0313	0.8936	0.6549	0.0089	0.0004	0.0361	
	3	3	5	8	4	3	4	4	4	3
ILE	7,004	11,852	15,299	8,922	11,694	11,023	14,118	12,418	10,136	9,914
	1.73	1.19	3.74	0.51	0.97	0.56	2.03	1.51	0.13	
	0.1593	0.2998	0.0134	0.6252	0.3709	0.5954	0.0887	0.1916	0.8993	
	3	3	5	5	5	5	5	4	4	3

Table 18: Primer sequences relevant to Chapter 4

Experiment	Primer name	Primer sequence
Primers for 5' RACE for LTM	GSPRACE-5'RW3	GATCACCGGAAGCGGCGAAGGTCCCGGA GCCGAAGATATCACCACCATTGTTCTGCTTC
	GSPRACE-5'RW4	CTAT TGATGGCGCGGACTCCGGCGAAGTTGAAAC
	GSPRACE-5' RW5	AT AGGCCGGTCTGCCAAATCAGCGTCTC
	GSPRACE-5' RW6	G GATGCTTCTTGCAGCTGCCAACATGTT
	GSPRACE-5' RW7	CAAT CTCTAACGTCCACCGTCTCGGACCGTT
Primers for 5' RACE for SAD12	GSPRACE5' RW1	GAT AGTGAACGTAAGCAATGAGATTATCC
	GSPRACE5' RW2	GCTTCGA GGTGATGGCGCGGACTCCGGCGAAAT
Primers for 5' RACE for ES913	GSPRACE-5'RW12	TGA GGTCGATCTGCCAAATCAGCGTCTCG
	GSPRACE-5'RW13	TTCTCGA
Primer for 3' RACE for LTM	GSPRACE3' FW2	TACGAGTTTATACCGTTCGGGTC
Primer for 3' RACE for SAD12	GSPRACE3'FW1	GGACAGGTAGCCACATTCATGTATGC
Primer for 3' RACE for ES913	GSPRACE3'FW4	ATGCCGCCGTGGACTAGGCCAGAA CGGGATCCATGATGATGATGAGCCTT
Amplification of BCMA copies from LTM	BCMA1	ACCACATC GGGGTACCCCGTTAAGGACGAATTTT
	BCMA1	TTTATAAAAG CGGGATCCATGATGATGATGAGCCTT
	BCMA2	ACCACATC
	BCMA2	GGGGTACCCCGTTAATGACGAAATTT

		TGGATAAAG
	BCMA3	CGGGATCCATGATGATGATGAGTCTT ACCACATCA
	BCMA3	TCCCCCGGGTTAAGGACGAAATTTTTT ATAAAGGTTTGG
Amplification of BCMA copies from SAD12	BCMA3	CGGGATCCATGATGATGATGAGCCTT ACCACATC
	BCMA3	GGGGTACCCCGTTAAGGACGAAATTT TTTATAAAG
	BCMA2	CGGGATCCATGATGATGATGAGCCTT ACCACATC
	BCMA2	GGGGTACCCCGTTAATGACGAAATTT TGGATAAAG
	BCMA2	TGGATAAAG
Amplification of BCMA from ES913	BCMA-ES913-FW	ATGATGATGATGAGCCTTACCACATC
	BCMA-ES913-RW	TTAAGGACGAAATTTTTTATAAAGG
Heterologous expression	LTM BCMA1,3-F	AAGGGAATTCCATATGGCTTCTCGTC CAACTAAAGCTAAAGATCGTTCTCGC
	SAD12 BCMA1-F	AAGGGAATTCCATATGGCTTCTCGTC CAACTAAAGCTAAAGATAGTTCTCGC
	LTM and SAD12 BCMA2-F	AAGGGAATTCCATATGGCTTCTCGTC CAACTAAAAGCTAAAGATCGTTCTCGC
	LTM BCMA1,3-R	CTGCTCTAGATTAGTGGTGATGATGA GGACGAAATTTTTTATAAAAGTGTTGG
	SAD12 BCMA3-R	CTGCTCTAGATTAGTGGTGATGATGA GGACGAAATTTTTTATAAAGGTTTGG
	BCMA2-R	CTGCTCTAGATTAGTGGTGATGATGA CGAAATTTTGGATAAAGG

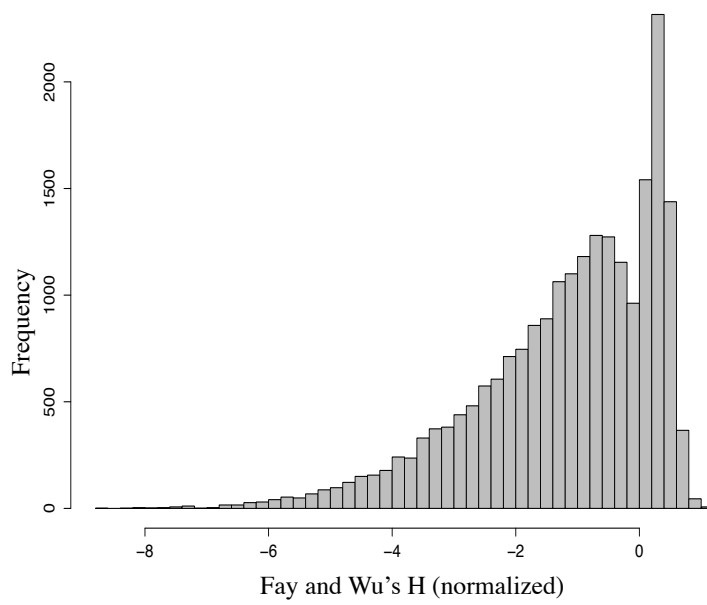
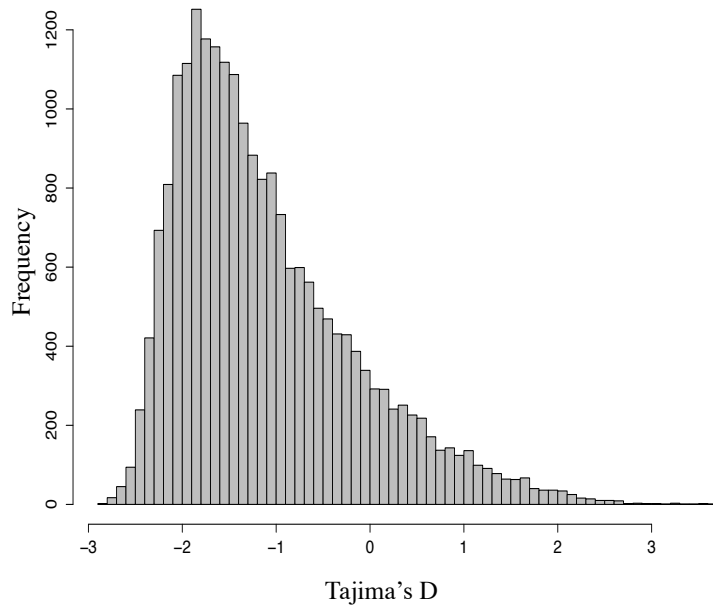


Figure 12: Genome wide distributions of Tajima's D and Fay and Wu's H values for *A.thaliana*.

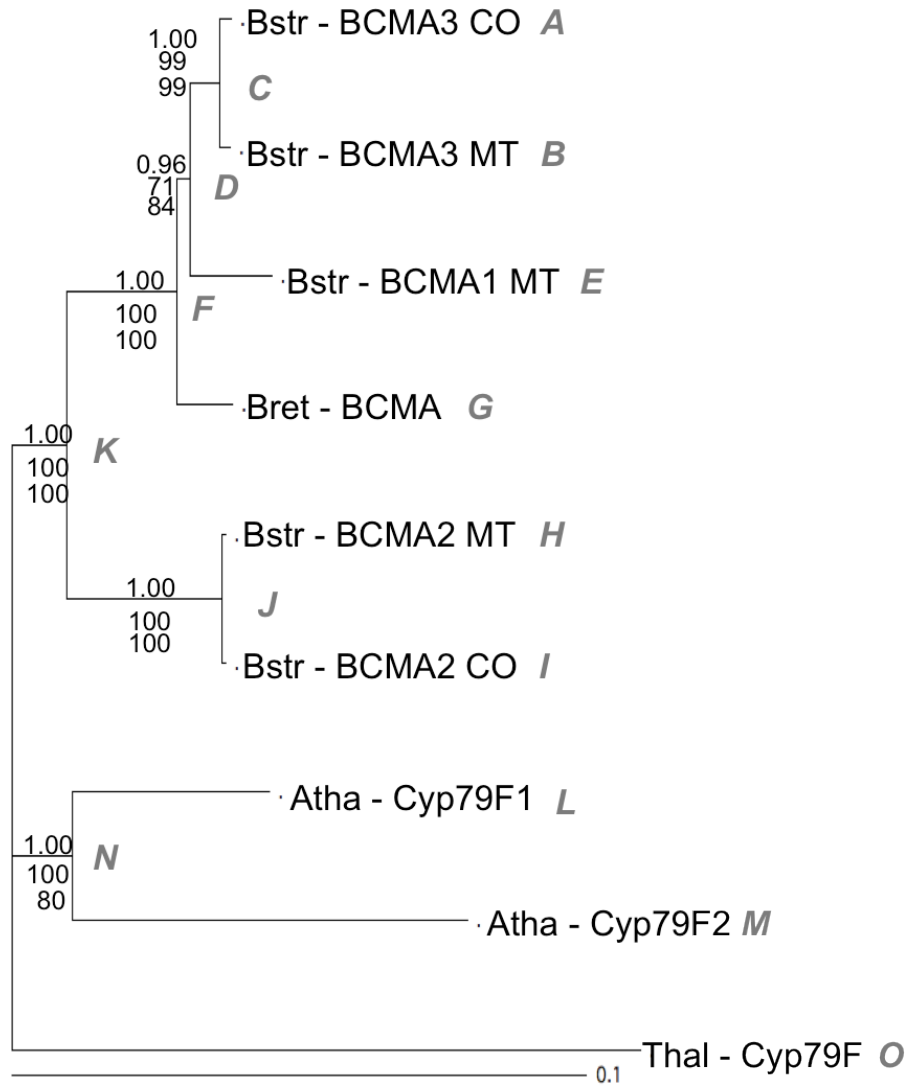


Figure 13: Phylogenetic tree of Cyp79 genes. Shown are sequences from *B. stricta* (Bstr), *B. retrofracta* (Bret), *A. thaliana* (Athal), and *Thellungiella halophila* (Thal). Tree is the consensus from Bayesian inference, maximum likelihood, and maximum parsimony, with bootstrap support indicated from top to bottom, respectively. Grey italic letters denote branches tested in PAML analysis (Table 16).

References

- Alexander, P. J., M. D. Windham, R. Govindarajulu, I. A. Al-Shehbaz & C. D. Bailey, (2010) Molecular phylogenetics and taxonomy of the genus *Thysanocarpus* (Brassicaceae). *Systematic Botany* 35: 559-577.
- Alonso, J. M., A. N. Stepanova, T. J. Leisse, C. J. Kim, H. Chen, P. Shinn, D. K. Stevenson, J. Zimmerman, P. Barajas, R. Cheuk, C. Gadrinab, C. Heller, A. Jeske, E. Koesema, C. C. Meyers, H. Parker, L. Prednis, Y. Ansari, N. Choy, H. Deen, M. Geralt, N. Hazari, E. Hom, M. Karnes, C. Mulholland, R. Ndubaku, I. Schmidt, P. Guzman, L. Aguilar-Henonin, M. Schmid, D. Weigel, D. E. Carter, T. Marchand, E. Risseuw, D. Brogden, A. Zeko, W. L. Crosby, C. C. Berry & J. R. Ecker, (2003) Genome-Wide Insertional Mutagenesis of *Arabidopsis thaliana*. *Science* 301: 653-657.
- Alvarez-Ponce, D., M. Aguadé & J. Rozas, (2009) Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Research* 19: 234-242.
- Alvarez-Ponce, D., M. Aguadé & J. Rozas, (2011) Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: A network-level analysis of selective pressures. *Genome biology and evolution* 3: 87-101.
- Barnes, H., (1996) Maximizing expression of eukaryotic cytochrome P450s in *Escherichia coli*. *Methods in Enzymology* 272: 3-14.
- Barrett, R. D. H. & H. E. Hoekstra, (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics Nat Rev Genet* 12: 767-780.
- Baudry, J., S. Rupasinghe & M. A. Schuler, (2006) Class-dependent sequence alignment strategy improves the structural and functional modeling of P450s. *Protein Engineering, Design and Selection* 19: 345-353.
- Beck, J. B., H. Schmuths & B. A. Schaal, (2008) Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Mol Ecol* 17: 902-915.
- Benveniste, I., B. Gabriac & F. Durst, (1986) Purification and characterization of the NADPH-cytochrome P-450 (cytochrome c) reductase from higher-plant microsomal fraction. *Biochem J* 235: 365-373.

- Blau, P. A., P. Feeny, L. Contardo & D. S. Robson, (1978) Allylglucosinolate and herbivorous caterpillars - contrast in toxicity and tolerance. *Science* 200: 1296-1298.
- Bodnaryk, R. P., (1994) Potent effect of jasmonates on indole glucosinolates in oilseed rape and mustard. *Phytochemistry* 35: 301-305.
- Bouchereau, A., N. Clossais-Besnard, A. Bensaoud, L. Leport & M. Renard, (1996) Water stress effects on rapeseed quality. *European Journal of Agronomy* 5: 19-30.
- Brader, G., M. D. Mikkelsen, B. A. Halkier & E. Tapio Palva, (2006) Altering glucosinolate profiles modulates disease resistance in plants. *The Plant Journal* 46: 758-767.
- Bressan, M., M.-A. Roncato, F. Bellvert, G. Comte, F. e. Z. Haichar, W. Achouak & O. Berge, (2009) Exogenous glucosinolate produced by *Arabidopsis thaliana* has an impact on microbes in the rhizosphere and plant roots. *ISME J* 3: 1243-1257.
- Brown, P. D., J. G. Tokuhisa, M. Reichelt & J. Gershenzon, (2003) Variation of glucosinolate accumulation among different organs and developmental stages of *Arabidopsis thaliana*. *Phytochemistry* 62: 471-481.
- Brunelle, A., C. Whitlock, P. Bartlein & K. Kipfmüller, (2005) Holocene fire and vegetation along environmental gradients in the Northern Rocky Mountains. *Quaternary Science Reviews* 24: 2281-2300.
- Cai, W., J. Pei & N. Grishin, (2004) Reconstruction of ancestral protein sequences and its applications. *Bmc Evolutionary Biology* 4: 33.
- Cao, J., K. Schneeberger, S. Ossowski, T. Gnther, S. Bender, J. Fitz, D. Koenig, C. Lanz, O. Stegle, C. Lippert, X. Wang, F. Ott, J. Müller, C. Alonso-Blanco, K. Borgwardt, K. J. Schmid & D. Weigel, (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet*.
- Casals, F., M. Sikora, H. Laayouni, L. Montanucci, A. Muntasell, R. Lazarus, F. Calafell, P. Awadalla, M. G. Netea & J. Bertranpetit, (2011) Genetic adaptation of the antibacterial human innate immunity network. *Bmc Evolutionary Biology* 11: 202.
- Chen, S., E. Glawischnig, K. Jorgensen, P. Naur, B. Jorgensen, C. E. Olsen, C. H. Hansen, H. Rasmussen, J. A. Pickett & B. A. Halkier, (2003) CYP79F1 and CYP79F2 have

- distinct functions in the biosynthesis of aliphatic glucosinolates in Arabidopsis. *Plant J* 33: 923-937.
- Cork, J. M. & M. D. Purugganan, (2004) The evolution of molecular genetic pathways and networks. *Bioessays* 26: 479-484.
- Czechowski, T., M. Stitt, T. Altmann, M. K. Udvardi & W.-R. Scheible, (2005) Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. *Plant Physiol.* 139: 5-17.
- Daxenbichler, M. E., G. F. Spencer, D. G. Carlson, G. B. Rose, A. M. Brinkler & R. G. Powell, (1991) Glucosinolate composition of seeds from 297 species of wild plants. *Phytochemistry* 30: 2623-2638.
- Delgado, J., Liao, J.C. , (1992) Determination of Flux Control Coefficients from transient metabolite concentrations. *Biochemical Journal* 282: 919.
- Duan, H. & M. A. Schuler, (2006) Heterologous expression and strategies for encapsulation of membrane-localized plant P450s. *Phytochemistry Reviews* 5: 507-523.
- Eanes, W. F., (1999) Analysis of selection on enzyme polymorphisms. *Annual Review of Ecology & Systematics* 30: 301.
- Eanes, W. F., (2011) Molecular population genetics and selection in the glycolytic pathway. *J Exp Biol* 214: 165-171.
- Eckert, A. J., J. D. Liechty, B. R. Tearse, B. Pande & D. B. Neale, (2010) DnaSAM: Software to perform neutrality testing for large datasets with complex null models. *Molecular Ecology Resources* 10: 542-545.
- Fahey, J. W., A. T. Zalcmann & P. Talalay, (2001) The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry* 56: 5-51.
- Fan, J., C. Crooks, G. Creissen, L. Hill, S. Fairhurst, P. Doerner & C. Lamb, (2011) Pseudomonas sax Genes Overcome Aliphatic Isothiocyanate-Mediated Non-Host Resistance in Arabidopsis. *Science Signaling* 331: 1185-1106.

- Flowers, J. M., E. Sezgin, S. Kumagai, D. D. Duvernell, L. M. Matzkin, P. S. Schmidt & W. F. Eanes, (2007) Adaptive evolution of metabolic pathways in *Drosophila*. *Molecular Biology And Evolution* 24: 1347-1354.
- Gaut, B., L. Yang, S. Takuno & L. E. Eguiarte, (2011) The Patterns and Causes of Variation in Plant Nucleotide Substitution Rates. *Annual Review of Ecology, Evolution, and Systematics* 42: 245-266.
- Gigolashvili, T., B. Berger & U.-I. Flügge, (2009) Specific and coordinated control of indolic and aliphatic glucosinolate biosynthesis by R2R3-MYB transcription factors in *Arabidopsis thaliana*. *Phytochemistry Reviews* 8: 3-13.
- Goldman, N., J. P. Anderson & A. G. Rodrigo, (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49: 652-670.
- Gotoh, O., (1992) Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *Journal of Biological Chemistry*.
- Grubb, C., B. Zipp, J. Ludwig-Muller, M. Masuno, T. Molinski & S. Abel, (2004) *Arabidopsis* glucosyltransferase UGT74B1 functions in glucosinolate biosynthesis and auxin homeostasis. *Plant Journal* 40: 893-908.
- Gustafsson, C., S. Govindarajan & J. Minshull, (2004) Codon bias and heterologous protein expression. *Trends In Biotechnology*.
- Haddrill, P. R., K. R. Thornton, B. Charlesworth & P. Andolfatto, (2005) Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* 15: 790-799.
- Haley, C. S. & S. A. Knott, (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69: 315-324.
- Halkier, B. A. & J. Gershenzon, (2006) Biology and biochemistry of glucosinolates. *Annual Review of Plant Biology*.
- Halkier, B. A., H. L. Nielsen, B. Koch & B. L. Møller, (1995) Purification and characterization of recombinant cytochrome P450TYR expressed at high levels in *Escherichia coli*. *Archives Of Biochemistry And Biophysics* 322: 369-377.

- Hansen, T. F. & G. P. Wagner, (2001) Modeling genetic architecture: A multilinear theory of gene interaction. *Theoretical Population Biology* 59: 61-86.
- Hartl, D. L., M. Medhura, L. Green & D. E. Dykhuizen, (1986) The evolution of DNA sequences in *Escherichia coli*. *Phil. Trans. T. Soc. Lond. B* 312: 191-204.
- Holloway, A., M. Lawniczak, J. Mezey, D. Begun & C. Jones, (2007) Adaptive Gene Expression Divergence Inferred from Population Genomics. *PLoS Genetics* 3: e187.
- Hopkins, R. J., N. M. van Dam & J. J. A. van Loon, (2009) Role of Glucosinolates in Insect-Plant Relationships and Multitrophic Interactions. *Annual Review of Entomology*. 54: 57-83.
- Horton, M. W., A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, A. Auton, N. W. Muliyati, A. Platt, F. G. Sperone & B. J. Vilhlmsson, (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics* 44: 212-216.
- Hu, T. T., P. Pattyn, G. B. Erica, J. Cao, J.-F. Cheng, R. M. Clark, N. Fahlgren, J. A. Fawcett, G. Jane, H. Gundlach, G. Haberer, J. D. Hollister, S. Ossowski, R. P. Ottillar, A. A. Salamov, K. Schneeberger, M. Spannagl, X. Wang, L. Yang, M. E. Nasrallah, J. Bergelson, J. C. Carrington, B. S. Gaut, S. Jeremy, K. F. X. Mayer, Y. van de Peer, I. V. Grigoriev, N. Magnus, D. Weigel & Y.-L. Guo, (2010) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. 1-22.
- Huelsenbeck, J. P. & F. Ronquist, (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754-755.
- Jovelin, R. & P. C. Phillips, (2011) Expression level drives the pattern of selective constraints along the insulin/tor signal transduction pathway in *Caenorhabditis*. *Genome biology and evolution* 3: 715-722.
- Kacser, H. & J. A. Burns, (1973) The control of flux. *Symp Soc Exp Biol* 27: 65-104.
- Kliebenstein, D. J., J. Kroymann & T. Mitchell-Olds, (2005) The glucosinolate-myrosinase system in an ecological and evolutionary context. *Current Opinion in Plant Biology* 8: 264-271.

- Kliebenstein, D. J., V. M. Lambrix, M. Reichelt, J. Gershenzon & T. Mitchell-Olds, (2001) Gene duplication in the diversification of secondary metabolism: Tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. *Plant Cell* 13: 681-693.
- Koritsas, V. M., J. A. Lewis & G. R. Fenwick, (1991) Glucosinolate responses of oilseed rape, mustard and kale to mechanical wounding and infestation by cabbage stem flea beetle (*Psylliodes chrysocephala*). *Annals of Applied Biology*. 118: 209-221.
- Kryazhimskiy, S. & J. B. Plotkin, (2008) The population genetics of dN/dS. *PLoS Genetics* 4: e1000304.
- Lambrix, V., M. Reichelt, T. Mitchell-Olds, D. Kliebenstein & J. Gershenzon, (2001) The *Arabidopsis* epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences *Trichoplusia ni* herbivory. *Plant Cell* 13: 2793-2807.
- Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel & E. Huala, (2011) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*. 40: D1202-D1210.
- Livingstone, K. & S. Anderson, (2009) Patterns of Variation in the Evolution of Carotenoid Biosynthetic Pathway Enzymes of Higher Plants. *Journal of Heredity* 100: 754-761.
- Loytynoja, A. & N. Goldman, (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* 102: 10557.
- Lu, Y. & M. D. Rausher, (2003) Evolutionary Rate Variation in Anthocyanin Pathway Genes. *Mol Biol Evol* 20: 1844-1853.
- Luisi, P., D. Alvarez-Ponce, G. M. Dall'Olio, M. Sikora, J. Bertranpetit & H. Laayouni, (2012) Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. *Mol Biol Evol* 29: 1379-1392.
- Ma, X., Z. Wang & X. Zhang, (2010) Evolution of dopamine-related systems: biosynthesis, degradation and receptors. *Journal of Molecular Evolution* 71: 374-384.

- MacKerell Jr, A., D. Bashford & M. Bellott, (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*.
- Mailer, R. J. & P. S. Cornish, (1987) Effects of water stress on glucosinolate and oil concentrations in the seeds of rape (*Brassica napus* L.) and turnip rape (*Brassica rapa* L. var. *silvestris*[Lam.] Briggs). *Australian Journal of Experimental Agriculture* 27: 707.
- McDonald, J. H. & M. Kreitman, (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351: 652-654.
- Mikkelsen, M., V. Fuller, B. Hansen, M. Nafisi, C. Olsen, H. Nielsen & B. A. Halkier, (2009) Controlled indole-3-acetaldoxime production through ethanol-induced expression of CYP79B2. *Planta*.
- Nielsen, R., (2005) Molecular signatures of natural selection. *Annual Review of Genetics* 39: 197-218.
- Olson-Manning, C. F., C.-R. Lee, M. D. Rausher & T. Mitchell-Olds, (2013) Evolution of Flux Control in the Glucosinolate Pathway in *Arabidopsis thaliana*. *Molecular Biology And Evolution* 30: 14-23.
- Olson-Manning, C. F., M. R. Wagner & T. Mitchell-Olds, (2012) Adaptive evolution: evaluating empirical support for theoretical predictions. *Nature Reviews Genetics* *Nat Rev Genet* 13: 867-877.
- Omura, T. & R. Sato, (1964) The carbon monoxide-binding pigment of liver microsomes I. Evidence for its hemoprotein nature. *Journal Of Biological Chemistry*.
- Pauwels, L., D. Inzé & A. Goossens, (2009) Jasmonate-inducible gene: what does it mean? *Trends In Plant Science* 14: 87-91.
- Pfeiffer, T., O. S. Soyer & S. Bonhoeffer, (2005) The Evolution of Connectivity in Metabolic Networks. *PLoS Biology* 3: e228.
- Poulos, T. & E. Johnson, (2005) Structures of cytochrome P450 enzymes. *Cytochrome P450*: 87-114.
- Prasad, K. V. S. K., B.-h. Song, C. Olson-Manning, J. T. Anderson, C.-R. Lee, M. E. Schranz, A. J. Windsor, M. J. Clauss, A. J. Manzaneda, I. Naqvi, M. Reichelt, J.

- Gershenzon, S. G. Rupasinghe, M. A. Schuler & T. Mitchell-Olds, (2012) A gain-of-function polymorphism controlling complex traits and fitness in nature. *Science* 337: 1081-1084.
- Ramsay, H., L. Rieseberg & K. Ritland, (2009) The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. *Molecular Biology And Evolution*.
- Rausher, M., Y. Lu & K. Meyer, (2008) Variation in Constraint Versus Positive Selection as an Explanation for Evolutionary Rate Variation Among Anthocyanin Genes. *Journal of Molecular Evolution* 67: 137-144.
- Rausher, M. D., R. E. Miller & P. Tiffin, (1999) Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Molecular Biology And Evolution* 16: 266-274.
- Rodman, J., P. Soltis, D. Soltis, K. Sytsma & K. Karol, (1998) Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. *American Journal of Botany* 85: 997-997.
- Rogers, R. L. & D. L. Hartl, (2012) Chimeric Genes as a Source of Rapid Evolution in *Drosophila melanogaster*. *Molecular Biology And Evolution* 29: 517-529.
- Rogers, S. & A. Bendich, (1988) Extraction of DNA from plant tissues. In: *Plant Molecular Biology Manual*. pp. A6:1-10.
- Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer & R. Rozas, (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-2497.
- Rupasinghe, S. & M. A. Schuler, (2006) Homology modeling of plant cytochrome P450s. *Phytochemistry Reviews* 5: 473-505.
- Rushworth, C. A., B.-h. Song, C.-R. Lee & T. Mitchell-Olds, (2011) *Boechera*, a model system for ecological genomics. *Mol Ecol*.
- Scheiner, S. M., (2001) Multiple response variables and multispecies interactions. *Design and analysis of ecological experiments*: 99-133.

- Schranz, M. E., A. J. Manzaneda, A. J. Windsor, M. J. Clauss & O. Mitchell, (2009) Ecological genomics of *Boechera stricta*: identification of a QTL controlling the allocation of methionine- vs branched-chain amino acid-derived glucosinolates and levels of insect herbivory. *Heredity* 102: 465-474.
- Schwanhausser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen & M. Selbach, (2011) Global quantification of mammalian gene expression control. *Nature* 473: 337-342.
- Sharbel, T. F., B. Haubold & T. Mitchell-Olds, (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and post-glacial colonization of Europe. *Molec. Ecol.* 9: 2109-2118.
- Slotte, T., T. Bataillon, T. T. Hansen, K. St Onge, S. I. Wright & M. H. Schierup, (2011) Genomic Determinants of Protein Evolution and Polymorphism in *Arabidopsis*. *Genome biology and evolution* 3: 1210-1219.
- Sonderby, I. E., F. Geu-Flores & B. A. Halkier, (2010) Biosynthesis of glucosinolates-- gene discovery and beyond. *Trends In Plant Science* 15: 283-290.
- Stoletzki, N. & A. Eyre-Walker, (2011) Estimation of the neutrality index. *Molecular Biology And Evolution* 28: 63-70.
- Sussman, M. R., R. M. Amasino, J. C. Young, P. J. Krysan & S. Austin-Phillips, (2000) The *Arabidopsis* knockout facility at the University of Wisconsin-Madison. *Plant Physiology* 124: 1465-1467.
- Swofford, D. L., (2002) *PAUP*: phylogenetic analysis using parsimony (*and other methods)*. Sinauer, Sunderland, Massachusetts, USA.
- Tantikanjana, T., M. D. Mikkelsen, M. Hussain, B. A. Halkier & V. Sundaresan, (2004) Functional analysis of the tandem-duplicated P450 genes SPS/BUS/CYP79F1 and CYP79F2 in glucosinolate biosynthesis and plant development by Ds transposition-generated double mutants. *Plant Physiology* 135: 840-848.
- Timm, N. H., (1975) *Multivariate analysis, with applications in education and psychology*. Brooks/Cole Monterey, CA.

- Vidmar, G. & M. Pohar, (2005) Augmented convex hull plots: Rationale, implementation in R and biomedical applications. *Computer methods and programs in biomedicine* 78: 69-74.
- Vogel, C. & E. M. Marcotte, (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics Nat Rev Genet* 13: 227-232.
- Watt, W. B. & A. M. Dean, (2000) Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. *Annu. Rev. Genet.* 34: 593-622.
- Windsor, A. J., M. Reichelt, A. Figuth, A. Svatos, J. Kroymann, D. J. Kliebenstein, J. Gershenzon & T. Mitchell-Olds, (2005) Geographic and evolutionary diversification of glucosinolates among near relatives of *Arabidopsis thaliana* (Brassicaceae). *Phytochemistry* 66: 1321-1333.
- Wright, K. M. & M. D. Rausher, (2010) The Evolution of Control and Distribution of Adaptive Mutations in a Metabolic Pathway. *Genetics* 184: 483-502.
- Yang, L. & B. S. Gaut, (2011) Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Molecular Biology And Evolution*.
- Yang, Z., (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24: 1586-1591.
- Zera, A. J., (2011) Microevolution of intermediary metabolism: evolutionary genetics meets metabolic biochemistry. *J Exp Biol* 214: 179-190.

Biography

Carrie Olson-Manning was born June 17th, 1985 in Fargo, ND. She received her Bachelor of Science with an emphasis on Genetics and Evolution from the University of Minnesota in May 2007. The titles of papers she published while pursuing her Ph.D. are: "A Gain-of-Function Polymorphism Controlling Complex Traits and Fitness in Nature" published in *Science*, "Evolution of flux control in the glucosinolate pathway in *Arabidopsis thaliana*" published in *Molecular Biology and Evolution*, and "Adaptive evolution: evaluating empirical support for theoretical predictions" a review published in *Nature Reviews Genetics*. Carrie received the following honors and fellowships: Duke Bass teaching fellowship, a twice invited speaker for the Society of Molecular Biology and Evolution, Biology Grant-in-Aid, Duke Graduate Travel Fellowship, and the National Science Foundation Doctoral Dissertation Improvement Grant.